Time-varying Media Coverage and Stock Returns

Hannan Zheng[*]

January 5, 2021[†]

Abstract

I show that news editors have state-dependent preferences for different types of firms. Using New York Times data and natural language processing techniques, I estimate the loadings of media coverage on eight common firms features and extract the corresponding editor preferences. I find that firms with high editor preference earn higher returns than those with low preference, on average. This is consistent with a theory that posits that, if investors delegate their information selection to news editors, the state-dependent coverage decisions signal risky features and hence covered firms require more risk compensation. The annualized excess return of around 12% due to coverage cannot be explained by standard risk factors. This excess return is also present among non-covered firms but is more short-lived. (*JEL*: G12, G14)

[*] Zheng is at the Department of Finance, Boston University Questrom School of Business: Phone (617) 955-0857, email: zhannan@bu.edu.

[†] I am grateful to Gustavo Schwenkler, Andrea Vedolin, Steven Kou, Evgeny Lyandres, Xing Hao, Fernando Zapatero, Max Reppen, and other seminar participants at Boston University for useful comments and suggestions. I also thank Gustavo Schwenkler for providing the New York Times' API.

1 Introduction

How do we choose what to read when there are thousands of papers and articles in the world? As described in Kahneman (1973), attention is a scarce resource. Therefore, as scholars, we delegate a part of our selection to editors at top journals. By choosing which papers to publish, editors inform us what topics they think have the most potential or largest impact from their professional points of view. The same thing happens to investors. With countless events occurring everyday, each of them having the potential to affect investors' consumption and investment, it is almost impossible for individuals to extract useful information. As a result, investors also delegate part of their selection to news editors.

Nimark and Pitschner (2019) show under general conditions that agents in an economy can reduce their information entropy by outsourcing their information choice to an entity that makes state-dependent reporting decisions, compared to a situation in which they select by themselves. In this model, agents receive information in two distinct ways: via the content of news reporting about events, and via the editorial decision on what events to cover. If this theory holds true, then editors' coverage decisions should have an impact on agents' investment decisions. For instance, under a similar theoretical framework, Chahrour et al. (2019) shows that if state-dependent editors believe riskier sectors are more newsworthy, then time-varying media focus can explain additional aggregate fluctuations seen in the data due to the under- or overreaction of firms. In this paper, I show more empirical evidence from stock returns to support this channel. My results are threefold. First, I show that news editors have time-varying preferences for different characteristics of firms. Second, I show that firms with high editor preference earn higher returns than firms with low preference. A related trading strategy attains an annualized alpha of 12% after controlling for common risk factors. Lastly, I show that such return patterns not only exist among firms that are covered in news, but also among those that are never reported in my news data. However, this excess return is relatively short-lived.

To estimate editor preferences that are important for investors, I use business news articles from the New York Times between January 1995 and December 2015. I apply a natural language processing methodology introduced in Schwenkler and Zheng (2019) to identify any mentions of firms and match them in the Compustat/CRSP database. Then, I use the natural logarithm of the number of mentions in a given month as a media coverage measure for each firm. I find that the news unconditionally prefers firms with large market capitalizations. Therefore, I restrict my focus to the largest 500 firms in the Compustat/CRSP universe sorted by market value every month. I monthly run a cross-sectional regression with firm fixed-effect to study the correlation between media coverage and 9 common features (including market value) of firms.

My regression first shows that, over 20 years, media coverage constantly has a significant and positive loading on market value, which is consistent with the existing literature. It implies that news editors consistently prefer larger firms and thus any related reporting decisions convey very limited information about current states to investors. I therefore primarily consider media coverage in excess of firm size.

Media coverage has highly changeable loadings on the other eight features, either in the sense of coefficient sign or of significance level. This shows that, at different times, editors focus on different types of firms. If the news selection function used by editors is consistent with Chahrour et al. (2019) – that is, riskier firms are considered more newsworthy – then such focus shift implies the editors' judgement on the risks that are associated with different types of firms. For example, if the news tends to cover value firms in month t, then it reflects editors' concurrent concern of the value factor. On the other hand, in month $t + k$ the concern can be flipped such that the news prefers growth firms instead. Consequently, I use the time-varying projection of media coverage on the eight firm features as my estimate of editor preferences for risks.

After sorting on editor preference at the end of each sample month, I find that firms in the high group outperform those in the low group in the following month, indicating that the former require more compensation in the market. This result is robust to controlling for risk factors in Fama and French (1993), Carhart (1997) and Frazzini and Pedersen (2014). More than that, this return difference exists among firms covered previously in the news and also non-covered firms, while for the former the difference is larger. A long-short strategy that holds covered firms with high editor preference and shorts non-covered firms with low editor preference earns a significant 12% annualized alpha. Such result should be expected if the theory in Nimark and Pitschner (2019) holds true. If the news starts to cover one type of firms more than others, it signals to investors that this type of firms are riskier than the rest. As long as investors believe this signal is credible, they will react accordingly and affect asset prices even though the underlying fundamentals may not change yet.[1]

I study what is the determinant of the coverage-induced return difference. First, I show that this return difference is not determined by any specific features that compose the editor

[1] Since the New York Times is one of the biggest news outlets in the United States, it is reasonable to believe this is true.

preferences. I construct high and low portfolios based on the time-varying projection of media coverage on each single feature.[2] I find that half of these features alone cannot produce any significant return difference, while the most successful one only obtains a 6.72% annualized alpha. This is about half the size of the alpha based on all features. I also apply the elastic regularization methodology of Fama and French (2020) to show that editor preferences explain the return variation between the 500 largest firms better than any single feature. Next, I show that time-varying selection by editors is crucial for my results. Based on the time series average of editor preferences, I show that the return difference between high and low portfolios shrinks largely and no longer exists without knowing which firms are covered in the news. These tests reflect that my results are caused by two combined facts. First, news editors monitor different aspects of firms. Second, news editors adjust their focus dynamically over time.

Lastly, I inspect the information quality of editorial selection. I test if editor preference is limited to the 500 largest firms that I choose every month, and cannot produce similar results among more other firms. Using the coefficients I estimated for the 500 largest firms, I calculate editor preferences for the largest 1000, 3000, and 5000 firms in the Compustat/CRSP by market value. I show that the results are very close to the baseline, especially among non-covered firms. This comparison suggests that, when making general coverage decisions, news editors judge on the basis of more comprehensive concerns about common risks, rather than any specific firms. Otherwise, my estimation for editor preferences will be sensitive to any outliers in my firm samples, and thus have limited return prediction power for a broader group of firms. Still, I find that the news editors in my data are myopic. If I hold the long-short portfolio open for 12 months, the annualized alpha will decrease to 5.36%. This suggests that editor preferences capture short-term risks. This myopia could be caused by the capacity of editors. But It could also be the case that investors who subscribe to newspapers mostly expect to learn about short-term risks. In order to fulfill that demand, editors need to provide myopic information accordingly (see Mullainathan and Shleifer (2005)). In an equilibrium between editors and investors, however, these two explanations are not distinguishable.

The rest of this paper is organized as follows. Section 2 introduces my data and methodology. Section 3 presents my empirical results. Section 4 present the robustness checks. Section 5 concludes.

[2] Here, the editors play an active and state-dependent signaling role for individual risks.

1.1 Related literature

This paper contributes to the literature on the relation between media coverage and asset prices. One strand of this literature studies how news sentiment affects the market. For example, Mullainathan and Shleifer (2005), Solomon and Soltes (2012), Niessner and So (2018) and others document that news media tends to cover negative news about firms. Tetlock (2007) and Tetlock et al. (2008) show that the media pessimism predicts downward stock returns in the short-term. Garcia (2013) verifies that this predictability concentrates in recessions[3]. Ahern and Sosyura (2015) also document that investors cannot identify merger rumors, which lead to short-term mispricings.

Another strand of the literature focuses on the interaction between attention-limited investors and media coverage, and how this affects investment decisions. For example, Engelberg and Parsons (2011) find that local media coverage for earnings announcements of S&P 500 Index firms strongly predicts the level and the timing of local trading. Barber and Odean (2008) and Peress (2014) find that investors tend to buy stocks in the news and cause buying pressure. Ben-Rephael et al. (2017) construct an index for abnormal institutional investor attention using news searching and news reading activity on Bloomberg terminals, and find that it correlates with the speed of price adjustments.

Although the literature agrees on the relationship between media coverage and trading behavior, when it comes to the cross-section of asset prices there are two seemly opposite conclusions. The first type of argument use media coverage as a proxy for investor recognition since the intensity of news coverage influences the cost of information acquisition (see Carroll (2003), for example). In an incomplete market modeled by Merton et al. (1987), investors only have information access to a limited number of securities, such that there exists an imperfect diversification in their investment allocations. As a consequence, securities that have lower investor recognition need to offer higher risk compensation. Fang and Peress (2009) document that stocks with no media coverage earn higher returns than stocks with high media coverage even after controlling for common risk factors. This return difference is particularly large among firms with more information friction such as firms with a high fraction of individual ownership. Gao et al. (2020) finds similar evidence in the bond market: media coverage is negatively correlated with firms' cost of debt and is robust to controlling for standard yield determinants. On the other hand, based on typical attention stories, Hillert et al. (2014) show that momentum is stronger

[3] In Section 4 I show that my results are not related to similar sentiment effect.

among covered firms than among non-covered firms, and is monotone in the degree of coverage. Hillert and Ungeheuer (2016) find that highly covered firms outperform less covered firm in the short run, which directly contradicts the recognition story. The authors argue that the different conclusions are caused by a longer holding period and the way to control for the size effect.

Although this paper is inspired by Nimark and Pitschner (2019) and Chahrour et al. (2019), which introduce a new editor-based mechanism, the empirical setting is close to Fang and Peress (2009), Hillert et al. (2014) and Hillert and Ungeheuer (2016). Nevertheless, my paper differs from them in three major ways. First and most importantly, the editor preference measure I construct is orthogonal to the definition of media coverage in their papers. In any given month, we all run regressions of media coverage on firm characteristics. While I use the projection part (except for the part that is related to market value) as my measure for editor preference, they use the residual part to measure characteristic-free coverage or "coverage shock". Theoretically, these two measures should have little correlation, which means a highly covered firm in their settings can be a firm with low editor preference in my definition.[4] Second, in Section 3.2, I show that the return difference I observe cannot be explained by any characteristic alone, and hence is beyond the exaggerated momentum effect discovered by Hillert et al. (2014). At last, editor preference can also be estimated for firms that are not covered in the news, and related return patterns are also discovered among non-covered firms. However, in their settings, there is no way to measure either investor attention or editor preference among non-covered firms, and consequently they are treated as the same.

My paper also contribute to a growing literature on the prediction power of news selection by editors on the economy. Blinder and Krueger (2004) and Curtin (2007) document that households mainly rely on either TV news shows or newspapers to get information of the economy. Therefore, the news has incentives to monitor the state of the economy for households. Using an LDA model and LASSO regression, Larsen et al. (2020) show that news topics are good predictors of both inflation and inflation expectations. With a more comprehensive topic modeling, Cong et al. (2019) and Bybee et al. (2020) show that news topics have incremental forecasting power for several macroeconomic outcomes, above and beyond standard numerical predictors. These discoveries altogether support the argument that news editors make state-dependent coverage decisions, and we can better understand the economy by studying these decisions. My paper

[4] In an unreported test, I also sort on the residual of my regression and find evidence consistent with Fang and Peress (2009).

analyzes such prediction power for stock returns.

2 Data & empirical setup

2.1 News data

I obtain daily news for the time between January 1, 1995, and December 30, 2015, through The New York Times' API (`https://api.nytimes.com`). I only select articles from the "Business" and "Business Day" sections and filter out any company announcements. I am left with 140,227 articles in total. Panel (a) of Table 1 provides summary statistics of the news articles and Panel (b) displays a sample article.

