Information vs sentiment: the impact of negative media reports on stock price

Prof. Fan Wang

School of Business, Sun Yat-sen University, Guangzhou 510275, PR China

wangfan5@mail.sysu.edu.cn

Phd. Yunpeng Zhang

School of Business, Sun Yat-sen University, Guangzhou 510275, PR China
zhangyp36@mail2.sysu.edu.cn

*Prof. Lijian Wei**

School of Business, Sun Yat-sen University, Guangzhou 510275, PR China

weilj5@mail.sysu.edu.cn

Phd. Junqin Lin
School of Business, Shantou University, Shantou 515000, PR China

junqinlin@stu.edu.cn

*Corresponding Author

Information vs sentiment: the impact of negative media reports on stock price

Abstract

To disentangle the interaction of information and sentiment in repetitive media hype, we introduce a media risk index (*MRI*) as a weighted average of five classified information-based risk measures by taking into account the repeated hype over 3 million online news from 453 media platforms through machine learning and natural language processing methods. We find that, with more investor attention, information in the media can be incorporated into the stock price more quickly, thereby reducing the degree of stock price underreaction. However, high sentiment hyped by the media can cause investors to overreact, leading to stock price overreaction. The findings provide insights for listed companies to manage information disclosure, and for market regulators to monitor market information and guide investment.

Keywords: Media Risk Index; Media Hype, Overreaction and Underreaction; Information and Sentiment.

EFM classification code: 320; 350; 620

1 Introduction

Media reports, in particular repetitive media hype, play an important role in conveying information and sentiment in financial markets by influencing investor behavior (Hillert et al., 2014). It has been well documented that, when media sentiment represents investors' sentiment, stock prices can overreact to news shocks (e.g., Tetlock, 2007; Schumaker et al., 2012; Garcia, 2013; Kim et al., 2022). However, when news carries valid information not yet reflected in stock price, the stock price can underreact to the news (e.g., Chan, 2003; Savor, 2012; Frank and Sanati, 2018). The current research primarily focuses on a few authoritative media reports from Internet media. However, there is a growing trend in numerous Internet media platforms and an abundance of news sources, which may distract investors' attention from fundamental information. In particular, repeated media reporting and hype may lead to investor sentiment, driving away from the fundamental information (Liang et al., 2020). Furthermore, the review period for Internet media is often short and the content release is uncontrollable. This may leave more room for fake news and media manipulation (Vosoughi et al., 2018). Very often, it is difficult to disentangle the interaction of information and sentiment in repetitive media hype.

To disentangle the interaction of information and sentiment, we first introduce a Media Risk Index (*MRI*) as a weighted average of five classified traditional information-based risk measures in this paper. We estimate the weights by taking into account repeated hype through over 3 million online news from 453 media platforms using machine learning and natural language processing methods. The news data includes 250 listed Chinese companies penalized by the China Securities Regulatory Commission between 2014 and 2022. We then decompose the information and sentiment in MRI into a *Media Risk Information Index (MRII)* and a *Media Risk Sentiment Index (MRSI)* through media reporting intensity. We show that the effect of media on stock prices depends on the interaction of the information effect and the sentiment effect.

The MRI provides a unified measure to capture both stock price underreaction and overreaction from three different aspects. First, we focus on the traditional types of risk information in the news, instead of emotional features in the previous studies (Garcia, 2013; Guégan & Renault, 2021; Maghyereh & Abdoh, 2022). We consider five classified information-based risk measures: *operational risk*, *accounting risk*, *stock market risk*, *legal policy risk*, and *other risks*, which are traditional risk factors in the literature. Each article is classified based on these risk measures. Second, to capture repetitive media hype, machine learning training features are implemented separately for article titles and contents. The commonly used dictionary method involves counting positive and negative words in article content (Tetlock, 2007; Tetlock and Macskassy, 2008; Loughran & McDonald, 2011). It is unable to accomplish complex risk classification tasks and does not align with our news browsing habits in the era of internet media. Third, for the weighting, we consider the intensity of the media coverage. According to agenda-setting theory (Cao et al., 2021), repeated reports and hype by the media enhance public attention. Therefore, the intensity of media coverage reflects the extent of the impact range of media risk.

The empirical test on the explanation power and the underlying mechanisms of the MRI (including MRII and MRSI) is conducted in three steps. First, we examine the effect of MRI on stock prices on the same-day and multi-day. Second, we collect posts from stock forum boards to construct investor attention and sentiment indicators and analyze the information and sentiment effects in the MRI. Finally, based on the effect of MRI, we classify companies and build profitable investment strategies. We find the following novel results.