For every article in my data, I apply the company identification methodology introduced in Schwenkler and Zheng (2019) to recognize any mentions of firms. This methodology has two basic steps: Firstly, I use Natural Language Processing (NLP) toolkit provided by Manning et al. (2014) and Arnold (2017) called *coreNLP* to identify any named entity appearing in my articles (I use the Named Entity Recognition algorithm, or NER). These entities will be tagged as different types, such as persons, locations, organizations, etc. For those entities that are marked as organizations, I follow a series of rules to clean and match them with the Compustat/CRSP database over the same period. The approach I use has been proven to have more than 87% matching accuracy.

Panel (a) of Table 1 gives some basic statistics of my recognition results. Notice that, in most of the cases, one company can be mentioned multiple times within an article. I believe that the number of mentions contains useful information. If two companies are both mentioned in the same article while the former one shows up more times, then in my analysis it will have a higher media coverage than the latter one instead of being equal. This difference separates my paper from former text-based asset pricing studies such as Hillert et al. (2014), Scherbina and Schlusche (2016), and Chahrour et al. (2019) to some extent. In those studies, each news article is marked with several company tags by the databases they use, meaning the article is mainly talking about these companies. They count the number of tags for each company as their media coverage. One can imagine that the outcomes of my method should be similar to such tag-counting method: the primary tagged firm usually will be mentioned more times in the article. The advantage of my counting method is, however, that I do not rely on these post processed tagging service, which may not be provided by all news databases. In Section 4.2, I show that these two measures for media coverage do not change the main results of this paper.

Also, notice that the total mentions of each firm in my data have heavy tailed distributions. To control for this issue, I define the measure

$$coverage_{i,t} = ln(1 + mention_{i,t}),$$

where i is the firm index and t stands for the counting period. Here, $mention_{i,t}$ is the number of mentions of firm i in the news articles in month t. I will heavily rely on this measure in the rest of this paper.

2.2 Prior coverage preference

In this paper, I study the state-dependent preferences of editors. I am not only interested in which companies are covered by news, but also in which are not. Ideally, I want to focus on a group of firms sharing similar prior probabilities of being covered. Then, if some of them (rather than others) are considered to be newsworthy by editors, it suggests that these firms have some features that need to be paid attention to in the current state of the world. Although it is almost impossible to estimate such group in practice, we can instead look for what common features of firms can constantly and consistently affect media coverage in any state. As pointed out in Fang and Peress (2009), Engelberg and Parsons (2011), Solomon and Soltes (2012), and Hillert et al. (2014), size (market value) is the most dominant feature of firms that determines prior media coverage on average: larger firms are more likely to get covered by any news outlet and in any situation. Therefore, I only consider the 500 largest firms by market value in the Compustat/CRSP universe. Since all of these firms are large, we can expect them to have similar chance of media coverage if everything else is equal. I therefore focus on editors' preferences after conditioning on size. There are two more reasons of doing this. First, the number of covered firms versus the non-covered ones is more balanced among largest 500 firms. Second, picking 500 rather than fewer firms yields a wider industry coverage (see Figure 2).

For every month t in my sample period, I use market value at the end of the previous month $t-1$ to select the 500 largest firms. This leaves me with 1,426 unique firms over the whole period; see Panel (a) of Table 2 for summary statistics. After I determined the top 500 group, I divide the group into two sub-groups: the covered group (all the firms that are reported at least once in month t) and the non-covered group (the rest of the top 500). In every month, roughly 20 to 40% of the largest 500 firms are in the covered group.[5] See Figure 1 for details.

[5] However, there is one outlier on February 2015, in which I was not able to download enough business news

There are 665 out of 1,426 firms that are never covered by my news data over the sample period. This may be expected because, even within the top 500 group, there is significant size dispersion and the news still prefers to report about larger firms. The characteristics comparison between the covered and non-covered groups can be found in Panel (b) and (c) of Table 2, and Figure 2 reports differences in their industry distributions.[6]. I conclude that indeed the covered firms on average are larger than the non-covered ones no matter what measures I use (market value, asset size, sales, etc.). Moreover, the covered group has a larger portion of firms operate internationally, which is also natural for big firms. Lastly, when compared to either the top 500 group or the CRSP/Compustat universe, the covered group tends to have more firms in both consumer staples and health care sectors, while less firms in information technology industry.

2.3 Editor preference and portfolio construction

I now estimate the state-dependent preference of editors. More plainly, I estimate what features of a firm (except for size) can cause a higher media coverage in different periods. Following the standard procedure as in Fang and Peress (2009), Hillert et al. (2014), and Hillert and Ungeheuer (2016), in every month t I run the following cross-sectional regression among the 500 largest firms:

$$coverage_{i,t} = \alpha_t + \gamma_t * ln(size_{i,t}) + \sum_k \beta_{k,t} * feature_{k,i,t} + \text{fixed effect}_i + \epsilon_{i,t} \tag{1}$$

where $size$ stands for market value and $feature$ includes other common features of firms that can affect media coverage. Here, fixed effects control for the GIC industry sector and also for whether a firm i is ever covered in the most recent 3 months. The reason I want to take those into consideration is that industry sector, along with other omitted firm features such as headquarter address and rating, can cause prior non-coverage in the recent period and hence the conditional non-coverage in month t. This potential endogeneity issue is also the reason why I run this regression using both covered and non-covered firms (which have zeros at the left hand side) instead of using only covered firms. Although the Hausman test is not applicable here, Figure 3 demonstrates the sample correlation between $\sum_k \beta_{k,t} * feature_{k,i,t}$ and $\epsilon_{i,t}$ when using different

from the New York Times archive.

[6] Notice that here the group definition is a little different: for such a long period, every top 500 firm has a fair chance to be covered at least once. Therefore in firm's characteristic comparisons I calculate how many times a firm is included in the covered group and compare the top 30% (6 times, "most covered") versus the firms that were never covered over the sample period.

firm pools. Including all 500 firms and fixed effect definitely mitigate the issue. Besides, I also want to eliminate any industry effect on media coverage.

In $feature_{k,i,t}$, I consider eight different variables that are correlated with media coverage: cumulative return in past three months ($CumRet$) and its absolute value ($|CumRet|$), idiosyncratic volatility ($IVOL$), book-to-market ratio (B/M), and firms' factor loadings in the Carhart (1997) 4-factor model.[7]

I verify the correlation between $size$ and $coverage$. Figure 4 reports the t-statistic of γ in my regression (1). One can see that indeed $size$ positively, persistently, and significantly affects the media coverage of a firm, even after controlling for other features and fixed effects. This observation justifies my narrowing down to the largest firms and not considering $size$ as a feature reflecting the editors' state-dependent preferences. On the other hand, $feature_{k,i,t}$ tell a different story. Figure 5 shows the t-statistics of their coefficients, β_k. The coefficients are not always significant over the sample period, implying that none of them are decisive factors that drive the media coverage. Moreover, the sign of the coefficients switches through different periods, which provides more supportive evidence for the argument that the co-movement between these features and media coverage is changing over time. Therefore, I consider this time-varying part $\sum_k \beta_{k,t} * feature_{k,i,t}$ to reflect the editors' preference for features of a firm that they believe investors should pay attention to at time t. I denote editors' preferences as EP and define this measure as follows:

$$EP_{i,t} = \sum_k \beta_{k,t} * feature_{k,i,t}$$

At the end of month t, I construct 4 equal-weighted portfolios and rebalance them one month later. I construct the high EP portfolio using the 30% percentile among the covered firms. This portfolio has the largest $EP_{i,t}$ measurement. I also construct the analogous low EP portfolio, as well as equivalent high and low EP portfolios among non-covered firms. This is the key difference between my methodology and the existing literature, which uses $\epsilon_{i,t}$ in Eq. (1) to construct the high and low portfolios. Since the dependent variables and the residual provide orthogonal information, the return patterns of my portfolios should be caused by different factors.

[7] I use a 24-month rolling window to estimate each firm's factor loadings and use the residual as $IVOL$. A longer rolling window has been tested, but it has negligible impact on the results.

3 Empirical results

3.1 Post-news: baseline performance

In the equilibrium of the models of Nimark and Pitschner (2019) and Chahrour et al. (2019), if risk-averse investors outsource their information choice to news editors, then news editors view more risky firms as more newsworthy. Based on these models, we expect high EP firms to earn a higher expected return than low EP firms. Moreover, this return difference should exist both among the covered firms and the non-covered firms. This is because EP captures editors' focus on firm features rather than idiosyncratic characteristics, and even non-covered firms are differently exposed to the risky features that editors are concerned with. This will be the major difference between media coverage and editor preference in the setting, since one cannot measure the former one among non-covered firms.

Figure 6 graphically demonstrate the corresponding return patterns. Some interesting observations can be made. The covered firms with high EP have the highest overall returns over my sample period and are followed by the non-covered firms with high EP. Meanwhile the low EP portfolios earn pronouncedly lower returns. These observations are in line with my conjecture that newsworthy firms are the most risky ones for investors. The fact that such return difference is more narrow among non-covered firms arguably suggests that there are some other risky features causing media coverage, which I did not consider in my regression when I estimated editor preferences.

One may argue that my analysis is misleading since EP contains features that are proven to carry risk premia. For instance, when the book-to-market ratio has a more dominant and positive impact on media coverage than other features, then sorting on EP is basically sorting on book-to-market ratio. Consequently, the return difference between high EP firms and low EP firms will mainly be caused by the value factor. In order to have a closer scrutiny, I use the 4-factor model in Carhart (1997) plus the betting-against-beta (BAB) factor in Frazzini and Pedersen (2014) to decompose the results as shown in Table 3. Even after controlling for these common factors, one can still observe significant return differences that are associated with EP and are consistent with the patterns in Figure 6. Notice that a long-short portfolio that longs the covered firms with high EP and shorts the non-covered ones with low EP has a sizable 12% annualized alpha. The factor model test provides substantial evidence for the argument that news editors are guiding investors attention to a basket of risky features, such that firms with larger exposure to these features requires higher return compensation by investors.

3.2 Sorting decomposition

At first glance, the baseline results (or, more specifically, the alphas) in Table 3 are surprising from a portfolio construction point of view. Why does sorting on EP give a much different return pattern than sorting on the features that compose EP? Nevertheless, two nuances can be found from the way I construct EP to explain this return pattern.

First, the loadings of *coverage* on each feature are state-dependent. More specifically, both the level and the sign of the coefficients are time-varying (see Figure 5). This implies that the EP-based sorting changes over time. For instance, in some months the loading of *coverage* on book-to-market ratio can be negative, hence a high EP firm is more likely to be a growth firm. However, we know that the value factor return is always negatively correlated with the return of growth firms. Therefore, compared to the traditional static sorting, sorting on EP is more similar to a factor timing strategy. The reason such timing is generally successful is that news editors perform well in closely tracking the current state of the market on behalf of investors. This timely reflection of the focus of news editors is also in line with topic shifting observed by Nimark and Pitschner (2019) and Bybee et al. (2020), while the latter one also shows that news editors keep a good record of predicting the macro environment.

To more formally test the importance of the timing dimension for my results, I construct a static measure of editor preference using the time series average of β_k in Regression (1):

$$\text{Static EP}_i = \sum_k \bar{\beta}_k * feature_{k,i,t}.$$

I then sort on this new static EP every month and construct corresponding portfolios. The factor model results can be found in Table 4. In Panel (a) we can see that, among non-covered firms, the static EP measure cannot contribute additional pricing information at all. Only covered firms with high static EP contain alpha, but it is less pronounced. This is consistent with the fact that being covered itself also reflects the timely focus of news editors. Panel (b) shows a clearer evidence: high static EP firms no longer hold any alpha against low static EP firms without knowing which firms are in the news. All combined, news editors' time-varying preferences indeed play a defining role in generating the excess returns observed in Table 3.

Second, news editors not only monitor one specific feature but a combination of several features. The dispersion in the coefficients β_k reflects the different importance of these features at that moment. Not only does the sign of β_k matter, its level also decides whether a firm of high $feature_k$ belongs to the high EP group or not. I show that one single feature cannot contribute enough pricing information. Suppose we only consider one feature that composes editor preference. Then, to construct the high/low EP portfolios, I only need to sort on the projection of $coverage_i$ on this feature, i.e. $\beta_{k,t} * feature_{k,i,t}$ for firm i in month t. Table 5 demonstrates the results. The alphas either no longer exist or take a large discount when only considering one specific feature of editors' preferences. To further emphasize the explanation power of editor preference on the variation of returns among my sample firms, I apply a methodology for estimating factor returns introduced in Fama and French (2020). In their paper, they run a cross-sectional regression at time t:

$$R_{i,t} - R_{z,t} = \beta_{1,t} * C_{1,i,t-1} + \beta_{2,t} * C_{2,i,t-1} + \cdots + \epsilon_{i,t},$$

where R_i, t is the stock return of firm i in month t, $R_{z,t}$ is the benchmark return, $C_{i,t-1}$s are the characteristics of firms observed previously and such that β_t's are the realized factor return related to that characteristic at time t. Notice that, here, the characteristics are the independent variables and the factor returns are the loadings. Once we have a time series of realized factor return $\beta_{k,t}$, we can test whether its time-series average is significantly different from zero or whether it really contains pricing information. The intuition behind this test is to see whether the variation of a feature k among I firms can explain the return variation among these firms, even after controlling for other features. Following the same idea, I design a similar test to verify the capability of editor preference in explaining the return differences among 500 largest firms. In every month of my sample period, I run the following regression:

$$R_{i,t} - R_{z,t} = \gamma_t * ln(size_{i,t-1}) + \sum_k \beta_{k,t} * feature_{k,i,t-1} + \alpha_t * EP_{i,t-1} + \epsilon_{i,t} \tag{2}$$

where $feature_{k,i,t-1}$ are the same as in Regression (1), $R_{i,t}$ is the monthly return of Firm i in the top 500 group, and $R_{z,t}$ is the average monthly return of the top 500 group.[8] It is obvious that there exists co-linearity between EP and $feature_k$. Therefore, I use the elastic net regularization which will push the coefficients of variables that have limited explanation power toward zero. Once we obtain the time series of γ_t, $\beta_{k,t}$ and α_t, t-tests can be applied to verify their significance. Table 6 gives the results. The only significant coefficients here are the ones

[8] Notice that, here, I do not use risk-free rate or other orthogonal portfolio returns. This is because the focus of this test is not to estimate factor returns.

related to B/M, $IVOL$ and editor preference (EP). Moreover, the average coefficient of EP is at least ten times the average coefficients of other variables. All combined, these results show that EP explains the cross-sectional return differences among the 500 largest firms better than any specific feature that it consists of.

3.3 Comprehensive but myopic editors

It could be counterintuitive to some that news editors have such comprehensive reporting preference on firms' characteristics, and that they not just chase specific firms and their events. One evidence I have shown to support this fact is that the EP-related return difference also exists among non-covered firms. To better test how universal the EP premium is, I expand the size of my sample to see if the premium still exists among more non-covered firms. More specifically, I estimate in every month the coefficients of different features in Regression (1) using the 500 largest firms. Then, I calculate EP for each of the N largest firms. Finally, based on their EP values and their coverage status, I construct equal-weighted "covered/non-covered" and "high EP/low EP" portfolios as usual. In Table 7, it is very interesting to see that the EP-related alphas persist to exist among more firms, especially considering the fact that the coefficients used to construct EP are estimated using only 500 firms. Besides, in both the covered and the non-covered group, the alphas observed are comparable with my previous results in the senses of size and of statistical significance. That is, constraining myself to the 500 largest firms every month is good enough to evaluate editorial preferences.

Nevertheless, news editors seem to be myopic. From Panel (a) in Table 8 we can see that the EP-related alphas die out quickly after 1 month. For instance, if I rebalance the portfolios monthly (baseline), then the largest annualized alpha I obtained is around $1\% * 12 = 12\%$. However, if I hold the same portfolios for a longer period, the annualized alpha decreases to 5.36%. This alpha discount suggests that editors are either more concerned about short-term risks rather than long-term ones, or they are not capable of capturing long-term trends. It does not help much if I expand my news formation period from one month to a longer period. More past news also bring in more noisy information and impairs even the short-term return predictability. This myopia should be expected due to the objective of daily news reporting: to report events in a timely fashion. Thus, its readers will expect the editors to be more sensitive to short-term risks such that in turn the editors want to fulfill that expectation. In the future research in which other forms of information medium (for instance journals) are used, we may observe editor preference over a longer time horizon.

Myopia also partially explains why the EP-related alpha is not exploited by active fund managers. The trading strategy that longs high EP covered firms and shorts low EP non-covered firms only achieves an annualized Sharpe ratio of 0.89, implying that the return volatility is too high and hence less attractive to investors than it appears.[9] Besides, I observe a common phenomenon in both panels of Figure 6: the return gap between high EP firms and low EP firms shrinks largely when the market is in distress.[10] Myopia can provide an explanation. In a distressed market, news editors will focus more on the current state of the economy rather than providing forward-looking opinions. In that case, editor preference will have very limited predictive powers but only reflect a current situation, especially for those firms with high EP. In Figure 7, for example, the covered and high EP portfolio experience large drawdowns during the two bear markets. Once we realize that the news does not always provide useful forward-looking information, the trading strategies can be adjusted accordingly. For instance, a strategy that constantly longs high EP covered firms and shorts low EP non-covered firms, if I flip the long-short positions after I observed two negative monthly market return in a row, the annualized alpha will increase from 12% to 16.44% with a higher Sharpe ratio of 1.33.[11] From this rough application one can see how important it is to distinguish the content of news. In the future research, more sophisticated NLP tools, like the topic modeling in Bybee et al. (2020), can be applied to address this issue.

4 Robustness check

4.1 Tone of news

In this section, I test if my previous results are caused by some sentiment effects on investors behavior (for example, see Shiller (2015), Tetlock (2007) and Tetlock et al. (2008)). The *coreNLP* toolkit I applied to identify entities also provides sentiment for each sentence in my news data. Based on a pre-trained database and the structure of target sentences, the toolkit will give an integer sentiment score for each sentence ranging from 0 to 4. 0 stands for very negative and

[9] Although compared to the Sharpe ratios of other media-coverage-based trading strategies, such as 0.63 in Hillert and Ungeheuer (2016), it is relatively high.

[10] There are two major bear markets in my data: the dot-com crash (2000-2002) and the financial crisis (2007-2009).

[11] It happened 46 times in my sample period of 252 months.

4 stands for very positive. There are in total 10,145,101 sentences in my data with an average sentiment score of 1.3, therefore it would be reasonable to set 1 as my benchmark neutral score. 819,109 (8.07%) of these sentences mention at least one of the 500 largest firms in that month. Among these 819,109 sentences, 82.72% of them have sentiment scores of 1 and 5.42% of them have scores of 0. One can already observe that most of the media coverage of firms is not tilted toward pessimism. Next, I only select those sentences that have scores of 1 to calculate *coverage* for each firm, estimate EP, and construct the corresponding portfolios as before. Panel (a) in Table 9 reports the regression results when only "neutral news" are used. Comparing to Table 3, we see that there is not much difference in the sense of factor loadings. The effect on alphas is minor.

Garcia (2013) documents that the prediction power of news sentiment for stock returns is concentrated in recessions. Following this idea, in Panel (b) I also include the NBER recession indicator in the regressions. Although the alphas shrink slightly, the loadings on the recession indicator are never significant. I conclude that the return differences I document are not caused by the sentiment of my news data.

4.2 Editor preference estimation

In this section, I test the robustness of my results to the way I estimate EP. Firstly, I show that my results do not rely on the definition of *coverage*. As I introduce in Section 2.1, one of the most popular methods to measure news focus of firm i is to count how many news articles are classified under the name of firm i using news tags. However news tagging requires editorial classification (usually manually) and not every news database provides such a service. To approximate this tag-based measure, if a firm is mentioned more than 2 times in a given article, I add tag "Firm i" to this article.[12] For each firm i, I count how many news articles during month t contain its tag and use the logarithm of this number as *coverage*. After that, I estimate EP_i and construct portfolios as usual. Panel (a) in Table 10 reports the regression results using this different definition of *coverage*. Again, both the alphas and the factor loadings are similar in terms of scale and significance level. This comparison implies that the results I document in this paper are robust to different measurements of media coverage measurement, and thus this discovery supplements previous results, such as those of Fang and Peress (2009)

[12] I choose this threshold because from Table 1 we know that the average mentions of a mentioned firm in an article is approximately 2.33.

and Hillert and Ungeheuer (2016).

Next, I use the ridge regularization to mitigate the effect of residuals in Regression 1. The reason I choose ridge over elastic net or lasso is that I want to keep as many $feature_k$ as possible in the regression, while having a correct order of the importance for these features. Since $size$ and fixed effect combined have dominant impact on media coverage, it is easy for the regularization to push the coefficients of any other variables to zero. Then most of the time, the corresponding editor preference will have very limited stock-picking capability. Panel (b) in Table 10 shows the results with ridge regularization and they are again very similar to the baseline model. In unreported tests, I also construct EP under elastic regularization with equal weighted ℓ^1-norm penalty (lasso) and ℓ^2-norm penalty (ridge). The alphas are only about one third of their original sizes, and increase as the weight on ℓ^2-norm penalty increases. In conclusion, similar to what I observe in Section 3.2, using all $feature$s to construct EP instead of one or two of them is essential to the results.

Finally, I test if the results hold when fitting an non-linear model. I run zero-inflated negative binomial regressions to model the frequency (count number) of media coverage and the probability of being covered at the same time:

$$E[mention_i|X_i, Z_i] = (1-\pi_i)\mu_i \quad \text{and} \quad Var[mention_i|X_i, Z_i] = (1-\pi_i)(\mu_i + \mu_i^2/\theta)$$

where $\pi_i = \frac{e^{b'X_i}}{1+e^{b'X}}$, $\mu_i = e^{b'_Z Z_i}$. In X_i I include all the variables on the RHS of Regression 1 while in Z_i I exclude fixed effects. Then analogously, I use the projection on the eight features in X_i as my estimation for editor preference. Panel (c) in Table 10 shows the return difference between portfolios sorting on this differently estimated EP. The alphas are very significant and are only slightly smaller than the baseline ones. This nuance is expected since compared to my original linear model, EP here only accounts for the correlation between firm features and number of mentions conditioning on being covered already.

5 Conclusion

Using articles from the New York Times over 20 years, I show that editors have state-dependent preference for different types of firms. Moreover, I construct a variable EP to measure the editor preference for common characteristics of firms. After that, I show that firms with high EP earn higher returns than firms with low EP, no matter if I consider news-covered firms or non-covered ones. By constructing corresponding long-short portfolios, I exploit sizable and significant alphas that cannot be explained by mainstream risk factors.

Our findings validate previous theories that argue that the news provide a larger service to their readers than just reporting about current events. The motivation behind editors' choices to cover some events over the others is also of importance. In the equilibrium between news editors and rational investors who delegate their information collection to them, editors help investors in monitoring what kind of firms are facing more risks. In this paper, however, I do not decompose the editor preferences into finer parts: whether it indeed reflects editors' opinions on the state of the economy, or it simply reflects a motivation to attract investors irrational attention. Although I have shown some positive evidence for the former argument, I will leave a more detailed discussion of this topic for future research.

To obtain my results, I use a natural language processing toolkit to identify companies from news. My research lies among a new genre of machine learning applications in asset pricing that considers alternative data that are difficult to analyze with traditional quantitative methods.