First, whether media risk has an underreaction or overreaction on stock prices depends on the balance between the information effect and the sentiment effect in the media. With more investor attention, the information in the media will be incorporated into the stock price more quickly, thereby reducing the degree of underreaction. Meanwhile, the sentiment in the media can cause an overreaction in stock prices with high investor attention and an underreaction in stock prices with low investor attention. The sentimental aspect of media coverage results in an investor sentiment to overreact on stock prices, though investor sentiment does not influence media sentiment. Different from the previous studies that have separated the media's sentiment theory and information theory (Tetlock, 2007), we differentiate the media's informational and sentimental effects under a unified framework.

Second, we find that companies with a higher level of investor cognition are less susceptible to media risk. That is, companies with high audit quality, information disclosure transparency, investor literacy, and analyst attention are less affected by media risk. This can guide listed companies in market value management and information disclosure.

Third, we test the predictability of MRI on stock prices. Empirical evidence suggests that the stock price underreacts to the MRI, indicating that the MRI can be used to predict stock prices. By constructing a long-short investment portfolio based on MRI, we can generate an excess return of 12.5% per annum. This provides investors with opportunities to profit from the MRI.

Furthermore, we validated two different mechanisms through which media influences investor sentiment and expanded on Ren et al.'s (2021) work. On one hand, the media pays attention to investors sentiment in the market and uses it as a form of information. On the other hand, the sentiment reported repeatedly by the media can correct the overreaction of investor sentiment. However, investor sentiment does not affect media sentiment.

Our choice to focus on the Chinese market is motivated by its status as an emerging development market, in which the A-share market exhibits two distinct characteristics. First, in comparison to the stock markets in developed countries, such as the US, the Chinese market is dominated by individual investors who account for a significant majority. According to statistics from the Shanghai Stock Exchange in 2020, natural person investors make up 99.76% of the market. Consequently, individual investors, lacking professional knowledge, are susceptible to media sentiment. Second, China's stock market imposes restrictions on short selling, which results in an asymmetric effect of stock prices on negative financial news. Therefore, investigating the effect of media risk on stock prices in the Chinese markets is particularly meaningful. The findings also provide some implications for listed companies to manage information disclosure, and for market regulators to monitor market information and guide investment.

In the remainder of this paper, we highlight the contributions of this paper to the related literature in Section 2, present the methodology and data in Section 3, and empirical results in Section 4. Section 5 concludes.

2 Related Literature

This paper mainly contributes to the literature on the impact of media reports on stock markets. Although related studies have examined overreaction, underreaction, and media attention, most research focused on a few authoritative media outlets. Research on the impact of media risk in the context of the Internet is lacking.

One strand of literature focuses on the impact of media sentiment. Tetlock (2007) uses news data from *The Wall Street Journal* to construct the negative media sentiment and finds that media sentiment leads to an overreaction in stock prices. Based on the noise trader theory, he proposes that media sentiment can represent investor sentiment. Tetlock (2011) further defines the concept of outdated news using news data from the *Dow Jones newswire* and finds that outdated news is often accompanied by a reversal in stock prices, providing further evidence of media sentiment. Garcia (2013) identifies positive and negative sentiments in *The New York Times* and finds that media sentiment leads to a partial reversal of stock prices (within four days). Birru and Young (2022) argue that news-based sentiment measuring highly predicts returns during economic downturns. By examining how media news affects consumer sentiment, Kim et al. (2022) find that investors overreact to bad news, resulting in a negative herd effect. Conversely, Ren et al. (2021) find that social media influences the sentiment of mass media toward financial news.

Another strand of literature focuses on the impact of information in the media. Chan (2003) finds that negative news leads to long-term drifting of stocks, implying an underreaction of investors to bad news. Zhang et al. (2016) also identify an underreaction in the short term, linking it to the degree of attention toward a company. Frank and Sanati (2018) further find that the stock market overreacts to good news and underreacts to bad news, reflecting the interaction between retail investors with attention bias and arbitrageurs with short-term capital constraints. Peress and Fang (2009) find that companies that receive little media attention gain substantial future returns, and explain this through investor recognition hypothesis. Peress (2014) finds that the media, by enhancing the information transmission among investors, includes information on stock prices, thereby increasing the efficiency of the stock market. Hillert et al. (2014) find that media attention leads to investor bias, and stocks of companies receiving attention have a significant momentum effect.

These studies typically consider media attention and sentiment separately and do not consider the media risk brought about by the intensity of media coverage and the hype. As agenda-setting theory suggests, the intensity of media coverage can affect the public’s perception and their degree of concern (Lippmann, 1922; Funkhouser, 1973; Cao et al., 2021). As negative news impacts more severely than positive news (Garcia, 2013), bad news tends to garner more media attention in the era of internet media, leading to frequent re-publication by multiple media outlets (Zhang et al., 2016). When the media actively highlights negative news, the odds of this media risk irrationally affecting investors increase.