References

Ahern, Kenneth R and Denis Sosyura (2015), 'Rumor has it: Sensationalism in financial media', *The Review of Financial Studies* **28**(7), 2050–2093.

Arnold, Taylor (2017), 'A tidy data model for natural language processing using cleannlp', *The R Journal* **9**(2), 1–20.
URL: *https://journal.r-project.org/archive/2017/RJ-2017-035/index.html*

Barber, Brad M and Terrance Odean (2008), 'All that glitters: The effect of attention and news on the buying behavior of individual and institutional investors', *The review of financial studies* **21**(2), 785–818.

Ben-Rephael, Azi, Zhi Da and Ryan D Israelsen (2017), 'It depends on where you search: Institutional investor attention and underreaction to news', *The Review of Financial Studies* **30**(9), 3009–3047.

Blinder, Alan S and Alan B Krueger (2004), What does the public know about economic policy, and how does it know it?, Technical report, National Bureau of Economic Research.

Bybee, Leland, Bryan T Kelly, Asaf Manela and Dacheng Xiu (2020), The structure of economic news, Technical report, National Bureau of Economic Research.

Carhart, Mark M (1997), 'On persistence in mutual fund performance', *The Journal of finance* **52**(1), 57–82.

Carroll, Christopher D (2003), 'Macroeconomic expectations of households and professional forecasters', *the Quarterly Journal of economics* **118**(1), 269–298.

Chahrour, Ryan, Kristoffer Nimark and Stefan Pitschner (2019), 'Sectoral media focus and aggregate fluctuations', *Available at SSRN 3477432* .

Cong, Lin William, Tengyuan Liang and Xiao Zhang (2019), 'Textual factors: A scalable, interpretable, and data-driven approach to analyzing unstructured information', *Interpretable, and Data-driven Approach to Analyzing Unstructured Information (September 1, 2019)* .

Curtin, Richard (2007), 'What us consumers know about economic conditions'.

Engelberg, Joseph E and Christopher A Parsons (2011), 'The causal impact of media in financial markets', *The Journal of Finance* **66**(1), 67–97.

Fama, Eugene F and Kenneth R French (1993), 'Common risk factors in the returns on stocks and bonds', *Journal of Financial Economics* .

Fama, Eugene F and Kenneth R French (2020), 'Comparing cross-section and time-series factor models', *The Review of Financial Studies* **33**(5), 1891–1926.

Fang, Lily and Joel Peress (2009), 'Media coverage and the cross-section of stock returns', *The Journal of Finance* **64**(5), 2023–2052.

Frazzini, Andrea and Lasse Heje Pedersen (2014), 'Betting against beta', *Journal of Financial Economics* **111**(1), 1–25.

Gao, Haoyu, Junbo Wang, Yanchu Wang, Chunchi Wu and Xi Dong (2020), 'Media coverage and the cost of debt', *Journal of Financial and Quantitative Analysis* **55**(2), 429–471.

Garcia, Diego (2013), 'Sentiment during recessions', *The Journal of Finance* **68**(3), 1267–1300.

Hillert, Alexander, Heiko Jacobs and Sebastian Müller (2014), 'Media makes momentum', *The Review of Financial Studies* **27**(12), 3467–3501.

Hillert, Alexander and Michael Ungeheuer (2016), 'Ninety years of media coverage and the cross-section of stock returns', *University of Mannheim, working paper* .

Kahneman, Daniel (1973), *Attention and effort*, Vol. 1063, Citeseer.

Larsen, Vegard H, Leif Anders Thorsrud and Julia Zhulanova (2020), 'News-driven inflation expectations and information rigidities', *Journal of Monetary Economics* .

Manning, Christopher D, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard and David McClosky (2014), The stanford corenlp natural language processing toolkit, *in* 'Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations', pp. 55–60.

Merton, Robert C et al. (1987), 'A simple model of capital market equilibrium with incomplete information'.

Mullainathan, Sendhil and Andrei Shleifer (2005), 'The market for news', *American Economic Review* **95**(4), 1031–1053.

Newey, Whitney K and Kenneth D West (1987), 'A simple, positive semi-definite, heteroskedasticity and autocorrelation', *Econometrica* **55**(3), 703–708.

Niessner, Marina and Eric C So (2018), 'Bad news bearers: The negative tilt of the financial press', *Available at SSRN 3219831* .

Nimark, Kristoffer P and Stefan Pitschner (2019), 'News media and delegated information choice', *Journal of Economic Theory* **181**, 160–196.

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

Scherbina, Anna and Bernd Schlusche (2016), 'Economic linkages inferred from news stories and the predictability of stock returns', *AEI Paper & Studies* p. Bi.

Schwenkler, Gustavo and Hannan Zheng (2019), 'The network of firms implied by the news', *Available at SSRN 3320859* .

Shiller, Robert J (2015), *Irrational exuberance: Revised and expanded third edition*, Princeton university press.

Solomon, David H and Eugene F Soltes (2012), 'Managerial control of business press coverage', *Available at SSRN 1918138* .

Tetlock, Paul C (2007), 'Giving content to investor sentiment: The role of media in the stock market', *The Journal of finance* **62**(3), 1139–1168.

Tetlock, Paul C, Maytal Saar-Tsechansky and Sofus Macskassy (2008), 'More than words: Quantifying language to measure firms' fundamentals', *The Journal of Finance* **63**(3), 1437–1467.

Panel (a)

Variables	Mean	Std Dev.	Min.	Max.
Number of articles per month	558	139	327	980
Number of articles per year	6677	1417	4782	9845
Number of unique firms per article	3	2	1	62
Number of mentions per article	7	8	1	167
Total mentions of each firm	344	1510	5	33440

Panel (b)

I.B.M. and Partner May Offer Broadband from a Wall Plug
2005-07-11T00:00:00Z
By Ken Belson
I.B.M. will announce a partnership today with CenterPoint Energy, a utility based in Houston, to develop broadband services to be delivered over electric power lines. The companies will open a technology center in Houston to test and demonstrate the technology for consumers and other utility providers. CenterPoint Energy will also set up a pilot program in about 220 Houston homes that will run through August. Because power lines can carry data as well as electricity, utilities and broadband companies are hoping the technology will allow consumers to get high-speed Internet connections simply by plugging a special adapter into a wall outlet. Some utilities, including Con Edison in New York, have started offering such services on a limited basis. By relying on the adapters – which currently cost about $200 but are expected to become less expensive – utilities do not need to send a worker to install equipment. Consumers can use the adapters in any room with an outlet. The Federal Communications Commission is backing the development of this technology in hopes of creating a counterweight to the cable and phone industries, which provide the bulk of the 36 million broadband lines now being used in American homes. The service could also be cheaply deployed in rural areas where phone and cable companies have not yet expanded. CenterPoint says it will be one of the first utilities to test new technology, including faster chips that roughly triple connection speeds. With these chips, consumers will be able to receive Internet connections at about 7 megabits a second, equal to some of the fastest speeds available from cable companies. Utilities are interested in offering broadband services, not only because it could help them generate new revenue but also because it would allow them to read meters remotely, pinpoint problems throughout their network and monitor power surges as they take place rather than long afterward. Utilities say they could save millions of dollars if they could avoid long power failures and if they did not have to send workers to read meters. "People don't understand how little the utilities can see of their network," said Ray Blair, vice president for broadband over power lines at I.B.M., which is advising CenterPoint on the project. "If your power goes out, they don't know about it until you call. This will tell them exactly where to go and what to fix."

Table 1: Panel (a) contains summary statistics of news articles in my data set and are rounded to be integers. I download 140,227 news articles between January 1, 1995, and December 30, 2015, from the "Business" and "Business Day" sections of The New York Times. In the last row, I only consider firms that are mentioned at least once in my data. Panel (b) provides a sample article in my data.

Panel (a) Top 500					
Variables	Mean	Median	Std Dev.	Min.	Max.
Market value (million USD)	16242.07	7052.98	31816.73	9.86	345013.39
Total asset (million USD)	3307.55	1157.91	7076.50	16.17	103359.08
Total debt (million USD)	6427.03	1724.90	25356.76	0.00	538837.27
Book leverage (long-term debt/total asset)	122.35%	60.41%	232.05%	0.00%	4832.04%
Cash holding (million USD)	1507.33	259.73	7840.15	0.00	224841.63
Net income (million USD)	670.12	218.80	1409.15	0.02	19173.23
Sales (million USD)	8482.00	2926.15	18837.09	7.86	361489.00
Number of firms that operate internationally	1013 (71.04%)				
Number of firms whose stocks are traded in U.S. exchanges	1295 (90.81%)				

Panel (b) Top 500 - most covered					
Variables	Mean	Median	Std Dev.	Min.	Max.
Market value (million USD)	34917.12	18020.97	50581.42	299.13	345013.39
Total asset (million USD)	6727.04	2792.57	10860.16	40.93	103389.08
Total debt (million USD)	11925.00	4736.46	27114.77	0.00	265261.26
Book leverage (long-term debt/total asset)	119.00%	57.00%	310.00%	0.00%	4832.00%
Cash holding (million USD)	3137.10	658.34	10095.92	0.00	155513.40
Net income (million USD)	1295.18	502.52	2098.38	1.67	19173.23
Sales (million USD)	17140.11	7962.21	28023.60	198.81	361489.00
Number of firms that operate internationally	233 (52.13%)				
Number of firms whose stocks are traded in U.S. exchanges	409 (91.50%)				

Panel (c) Top 500 - non-covered					
Variables	Mean	Median	Std Dev.	Min.	Max.
Market value (million USD)	6592.87	4752.02	6983.87	9.86	62911.71
Total assset (million USD)	1579.18	728.14	3303.56	17.80	36373.00
Total debt (million USD)	3118.36	940.18	19299.86	0.00	460952.00
Book leverage (long-term debt/total asset)	123.00%	63.00%	167.00%	0.00%	1482.00%
Cash holding (million USD)	754.39	154.20	7755.56	0.00	224841.63
Net income (million USD)	294.90	143.79	599.47	0.92	7515.06
Sales (million USD)	3968.72	1659.77	10600.06	7.86	225987.20
Number of firms that operate internationally	150 (22.56%)				
Number of firms whose stocks are traded in U.S. exchanges	598 (89.92%)				

Table 2: Panel (a): Summary statistics of all firms from Compustat/CRSP database that at least once be counted as one of 500 largest firms by monthly market value. The 500 largest firms are updated on a monthly basis between January 1995 and December 2015 using data from the previous month, which leaves me with 1,426 unique firms over the sample period. The above statistics are time series moments over firm lifetimes that overlapped with my sample period. Total debt is the sum of current and long-term debt. Panel (b): summary statistics of the 447 firms in Panel (a) that are covered by media in at least 6 months of the sample period. Panel (c): summary statistics of the 665 firms in Panel (a) that are never covered by media during the sample period.

	Covered high -	Covered high -	Non-covered high	Non-covered high	Covered high -
			-	-	
	RF	Covered low	RF	Non-covered low	Non-covered low
Mkt -RF	1.2508^{***}	0.3428^{***}	1.1858^{***}	0.2348^{***}	0.2998^{***}
	(0.0402)	(0.0699)	(0.0503)	(0.0703)	(0.0778)
SMB	0.0450	0.0792	0.1401^{***}	0.0151	-0.0800
	(0.0636)	(0.0917)	(0.0515)	(0.0945)	(0.1025)
HML	0.3855^{***}	0.5286^{***}	0.0868	0.3558^{***}	0.6546^{***}
	(0.0967)	(0.1330)	(0.0907)	(0.1365)	(0.1494)
Mom	-0.2949^{***}	-0.3333^{***}	-0.1652^{***}	-0.2082^{**}	-0.3379^{***}
	(0.0636)	(0.0986)	(0.0475)	(0.0909)	(0.1000)
BAB	-0.1233^{*}	-0.1954^{*}	-0.0851	-0.2413^{**}	-0.2795^{**}
	(0.0691)	(0.1130)	(0.0724)	(0.0987)	(0.1126)
Alpha	0.0061^{***}	0.0074^{**}	0.0026	0.0065^{**}	0.0100^{***}
	(0.0020)	(0.0030)	(0.0019)	(0.0028)	(0.0030)
Observations	252	252	252	252	252
Adjusted R^2	0.9063	0.4907	0.8998	0.3982	0.4924
Note:					$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

Table 3: Results of time series regression from February 1995 to January 2016. Dependent variables are the return differences between different high EP portfolios and low EP portfolios as defined in Section 2.3, or their excess returns with respect to the 3-month Treasury bill. Independent variables are common factors including the Fama-French 3 factors in Fama and French (1993) (Mkt-RF, SMB, HML), the momentum factor in Carhart (1997) (Mom), and the betting-against-beta factor in Frazzini and Pedersen (2014) (BAB). Standard errors (in parentheses) are adjusted for serial autocorrelation using Newey and West (1987) with a lag of 3 months.

Panel (a) Sort on static *EP*

	Covered high -	Covered high -	Non-covered high	Non-covered high	Covered high -
			-	-	
	RF	Covered low	RF	Non-covered low	Non-covered low
Mkt-RF	1.2383***	0.3416***	1.1954***	0.2735***	0.3164***
	(0.0336)	(0.0588)	(0.0503)	(0.0515)	(0.0436)
SMB	0.1527***	0.2741***	0.1478***	0.0563	0.0612
	(0.0581)	(0.0612)	(0.0569)	(0.0993)	(0.0928)
HML	0.2470***	0.2193***	−0.1269*	−0.0525	0.3214***
	(0.0689)	(0.0843)	(0.0690)	(0.0817)	(0.0766)
Mom	−0.3306***	−0.4055***	−0.2734***	−0.3542***	−0.4114***
	(0.0547)	(0.0531)	(0.0428)	(0.0722)	(0.0517)
BAB	−0.2077***	−0.2857***	−0.1205*	−0.2845***	−0.3717***
	(0.0515)	(0.0651)	(0.0646)	(0.0624)	(0.0541)
Alpha	0.0050***	0.0031	0.0015	0.0021	0.0057***
	(0.0017)	(0.0021)	(0.0017)	(0.0019)	(0.0019)
Observations	252	252	252	252	252
Adjusted R^2	0.8891	0.6268	0.8889	0.5875	0.6297

Panel (b) Pooled and sort on static *EP*

	High -	High -
	RF	Low
Mkt-RF	1.2142***	0.2999***
	(0.0410)	(0.0489)
SMB	0.1514***	0.1163
	(0.0538)	(0.0894)
HML	−0.0169	0.0267
	(0.0572)	(0.0733)
Mom	−0.2893***	−0.3651***
	(0.0391)	(0.0584)
BAB	−0.1519***	−0.2966***
	(0.0543)	(0.0585)
Alpha	0.0024*	0.0025
	(0.0014)	(0.0017)
Observations	252	252
Adjusted R^2	0.9118	0.6464

Note: $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

Table 4: Results of time series regression from February 1995 to January 2016. Dependent variables are the return differences between different high static editor preference portfolios and low static editor preference portfolios among the 500 largest firms, or their excess returns with respect to the 3-month Treasury bill. In Panel (b), I do not distinguish between covered firms and non-covered firms. Independent variables are common factors including the Fama-French 3 factors in Fama and French (1993) (Mkt-RF, SMB, HML), the momentum factor in Carhart (1997) (Mom), and the betting-against-beta factor in Frazzini and Pedersen (2014) (BAB). Standard errors (in parentheses) are adjusted for serial autocorrelation using Newey and West (1987) with a lag of 3 months.

Panel (a): project on |CumRet|					
	Covered high -	Non-covered high -	Covered high -	Non-covered high -	Covered high -
	RF	RF	Covered low	Non-covered low	Non-covered low
Alpha	0.0052**	0.0022	0.0045	0.0031	0.0061*

Panel (b): project on CumRet					
	Covered high -	Non-covered high -	Covered high -	Non-covered high -	Covered high -
	RF	RF	Covered low	Non-covered low	Non-covered low
Alpha	0.0029	−0.0018	−0.0001	−0.0038	0.0009

Panel (c): project on IVOL					
	Covered high -	Non-covered high -	Covered high -	Non-covered high -	Covered high -
	RF	RF	Covered low	Non-covered low	Non-covered low
Alpha	0.0026*	0.0004	0.0003	0.0005	0.0027

Panel (d): project on B/M					
	Covered high -	Non-covered high -	Covered high -	Non-covered high -	Covered high -
	RF	RF	Covered low	Non-covered low	Non-covered low
Alpha	0.0037***	0.0009	0.0034*	0.0028*	0.0056***

Panel (e): project on market beta					
	Covered high -	Non-covered high -	Covered high -	Non-covered high -	Covered high -
	RF	RF	Covered low	Non-covered low	Non-covered low
Alpha	0.0023	−0.0016	−0.0012	−0.0018	0.0022

Panel (f): project on SMB beta					
	Covered high -	Non-covered high -	Covered high -	Non-covered high -	Covered high -
	RF	RF	Covered low	Non-covered low	Non-covered low
Alpha	0.0037***	0.0008	0.0031*	0.0027**	0.0055***

Panel (g): project on HML beta					
	Covered high -	Non-covered high -	Covered high -	Non-covered high -	Covered high -
	RF	RF	Covered low	Non-covered low	Non-covered low
Alpha	0.0018	0.0011	−0.0009	0.0016	0.0022

Panel (h): project on Mom beta					
	Covered high -	Non-covered high -	Covered high -	Non-covered high -	Covered high -
	RF	RF	Covered low	Non-covered low	Non-covered low
Alpha	0.0032**	0.0007	0.0020	0.0017	0.0042*
Note:					*p<0.1; **p<0.05; ***p<0.01

Table 5: Alphas of time series regression from February 1995 to January 2016. Dependent variables are the return differences between different portfolios among 500 largest firms, or their excess returns with respect to the 3-month Treasury bill. Here "high" and "low" mean that I sort on the projection of *coverage* on a certain feature k, i.e. $\beta_{k,t} * feature_{k,i,t}$ for each company i and choose the top and the bottom terciles. The definitions of these eight features are consistent with the definitions in Section 2.3. Independent variables are common factors including the Fama-French 3 factors in Fama and French (1993) (Mkt-RF, SMB, HML), the momentum factor in Carhart (1997) (Mom), and the betting-against-beta factor in Frazzini and Pedersen (2014) (BAB). Standard errors are adjusted for serial autocorrelation using Newey and West (1987) with a lag of 3 months when calculate significance levels.

Features	Coefficient mean	T-statistic
$ln(size)$	~ 0	-0.9827
B/M	-0.0047	-2.1689
Market beta	-0.0001	-0.1109
$CumRet$	-0.0090	-1.4933
$\|CumRet\|$	0.0048	0.6514
$IVOL$	-0.0356	-3.5315
SMB beta	-0.0009	-1.4879
HML beta	-0.0006	-0.8393
Mom beta	0.0006	0.4644
EP	0.3400	2.2976

Table 6: Results of t-tests for the time series of the variable coefficients in Regression 2, adjusted using Newey and West (1987) with a lag of 3 months. For Regression 2, the sample period is February 1995 to January 2016. The sample includes the 500 largest firms and is updated every month. The coefficients are estimated using the elastic net method and hence the independent variables are standardized to be under the same scale.

Panel (a): Among top 500 firms (baseline)

	Covered high -	Non-covered high	Covered high -	Non-covered high	Covered high -
		-		-	
	RF	RF	Covered low	Non-covered low	Non-covered low
Alpha	0.0061***	0.0026	0.0074**	0.0065**	0.0100***
	(0.0020)	(0.0019)	(0.0030)	(0.0028)	(0.0030)
Adjusted R^2	0.8587	0.8738	0.4294	0.3249	0.4490

Panel (b): Among top 1000 firms

	Covered high -	Non-covered high	Covered high -	Non-covered high	Covered high -
		-		-	
	RF	RF	Covered low	Non-covered low	Non-covered low
Alpha	0.0061***	0.0033*	0.0078**	0.0067**	0.0095***
	(0.0020)	(0.0019)	(0.0031)	(0.0028)	(0.0031)
Adjusted R^2	0.8643	0.9013	0.4111	0.3394	0.3986

Panel (c): Among top 3000 firms

	Covered high -	Non-covered high	Covered high -	Non-covered high	Covered high -
		-		-	
	RF	RF	Covered low	Non-covered low	Non-covered low
Alpha	0.0066***	0.0045**	0.0089***	0.0088***	0.0108***
	(0.0021)	(0.0020)	(0.0029)	(0.0032)	(0.0032)
Adjusted R^2	0.8615	0.9225	0.4295	0.4337	0.4389

Panel (d): Among top 5000 firms

	Covered high -	Non-covered high	Covered high -	Non-covered high	Covered high -
		-		-	
	RF	RF	Covered low	Non-covered low	Non-covered low
Alpha	0.0074***	0.0050**	0.0090***	0.0082**	0.0106***
	(0.0024)	(0.0025)	(0.0030)	(0.0034)	(0.0031)
Adjusted R^2	0.8541	0.8946	0.4495	0.4967	0.4922

Note: *p<0.1; **p<0.05; ***p<0.01

Table 7: Alphas of time series regression from February 1995 to January 2016. Dependent variables are return differences between different portfolios among N largest firms, or their excess returns to the 3-month Treasury bill. In every month, I estimate the coefficients in Regression (1) using the 500 largest firms, and calculate EP for each of the N largest firms. Based on their EP values and being covered or not, I construct the corresponding equal-weighted portfolios as usual. Independent variables are common factors including the Fama-French 3 factors in Fama and French (1993) (Mkt-RF, SMB, HML), the momentum factor in Carhart (1997) (Mom) and the betting-against-beta factor in Frazzini and Pedersen (2014) (BAB). Standard errors are adjusted for serial autocorrelation using Newey and West (1987) with a lag of 3 months.

Panel (a): 1 month news formation					
	Covered high -	Non-covered high -	Covered high -	Non-covered high -	Covered high -
	RF	RF	Covered low	Non-covered low	Non-covered low
$t+1$	0.0061***	0.0026	0.0074**	0.0065**	0.0100***
$t+3$	0.0120**	0.0042	0.0084	0.0086*	0.0164**
$t+6$	0.0240***	0.0138*	0.0144	0.0204**	0.0306***
$t+12$	0.0424***	0.0307**	0.0275*	0.0419***	0.0536***
Panel (b): 3 months news formation					
	Covered high -	Non-covered high -	Covered high -	Non-covered high -	Covered high -
	RF	RF	Covered low	Non-covered low	Non-covered low
$t+1$	0.0047***	0.0029*	0.0075***	0.0076**	0.0094***
$t+3$	0.0102***	0.0008	0.0113***	0.0062	0.0155**
$t+6$	0.0191***	0.0012	0.0148**	0.0055	0.0233*
$t+12$	0.0397***	0.0098	0.0299**	0.0220	0.0519***
Panel (c): 6 months news formation					
	Covered high -	Non-covered high -	Covered high -	Non-covered high -	Covered high -
	RF	RF	Covered low	Non-covered low	Non-covered low
$t+1$	0.0017	0.0005	0.0007	0.0009	0.0020
$t+3$	0.0062	0.0021	0.0029	0.0045	0.0086
$t+6$	0.0119**	0.0079	0.0012	0.0071	0.0110
$t+12$	0.0266***	0.0127	0.0069	0.0151	0.0290*
Note:					*p<0.1; **p<0.05; ***p<0.01

Table 8: Portfolio alphas from my 5-factor model under different news formation period. N months news formation period means that in month t, I use the average media coverage in most recent N months to calculate $coverage_t$ and EP_t, then I constructed the corresponding portfolios at the end of month t and rebalance it again at the end of month $t+k$. Therefore in Panel (a) the first row is the baseline performance from Table 3. For portfolios held more than one month, factor returns are also extended to a longer horizon accordingly. Standard errors are adjusted for serial autocorrelation using Newey and West (1987) with a lag of 3 months as before.