In this paper, we construct the MRI using machine learning and natural language processing methods. In the previous literature, the majority of studies use the dictionary method to construct media sentiment, including “The Harvard IV-4 Dictionary” (Tetlock, 2007; Tetlock & Macskassy, 2008), “Henry Dictionary” (Henry, 2008), “Diction Dictionary” (Davis et al., 2012), and “Loughran and McDonald Dictionary” (Loughran & McDonald, 2011), which are commonly used English dictionaries. The dictionary method can be useful to judge sentiment, but is limited to depict highly complex features within the text. As machine learning methods have advanced, supervised machine learning methods have been widely used for text analysis in finance research. For example, Manela and Moreira (2017) utilize the Support Vector Machine (SVM) method and one-hot coding to extract the bank volatility index from media reports. Antweiler and Frank (2004) and Das and Chen (2007) also use various machine learning methods to perform sentiment analysis on posts from Yahoo Finance and to predict stock prices. Following the advancements in natural language processing technology, some studies have begun to utilize deep learning techniques for sentiment analysis (Pathak et al., 2021; Costola et al., 2023), to which this paper contributes as well.

3 Media Risk Index (MRI)

We construct the media risk index (MRI) in three steps. First, we define features of the sentiments and risks in the media. Second, using machine learning methods, we extract the sentiment and risk features from the news title and contents. Finally, we build the MRI based on machine learning results.

3.1 Feature Definition

The features we extract from news reports include two categories. First, we need to determine whether media reports are generally positive or negative. Second, we need to categorize the risk type for every media report. The Basel Agreement categorizes bank risks as market, credit, liquidity, and operational risks. According to Leo et al. (2019), the risks that banks and other financial institutions face include legal, product, reputation, policy, and other non-accounting risks. Accordingly, we categorize risk types in media into the following five categories in light of the prior literature analysis and the characteristics of media reports[1]: operational risk, accounting risk, stock market risk, legal policy risk, and other risks. Table 1 lists the definitions and typical examples of the five categories of risks.

Table 1: Specific definitions and common examples of the five risk types

Risk types	*Specific explanations*
Operational risk	Risks related to business operations, such as company's product and operational process problems
Accounting risk	Risks related to accounting and finance, such as asset structure, receivables, and financing risks
Stock market risk	Risks related to the stock market, such as shareholder reduction, stock price fluctuation, and equity pledge
Legal policy risk	Risks related to laws, regulations, and policies, such as litigation, disputes, inquiries, and administrative penalties
Other risks	Risks not belonging to the preceding four types

News of different risk types contain different amounts of information and should be weighted differently. Fundamental investors base their investment choices on a company's intrinsic value. Their assessments are most influenced by financial statements. Therefore, accounting risk is highly weighted, while the other risks are less weighted due to the limited residual information. We use the Analytic Hierarchy Process[2] (AHP) method to assign weights to the importance of the information contained in different risks. *Table 2* lists the weights of the five categories of the risks.

Table 2: Weights of Media Risks

Risk types	*Weights (θ)*
Operational risk	16.08%
Accounting risk	31.01%
Stock market risk	23.50%
Legal policy risk	24.39%
Other risks	5.02%

Notes: Through the AHP, experts compare and score the importance of each pair of risks in their judgment matrices. With the premise of passing the consistency test, the weights of each type of risk are calculated.

[1] According to the Basel Agreement, listed companies also face operational risk and market risk, whereby market risk specifically refers to stock-market risk, whereas credit and liquidity are risks specific to banks. Following Leo et al. (2019), we classify the legal and policy as legal policy risk, and reputation and product as operational risk. In addition, owing to the role that media reports play as intermediaries in conveying information about listed companies, and as the main disclosure of company information coming from corporate financial statements, accounting risk is set up to assess news-related risks associated with the accounting and finance of the company.

[2] AHP is a method that assigns weights to evaluation indexes in a subjective manner (Vaidya & Kumar, 2006). We ask three experts to provide subjective assessments. Experts' assessments of the relative weight of each factor are compared and scored in pairs to create a judgment matrix. When the judgment matrix passes the consistency test, the weight of each evaluation index can be determined. From the judgment matrix, one can learn the precise weight assigned to each risk.

3.2 Feature Extraction

To extract the sentiment and risk features in the news, the documents are vectorized, thereby transforming the unstructured data into structured data. After data cleaning, word segmentation, part of speech tagging, and removing stop words for each news article, we use One-hot representation and TF-IDF representation to convert document data to vector[3].