Panel (a) Use neutral news

	Covered high -	Covered high -	Non-covered high	Non-covered high	Covered high -
			-	-	
	RF	Covered low	RF	Non-covered low	Non-covered low
Mkt -RF	1.2669***	0.3619***	1.1910***	0.2289***	0.3049***
	(0.0378)	(0.0617)	(0.0495)	(0.0723)	(0.0772)
SMB	0.0485	0.1187	0.1362***	0.0215	−0.0661
	(0.0632)	(0.0891)	(0.0523)	(0.0923)	(0.1008)
HML	0.3722***	0.4800***	0.1057	0.3719***	0.6384***
	(0.0969)	(0.1353)	(0.0875)	(0.1322)	(0.1471)
Mom	−0.2853***	−0.3177***	−0.1584***	−0.1934**	−0.3203***
	(0.0679)	(0.0956)	(0.0508)	(0.0944)	(0.1036)
BAB	−0.1114	−0.1956*	−0.0916	−0.2424**	−0.2622**
	(0.0697)	(0.1071)	(0.0721)	(0.1042)	(0.1161)
Alpha	0.0059***	0.0068**	0.0026	0.0064**	0.0097***
	(0.0020)	(0.0029)	(0.0019)	(0.0027)	(0.0030)
Observations	252	252	252	252	252
Adjusted R^2	0.8603	0.4449	0.8780	0.3178	0.4386

Panel (b) Include recession indicator

	Covered high -	Covered high -	Non-covered high	Non-covered high	Covered high -
			-	-	
	RF	Covered low	RF	Non-covered low	Non-covered low
Mkt -RF	1.2700***	0.3644***	1.1973***	0.2449***	0.3175***
	(0.0420)	(0.0656)	(0.0472)	(0.0764)	(0.0823)
SMB	0.0472	0.1176	0.1335**	0.0147	−0.0715
	(0.0623)	(0.0888)	(0.0527)	(0.0933)	(0.1011)
HML	0.3732***	0.4808***	0.1078	0.3770***	0.6424***
	(0.0992)	(0.1371)	(0.0901)	(0.1320)	(0.1484)
Mom	−0.2836***	−0.3163***	−0.1548***	−0.1845*	−0.3133***
	(0.0648)	(0.0951)	(0.0509)	(0.0953)	(0.1028)
BAB	−0.1101	−0.1946*	−0.0890	−0.2361**	−0.2572**
	(0.0696)	(0.1074)	(0.0726)	(0.1067)	(0.1166)
I_{Rec}	0.0016	0.0013	0.0033	0.0083	0.0066
	(0.0068)	(0.0070)	(0.0068)	(0.0062)	(0.0066)
Alpha	0.0057***	0.0066**	0.0022	0.0053*	0.0088***
	(0.0021)	(0.0033)	(0.0018)	(0.0030)	(0.0033)
Observations	252	252	252	252	252
Adjusted R^2	0.8598	0.4427	0.8777	0.3187	0.4377

Note: *p<0.1; **p<0.05; ***p<0.01

Table 9: Results of time series regression considering sentiment from February 1995 to January 2016. Dependent variables are return differences between different high EP portfolios and low EP portfolios as defined in Section 4.1, or their excess returns to 3-month Treasury bill. Independent variables are common factors including the Fama-French 3 factors in Fama and French (1993) (Mkt-RF, SMB, HML), the momentum factor in Carhart (1997) (Mom), the betting-against-beta factor in Frazzini and Pedersen (2014) (BAB) and the NBER based recession indicator (I_{Rec}). Standard errors (in parentheses) are adjusted for serial autocorrelation using Newey and West (1987) with a lag of 3 months.

Panel (a) Using number of articles

	Covered high -	Covered high -	Non-covered high	Non-covered high	Covered high -
			-	-	
	RF	Covered low	RF	Non-covered low	Non-covered low
Mkt-RF	1.2999***	0.3945***	1.1775***	0.2244***	0.3469***
	(0.0538)	(0.0692)	(0.0439)	(0.0537)	(0.0799)
SMB	−0.0196	0.1276	0.1704***	0.1242*	−0.0659
	(0.0702)	(0.0982)	(0.0460)	(0.0693)	(0.0821)
HML	0.3960***	0.4002***	0.0476	0.1577	0.5062***
	(0.1114)	(0.1516)	(0.0729)	(0.1031)	(0.1147)
Mom	−0.2988***	−0.2850***	−0.1706***	−0.1997***	−0.3279***
	(0.0620)	(0.0776)	(0.0468)	(0.0614)	(0.0710)
BAB	−0.2182**	−0.3431***	−0.1242**	−0.2915***	−0.3855***
	(0.0897)	(0.1155)	(0.0611)	(0.0674)	(0.0930)
Alpha	0.0059***	0.0072**	0.0026	0.0063**	0.0095***
	(0.0027)	(0.0030)	(0.0016)	(0.0021)	(0.0033)
Adjusted R^2	0.9063	0.4907	0.8998	0.3982	0.4924

Panel (b) Ridge regularization

	Covered high -	Covered high -	Non-covered high	Non-covered high	Covered high -
			-	-	
	RF	Covered low	RF	Non-covered low	Non-covered low
Mkt-RF	1.2420***	0.3406***	1.1708***	0.2576***	0.3288***
	(0.0444)	(0.0644)	(0.0524)	(0.0653)	(0.0695)
SMB	0.0914	0.1485*	0.1256**	0.0182	−0.0160
	(0.0673)	(0.0864)	(0.0574)	(0.0961)	(0.0996)
HML	0.3233***	0.3222***	−0.0242	0.1001	0.4476***
	(0.1010)	(0.1121)	(0.0804)	(0.1012)	(0.1236)
Mom	−0.2945***	−0.3515***	−0.1903***	−0.2461***	−0.3503***
	(0.0710)	(0.0858)	(0.0478)	(0.0923)	(0.0961)
BAB	−0.1576**	−0.2386**	−0.1115	−0.2950***	−0.3411***
	(0.0738)	(0.0964)	(0.0747)	(0.0901)	(0.1003)
Alpha	0.0065***	0.0069***	0.0036*	0.0068**	0.0096***
	(0.0020)	(0.0026)	(0.0020)	(0.0027)	(0.0027)
Adjusted R^2	0.8603	0.4449	0.8780	0.3178	0.4386

(continue)

Panel (c) Zero inflated negative binomial regression

	Covered high -	Covered high -	Non-covered high	Non-covered high	Covered high -
			-	-	
	RF	Covered low	RF	Non-covered low	Non-covered low
Mkt-RF	1.2177***	0.3169***	1.1475***	0.1861***	0.2563***
	(0.0386)	(0.0545)	(0.0391)	(0.0617)	(0.0880)
SMB	0.1176**	0.1647**	0.1381***	0.0046	−0.0159
	(0.0581)	(0.0806)	(0.0471)	(0.0824)	(0.1466)
HML	0.3290***	0.3967***	0.0711	0.2808**	0.5387***
	(0.0911)	(0.1144)	(0.0979)	(0.1399)	(0.1562)
Mom	−0.2468***	−0.3110***	−0.1561***	−0.2187**	−0.3094***
	(0.0626)	(0.0873)	(0.0482)	(0.0992)	(0.1071)
BAB	−0.1184**	−0.1664*	−0.0813	−0.2050**	−0.2422**
	(0.0560)	(0.0866)	(0.0633)	(0.0864)	(0.1197)
Alpha	0.0053***	0.0055***	0.0027*	0.0055***	0.0081***
	(0.0017)	(0.0021)	(0.0015)	(0.0021)	(0.0028)
Adjusted R^2	0.8587	0.4434	0.8786	0.3052	0.4013
Observations	252	252	252	252	252
Note:					*p<0.1; **p<0.05; ***p<0.01

Table 10: Results of time series regression using different methods to estimate editor preference as in Section 4.2 from February 1995 to January 2016. Dependent variables are return differences between different high *EP* portfolios and low *EP* portfolios, or their excess returns with respect to the 3-month Treasury bill as before. Independent variables are common factors including the Fama-French 3 factors in Fama and French (1993) (Mkt-RF, SMB, HML), the momentum factor in Carhart (1997) (Mom), and the betting-against-beta factor in Frazzini and Pedersen (2014) (BAB). Standard errors (in parentheses) are adjusted for serial autocorrelation using Newey and West (1987) with a lag of 3 months.

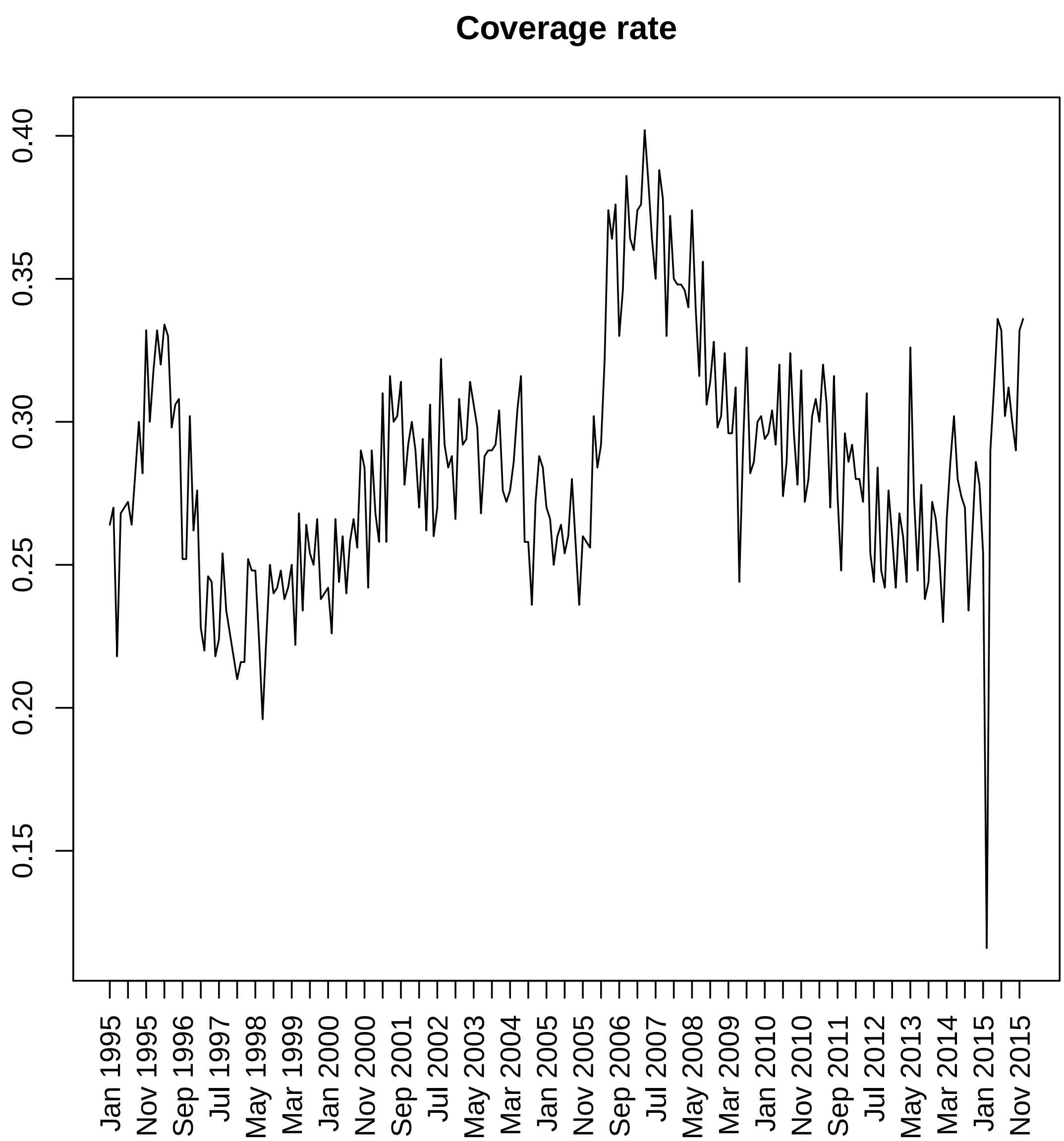

Figure 1: Graphical summary of how many firms in top 500 group are covered by the news data in every month between January 1995 and December 2015.