One-hot representation is simply assigning a vector value based on all words in the word bag (Manela and Moreira, 2017). The position of the vector is given a value of 1 when a word in this document appears in the word bag; otherwise, a value of 0 is given. One-hot coding helps to expand features and addresses the issue of unstructured data that the classifier cannot handle. The weight assigned is the same regardless of the frequency of word occurrence, which is a clear flaw in this method. As a result, Loughran and Mcdonald (2011) propose the TF-IDF method to consider the rarity of the words and their frequency of occurrence in various documents,

$$w_{i,j} = \begin{cases} \dfrac{(1 + \ln tf_{i,j})}{(1 + \ln a_j)} * \ln \dfrac{N}{df_i} , & tf_{i,j} \geq 1, \\ 0 , & other \end{cases} \tag{1}$$

where $w_{i,j}$ represents the weight of word i in document j, which is composed of two parts. The first part is TF (term frequency), which reflects the frequency of the keywords. In this part, $tf_{i,j}$ represents the frequency of occurrence of keywords i in document j and a_j represents the number of all the words in document j, which is used for normalization. The second part of IDF (inverse document frequency) reflects the prevalence of keywords, where N represents the total number of the documents, and df_i represents the number of documents in which the word i appears. As a particular word becomes common, its IDF value decreases, and this word becomes less important. Therefore, TF-IDF can substantially extract the keyword weight of the document by multiplying TF and IDF to take into account of word frequency and freshness. We use TF-IDF to find keywords for risk types to assess the classification's accuracy, see Appendix A for the details.

The method that integrates the TF-IDF and one-hot encoding can be employed for document vectorization, accurately delineating the attributes of each media report text and mitigating the high sparsity of each text vector. With the TF-IDF, we first extract the keywords from each media report's text. One-hot encoding is then performed for all the keywords.

After building the document vector and manual annotation of the results, we use supervised machine learning models such as Naive Bayes (NB), Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), and Adaboost for feature extraction (Antweiler and Frank, 2004; Das and Chen, 2007; Manela & Moreira, 2017). We use traditional machine learning methods, rather than simple dictionary methods, or the more complex deep learning methods. The reason is that the current

[3] Stop words are typically referred to as articles and conjunctions without any informational value.

dictionary method can only determine sentiment and cannot classify the risk type that the media belongs to. Furthermore, deep learning models have poor interpretability and are prone to the issue of overfitting.

We manually label all news reports on a typical listed company, including its risk and sentiment types. The degree of negativity is determined using the document vector of the news title. Given that the document title frequently summarizes the media report in a single sentence, it can condense the most crucial details and usually has clear sentimental characteristics. The risk type is determined through the document vector of the news content. Obviously, the body of an article contains the most information because its content is a thorough report and analysis of the company.

We find that the Naive Bayes model performs better than other models in judging the degree of negativity, while the Random Forest model outperforms other models in judging risk, see Appendix B for detailed information on the training of specific machine-learning models.

3.3 Index Construction

We construct document vectors using the feature representation method described in Section 3.2 and input them into the machine learning model (trained according to Appendix B) to determine the negativity level of all the news articles and the probability of the risk type they belong to. Next, for each stock k and each risk type m, we construct a media risk indicator, $risk_{k,t,m}$, from all the articles related to stock k over the time period t (say a day),

$$risk_{k,t,m} = \sum_{a=1}^{n} neg\_sentiment_{k,t}^{a} * risk_{k,t,m}^{a} \; , \qquad (2)$$

where $neg\_sentiment_{k,t}^{a} \in (0,1)$ refers to the negative degree of article a predicted, $risk_{k,t,m}^{a}$ refers to the media risk indicator of article $a,$ measured by the probability of risk m article a belongs to, and n denotes the total number of the articles related to stock k over the time period t. Both $neg\_sentiment_{k,t}^{a}$ and $risk_{k,t,m}^{a}$ are respectively the predicted probabilities of the negativity and different types of risks of the news through the machine learning model. Finally, we aggregate the individual media risk and generate a media risk index, $MRI_{k,t}$, as follows,

$$MRI_{k,t} = \sum_{m=1}^{5} risk_{k,t,m} * \theta_m = \sum_{a=1}^{n} [\left(neg\_sentiment_{k,t}^{a}\right) * (\sum_{m=1}^{5} risk_{k,t,m}^{a} * \theta_m)], \qquad (3)$$

here θ_m represents the relevant risk weight in *Table 2*.