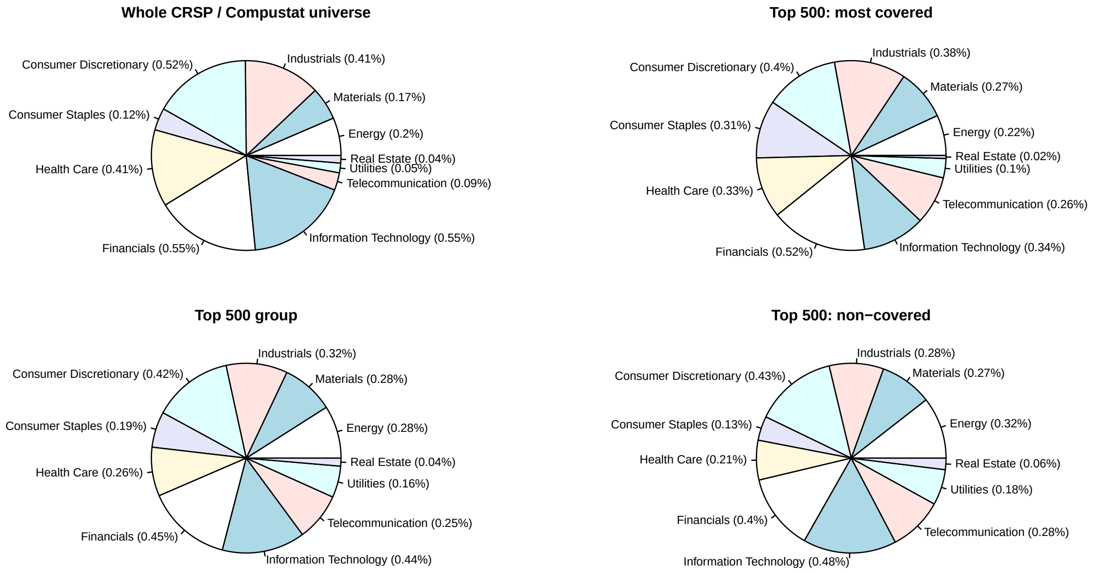

Figure 2: Graphical summary of industry distributions of 4 different firms groups from January 1995 and December 2015. I use the last available 2-digit GIC Sectors code for each firm before December 2015.

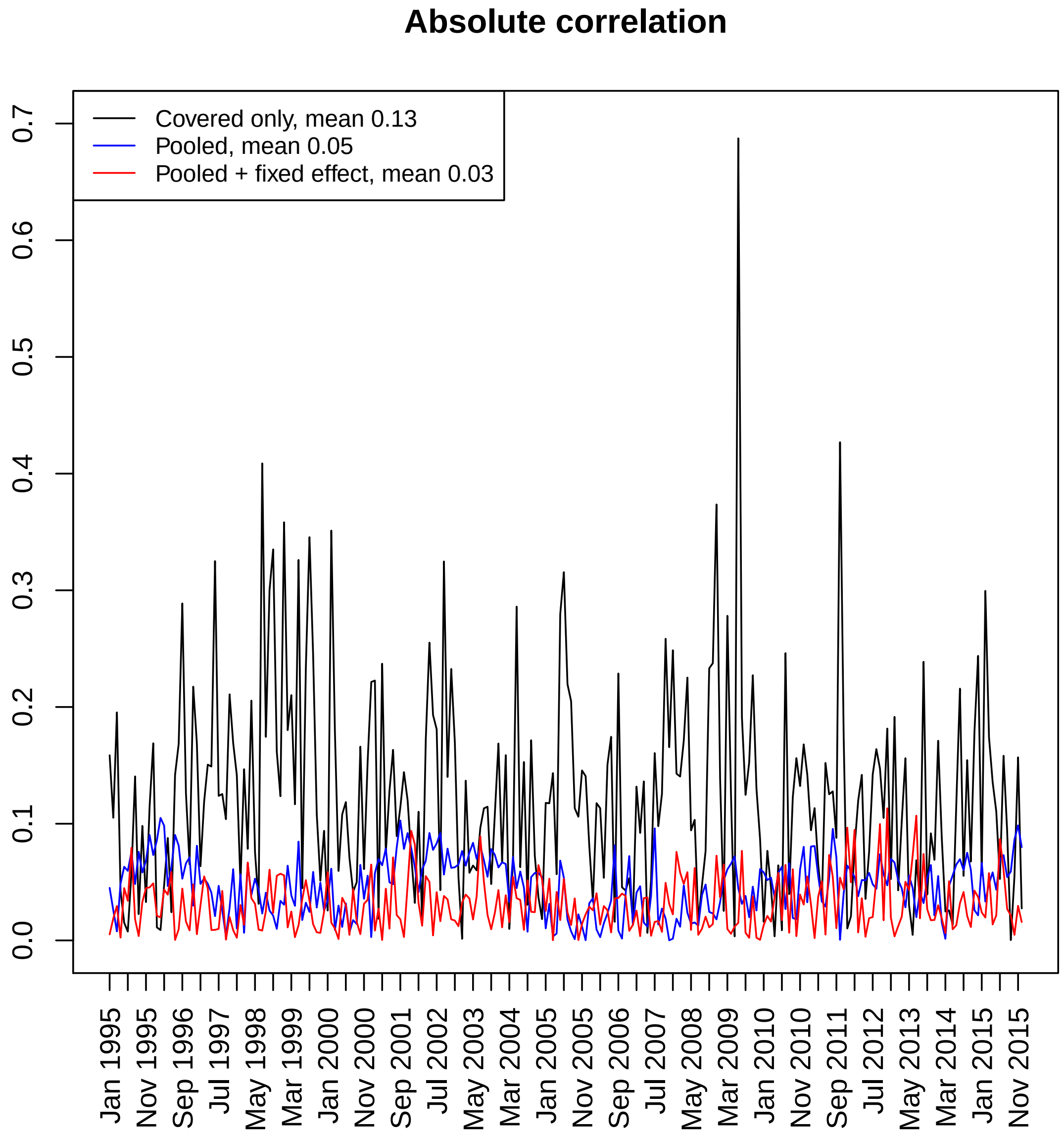

Figure 3: The time series of sample absolute correlation between $\sum_k \beta_{k,t} * feature_{k,i,t}$ and $\epsilon_{i,t}$ in regression (1) when I use covered firms only, when I use all firms, or when I use all firms plus fixed effect. The sample period is from January 1995 to December 2015.

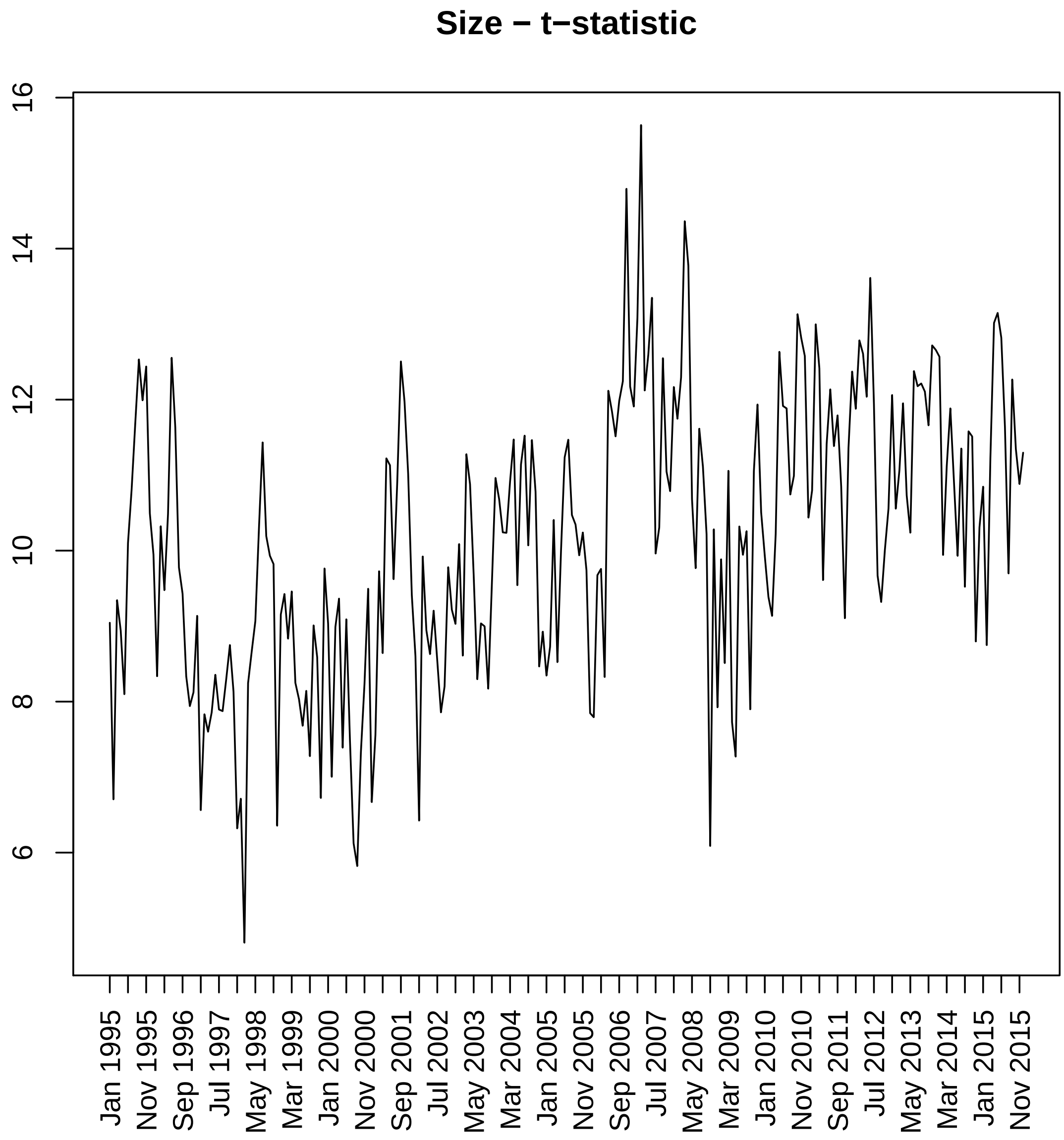

Figure 4: The monthly time series of the t-statistics of the coefficient γ_t in the cross-sectional regression (1). The data is from January 1995 to December 2015.