Intuitively, the risk type index ($risk_{k,t,m}$) reflects the risk value of each type of risk contained in all the news of company k each day. MRI ($MRI_{k,t}$) is a weighted average of different types of risks, characterizing the interaction of information and sentiment from three different aspects. First, it reflects the intensity of the media coverage of the day. In the Internet age, investors can access many sources of information. A company may have multiple news articles in a single day, and different platforms may have duplicate coverage of the same hot event. As suggested by agenda-setting theory, it is simple to draw the public's attention with highly frequent mass media reports (Funkhouser, 1973; Cao et al., 2021). All relevant news of the company should be summarized every day to consider the risk of media hype caused by the intensity of media reporting. Second, it uses the news title to reflect the degree of negativity. Because of limited attention, the first and most important thing that investors focus on is the news title. With the article content remaining unchanged, different titles directly indicate the degree of an investor's risk perception. Third, it uses the content of the news to reflect the amount of information about different types of risks. When amusing titles attract investors to read the content, rich information would come to them. Keeping the article title the same, different article contents determine their risk type and the amount of information contained.

Intuitively, the sentiment captures the hype caused by repeated media reporting, while the information refers to the stock price information in the daily media. To disentangle sentiment from information, for each stock each year, we regress the MRI towards the intensity of media reporting. We refer the part that can be explained by the intensity of media reporting as *Media Risk Sentiment Index (MRSI*) and the unexplainable part as *Media Risk Information Index (MRII)*.

Previous studies have mostly separated the influence of media sentiment and media reporting intensity (Peress et al., 2009; Zou et al., 2019; Umar et al., 2021; Biktimirov, 2021). The construction of media sentiment mainly uses the method of counting the number of positive and negative words in article content (Tetlock, 2007; Tetlock & Macskassy, 2008; Zhang et al, 2016). However, this construction method has two flaws: Firstly, it does not effectively reflect the public's news reading habits in the era of Internet media. People receive considerable news every day through Internet media. Owing to their limited attention, most of the news is only seen through titles without carefully reading the contents of the articles. When investors are interested in the content of the article, they can acquire additional rich information. Secondly, it overlooks the hype phenomenon brought about by the media's repeated reporting. When a company frequently appears in public view, the agenda-setting function of the media can enhance investors' attention towards it, thereby triggering irrational sentiment in investors. The MRI combines the intensity of media reports and the negativity of each news article while considering the information content of different types of risks. The negativity of each news article and the weight of the risk type reflect the information contained in the news, while the reporting intensity reflects the emotional hype in the media. Furthermore, we dissect the information and sentiment in media reports through the intensity of media coverage, which allows us to consider the functions of information and emotions separately in subsequent explorations.

3.4 Data

The news text from ChinaScope[4] covers 3,285,120 media reports that were

[4] https://chinascope.com/

gathered from 453 media platforms for 250 listed companies between January 1, 2014, and October 31, 2022. The 250 listed companies we selected are those that have been fined by the CSRC[5], which are likely to produce a barrage of bad press and run the danger of a stock price collapse. News sources are all well-known we-media public account platforms and major financial media platforms in China. Article names, body contents, publication dates, and pertinent platforms are all included in the data.

Our explained variable is the daily return of the company. The daily return of a stock (*r*) is calculated using the opening prices of two consecutive days[6]. The main explanatory variables are the company's daily MRI, MRII and MRSI. The control variables consist of two categories. The first category includes daily frequency time series data, including the Fama-French three factors. The second category is the company panel data with annual frequency, including whether audited by a Big Four accounting firm (BIG4), whether it is the nature of state-owned enterprises (SOE), percentage of top five shareholders (TOP5), percentage of institutional investors (POI), log of the total asset (SIZE), market-to-book ratio (MTB), return on asset (ROA), book leverage (LEV) and analyst attention (Analyst). The above data are obtained from CSMAR[7] except for our constructed media risks and MRI. To explore the impact of investor attention and sentiment, we collect stock bar data for the 250 companies using CNRDS[8], which include daily posts (Tpostnum) that are either positive (Pospostnum) or negative (Negpostnum). The definitions and descriptive statistics of the variables are provided in Appendix C.

4. Empirical Analysis

This section first examines the impact of the MRI on stock prices. We then conduct a mechanism analysis by examining the investor attention effect and investor sentiment effect, respectively.

4.1 Impact of MRI on Stock Prices

We first examine the impact of the MRI on stock prices on the same day. To explore stock price underreaction and overreaction, we also examine the impact of the MRI over multi-day.

4.1.1 Same-day Impact

We use the Fama-French three-factor model to examine the same-day impact of

[5] China Securities Regulatory Commission (CSRC).

[6] The opening price reflects the initial buying and selling sentiment of the market towards a particular stock, while the closing price may be influenced by other factors within the day, such as changes in market sentiment and fluctuations in trading volume. Considering media reports after the stock market closes, we calculate the daily return using the opening price of two consecutive days.

[7] CSMAR: China Stock Market & Accounting Research database, see https://data.csmar.com.

[8] CNRDS: Chinese Research Data Services Platform, see https://www.cnrds.com.

the MRI, MRII, and MRSI on stock prices. In model (1), we examine the impact of MRI on stock return. Model (2) investigates the effect of MRII and MRSI on stock returns. Models (3) and (4) extend the analysis further to include the Fama-French three factors and control variables. The data in the control variables are obtained from the company's annual report, which is typically disclosed in the first half of the second year, leaving the current year unknown. Thus, we use the previous year's annual report data as the control variables for the current year. We control the year-time fixed effect (Year) to prevent the endogenous problems caused by missing variables. The complete regression is as follows,

$$r_{k,t} = \alpha + \beta * MRI_{k,t} + \sum_{x} \gamma_x * factor_{x,t} + \sum_{y} \vartheta_y * control_{y,k,t} + Year + \varepsilon_{k,t}, \tag{4}$$

$$r_{k,t} = \alpha + \beta_1 * MRII_{k,t} + \beta_2 * MRSI_{k,t} + \sum_{x} \gamma_x * factor_{x,t} + \sum_{y} \vartheta_y * control_{y,k,t} + Year + \varepsilon_{k,t} \tag{5}$$

where $factor_{x,t}$ represents the Fama-French factors and $control_{y,k,t}$ represents the control variables, including BIG4, SOE, TOP5, POI, SIZE, BM, ROA, LEV and Analyst.

Table 3: Impact of MRI on stock price returns

	(1) r	(2) r	(3) r	(4) r
MRI	**-0.047^{***}**		**-0.044^{***}**	
(e−2)	**(−26.22)**		**(−26.68)**	
MRII		**-0.039^{***}**		**-0.038^{***}**
(e−2)		**(−16.72)**		**(−17.38)**
MRSI		**-0.057^{***}**		**-0.053^{***}**
(e−2)		**(−20.85)**		**(−20.78)**
MRK_RF			0.636^{***} (274.18)	0.636^{***} (274.19)
SMB			0.579^{***} (125.06)	0.579^{***} (125.03)
HML			-0.120^{***} (−25.40)	-0.120^{***} (−25.41)
Control	No	No	Yes	Yes
Year FE	Yes	Yes	Yes	Yes
N	795,750	795,750	795,750	795,750
R^2	0.0012	0.0012	0.1875	0.1875

t statistics in parentheses. * *$p < 0.10$,* ** *$p < 0.05$,* *** *$p < 0.01$*

Notes: This table describes the impact of MRI, MRII and MRSI on stock price returns, with the addition of the Fama-French factors and control variables in the explanatory variables in columns (3) and (4), where e−2 represents the reported numbers in the unit of 10^{-2} *, e.g., −0.063 in column (1) represents −0.063%.*

The daily effect of MRI, MRII, and MRSI on returns are negative and highly significant, as shown in Table 3. The impact of MRI on daily returns is about -0.045%, which is not affected by Fama-French factors and other variables. Among them, MRSI has a bigger impact on stock return rates, reaching -0.055%, while the influence of MRII is -0.038%. This shows that negative sentiment caused by repetitive media reporting is more likely to result in a risk of stock price decline. Further analysis (see the descriptive statistics in Appendix C) shows that the MRSI has a maximum value of 181. This indicates that the MRSI can lead to a drop of approximately 10% (181*0.055%=10%) in the stock price return on that day, which can be significant enough to trigger a stock market crash. All the Fama-French three factors are significant, indicating that the market, company size, and book value have an impact on these stocks.

4.1.2 Multi-days Impact

Chan (2003) introduces the concepts of underreaction and overreaction to describe the scenarios where the abnormal returns following an event exhibit the same/opposite sign as the returns on the event date. The current research only pays attention to a few authoritative media outlets and has not reached a consensus on whether media sentiment causes an overreaction or underreaction in stock prices (Chan, 2003; Tetlock, 2007; Tetlock, 2011; Garcia, 2013; Kim et al., 2022; Frank and Sanati, 2018). Given the timeliness and high frequency of Internet media, we investigate whether stock prices underreact or overreact to the MRI in multi-day.