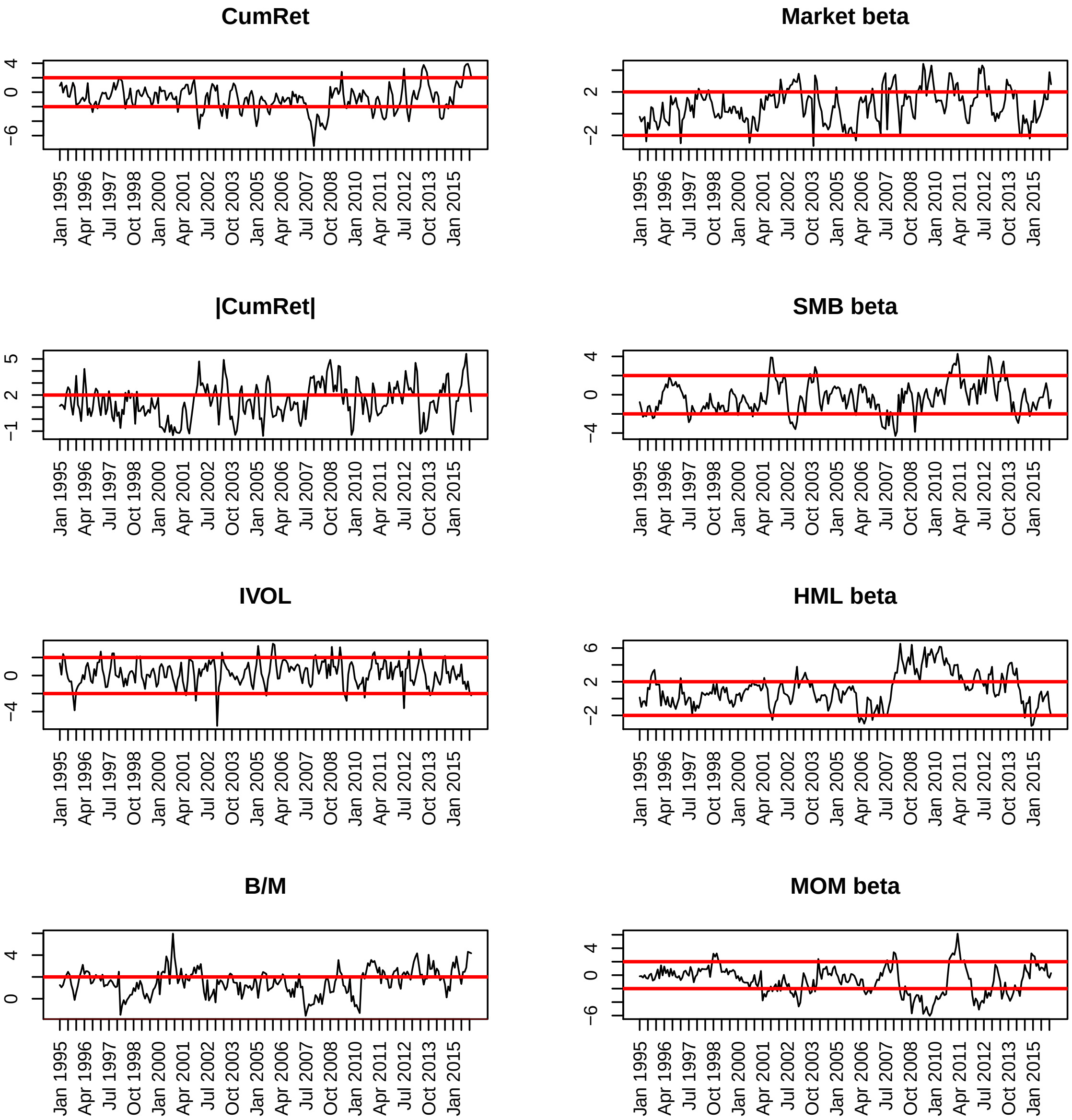

Figure 5: The monthly time series of t-statistics of the coefficients $\beta_{k,t}$ in the cross-sectional Regression (1) and the variables are defined in 2.3. The data is from January 1995 to December 2015. The red horizontal lines mark the critical values of 2 and -2.

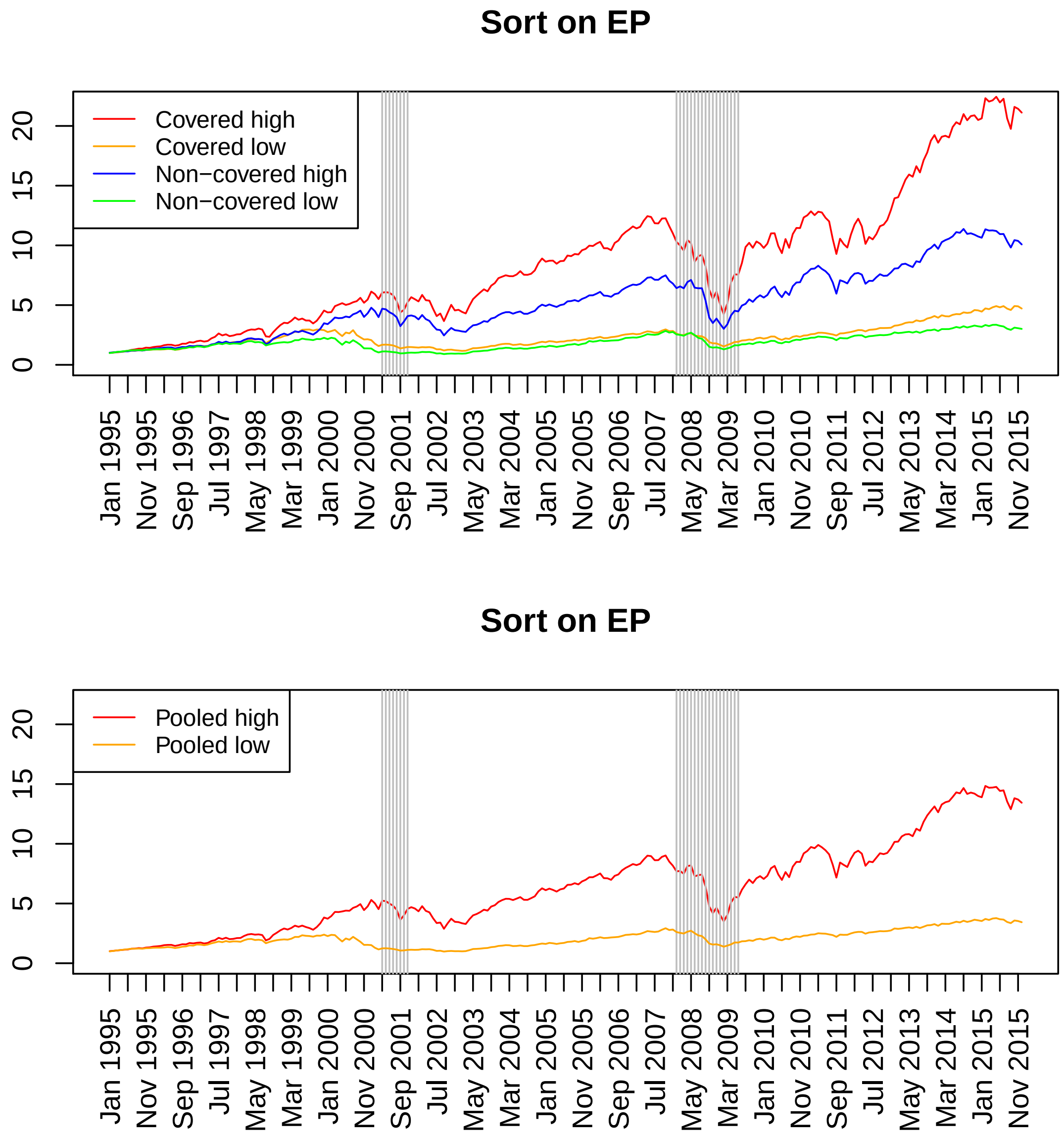

Figure 6: Cumulative returns of investing \$1 into the portfolios with monthly rebalancing from February 1981 to January 2016. These portfolios are defined as in Section 2.3. All portfolios are constructed at the end of each news forming month t in my data and the investment returns are calculated using the simple stock return in the next month $t+1$. Pooled portfolios mean that the sorting is conducted among all 500 largest firms without identifying covered firms. The grey lines mark months that are in recession according to the NBER based Recession Indicators from FRED https://fred.stlouisfed.org/series/USREC.

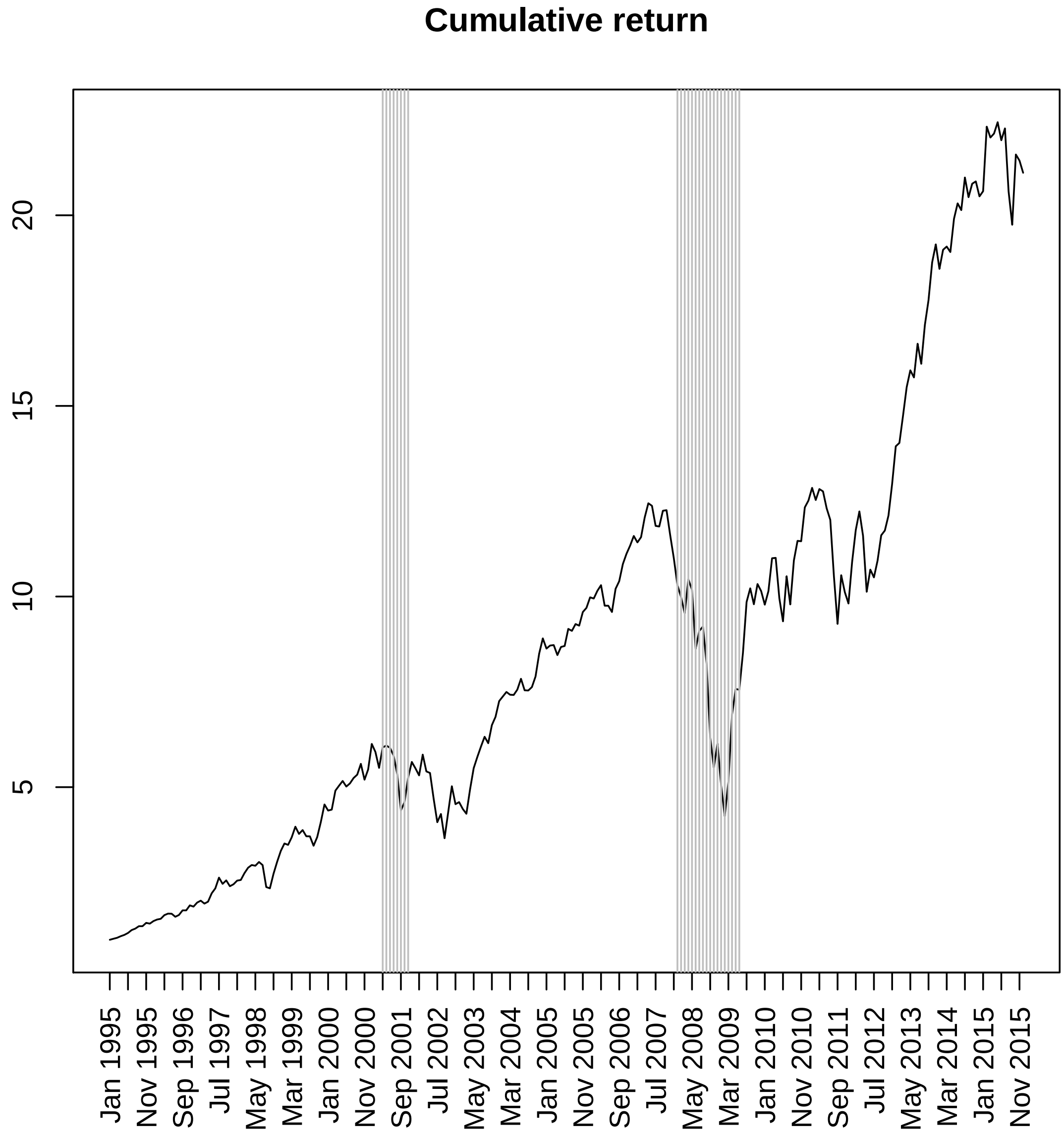

Figure 7: Cumulative returns of investing $1 into the portfolio with monthly rebalancing from February 1981 to January 2016. The portfolio consists of news-covered firms in the previous month with high EP using equal weights. The grey lines mark months that are in recession according to the NBER based Recession Indicators from FRED `https://fred.stlouisfed.org/series/USREC`.