We use vector autoregression (VAR) to run the Granger causality test of return and the MRI (MRII and MRSI are included), as in equations (6) and (7). We accept lag variables of up to four days because the market only opens on weekdays. We define a lag operator L4 (Tetlock, 2007), that is, $L4(x_t) = [x_{t-1,}x_{t-2,}x_{t-3,}x_{t-4,}]$. *Table 4* reports the results of the regression.

$$r_{k,t} = \alpha_1 + \beta_1 * L4(r_{k,t}) + \beta_2 * L4(MRI_{k,t}) + \varepsilon_{1,t}, \tag{6}$$

$$MRI_{k,t} = \alpha_1 + \beta_1 * L4(r_{k,t}) + \beta_2 * L4(MRI_{k,t}) + \varepsilon_{2,t}\ . \tag{7}$$

Table 4: Vector autoregressive models for returns and MRI

	r_t	MRI_t
r_{t-1}	-0.004^{**} (-2.01)	$\mathbf{-0.196^{**}}$ $\mathbf{(-2.02)}$
r_{t-2}	0.004^{**} (2.07)	-0.001 (-0.01)
r_{t-3}	0.024^{***} (12.80)	-0.153^{*} (-1.82)
r_{t-4}	0.012^{***} (7.06)	0.044 (0.68)
$MRI_{t-1}(\mathrm{e}-2)$	$\mathbf{-0.016^{***}}$ $\mathbf{(-4.70)}$	0.547^{***} (23.18)
$MRI_{t-2}(\mathrm{e}-2)$	-0.003 (-0.99)	-0.049^{***} (-3.07)
$MRI_{t-3}(\mathrm{e}-2)$	-0.004 (-1.41)	0.095^{***} (6.38)
$MRI_{t-4}(\mathrm{e}-2)$	0.005 (1.60)	0.065^{***} (5.10)
N	777,250	777,250

t statistics in parentheses. * p < 0.10, ** p < 0.05, *** p < 0.01

Notes: This table describes the vector autoregressive model of the four-day lagged return and MRI. It respectively uses the return and MRI of the same day as the dependent variables and uses the returns and MRI of the previous four days as independent variables, where e−2 represents the reported numbers in the unit of 10^{-2}, e.g., −0.016 in column 2 for MRI represents −0.016%.

In the regression of equation (5), the first lagged variable of MRI is significantly negative (−0.016), whereas the second, third, and fourth lagged MRI variables are not significant. Given that the sign is consistent with the effect of the MRI on the stock price on the same day in Table 3, the stock price underreacts to the MRI, and this insufficient reaction only occurs in the media reports from the previous day. In equation (6) of the MRI regression, the lagged variable of the return is significantly negative in the first order (−0.196). This finding suggests that news media also report the risk of stock price decline from the previous day, though MRI has an insufficient reaction to the stock price. Additional analysis without considering the auto-correlation of stock prices indicator shows that the MRI underreacts to the stock prices over the next 10 trading days. However, as time passes, both the degree and significance of this underreaction decrease (see Appendix D for the details). Compared with Table 4, the stock price absorbs the information contained in the media reports, resulting in returns being affected only by the previous day’s media.

Table 5: Vector autoregressive models for returns and MRII/MRSI

	(1) Information effect			(2) Sentiment effect	
	r_t	$MRII_t$		r_t	$MRSI_t$
r_{t-1}	-0.004^{*}	-0.151^{*}	r_{t-1}	-0.004^{**}	0.007^{*}
	(−1.88)	(−1.75)		(−2.06)	(0.11)
r_{t-2}	0.004^{**}	−0.091	r_{t-2}	0.004^{**}	$\mathbf{0.146^{***}}$
	(2.20)	(−1.50)		(2.04)	**(3.95)**
r_{t-3}	0.024^{***}	$\mathbf{-0.155^{**}}$	r_{t-3}	0.024^{***}	0.017
	(12.91)	**(−2.04)**		(12.80)	(0.47)
r_{t-4}	0.012^{***}	0.090	r_{t-4}	0.013^{***}	−0.021
	(7.11)	(1.56)		(7.07)	(−0.58)
$MRII_{t-1}(\mathrm{e}-2)$	$\mathbf{-0.006^{**}}$	0.447^{***}	$MRSI_{t-1}(\mathrm{e}-2)$	$\mathbf{-0.046^{***}}$	0.712^{***}
	(−1.96)	(23.47)		**(−4.61)**	(13.09)
$MRII_{t-2}(\mathrm{e}-2)$	−0.002	−0.01	$MRSI_{t-2}(\mathrm{e}-2)$	0.001	$\mathbf{-0.098^{**}}$
	(−0.72)	(−1.11)		(0.10)	**(−2.45)**
$MRII_{t-3}(\mathrm{e}-2)$	−0.001	0.093^{***}	$MRSI_{t-3}(\mathrm{e}-2)$	−0.014	0.059^{*}
	(−0.43)	(6.89)		(−1.23)	(1.69)
$MRII_{t-4}(\mathrm{e}-2)$	0.004	0.047^{***}	$MRSI_{t-4}(\mathrm{e}-2)$	$\mathbf{0.015^{**}}$	0.124^{***}
	(1.21)	(4.33)		**(1.96)**	(3.56)
N	777,250	777,250	N	777,250	777,250

t statistics in parentheses. * *$p < 0.10$,* ** *$p < 0.05$,* *** *$p < 0.01$*

Notes: This table describes the vector autoregressive model of the four-day lagged return and MRII/MRSI. For the information effect, we use vector autoregression (VAR) of MRII and returns. For the sentiment effect, we use VAR of MRSI and returns, where e−2 represents the reported numbers in the unit of 10^{-2} .

Furthermore, we investigate the information effect (MRII) and sentiment effect (MRSI) in MRI, and the results are shown in Table 5. For the information effect, we use a vector autoregressive model of MRII and returns, while for the sentiment effect, we replace it with MRSI. In the regression of returns, the first lag of MRII is negative (−0.006), which is consistent with the sign in Table 3 (−0.038), indicating that the stock price has an underreaction to the information component of media risk. On the other hand, the first lag of MRSI is negative (−0.046), and the fourth lag is positive (0.015), suggesting that the stock price has an overreaction to the sentiment component of media risk. In the regression of MRII, the third lag of returns is significantly negative (−0.155), and its autocorrelation coefficients are positive, indicating that MRII has an insufficient reaction to the stock price, and the information in media risk has some continuity. In the regression of MRSI, the second lag of returns is significantly positive (0.146), and negative autocorrelation coefficients are observed as well (−0.098), suggesting that MRSI has an overreaction to the stock price.

In conclusion, whether media risk has an underreaction or overreaction on stock prices depends on the balance between the information effect and the sentiment effect it contains. The information effect of the media indicates that stock prices have an insufficient reaction to absorb the information, while the sentiment effect suggests that the repeated hype in media coverage can cause drastic fluctuations and an excessive reaction in stock prices. Next, we will continue to explore its impact mechanism from the perspectives of investor attention and investor sentiment.

4.2 Investor Attention Effect

The transmission of information is related to attention, so it is necessary to consider the role of investor attention in the impact of media coverage on stock prices. According to Kahneman (1972), owing to attention biases in retail investors when new information becomes available, the investors' attention shifts to new information, leading to transactions based on it. The advent of Internet media has widened investors' access to information, but the abundance of mixed information has caused their attention to be dispersed. This observation is particularly applicable to emerging markets, where a substantial majority of retail investors cannot effectively analyze and process information. Consequently, valuable information contained in media reports does not promptly manifest in stock price.

We investigate the impact of MRII and MRSI on stock prices under different investor attentions. Following Dong et al. (2022), we measure investor attention (IA) based on the number of stock forum posts. we categorize all trading days into two distinct groups according to the level of investor attention. The first group encompasses those trading days with investor attention exceeding the average, whilst the second embodies those trading days where attention falls beneath this average. Accordingly, we replace MRI in equations (6) to (7) by MRII and MRSI, respectively, and apply vector autoregressive models to each group. The regression results are reported in Table 6 and Table 7.

In the regression of the rate of return in Table 6, we find that for the group with high investor attention, the lagged variables of MRII are not significant. In the group with lower investor attention, the first and second lagged variables of MRII are still insignificant, however the third lagged variable is significant −0.011 (0.011>0.006). Comparing the results in Table 5, the degree of underreaction of stock prices to MRII increases. This indicates that the lower the investor's attention to the company, the slower the information transmission speed of the company, and the greater the degree of underreaction. At the same time, we find that under high investor attention, the autocorrelation coefficients of return rate are all positive. However, under low investor attention, the first and second-order autocorrelation coefficients of return rate are negative. This suggests that stocks with low investor attention are more susceptible to media influence and experience larger fluctuations. In the regression of MRII, under high investor attention, media information is not influenced by past stock market, but under low investor attention, the lagged third-order variable of return rate is significant (-0.303), indicating that the information in media reports also has insufficient response to return rate at this time.

Table 6: Vector autoregression about the return and MRII according to investor attention groups

	Panel A: High IA		Panel B: Low IA	
	r_t	$MRII_t$	r_t	$MRII_t$
r_{t-1}	$\mathbf{0.014}^{**}$	-0.175	$\mathbf{-0.032}^{***}$	0.080
	(2.37)	(-0.48)	$\mathbf{(-9.47)}$	(1.03)
r_{t-2}	$\mathbf{0.013}^{***}$	-0.079	$\mathbf{-0.010}^{***}$	-0.013
	(2.69)	(-0.32)	$\mathbf{(-3.60)}$	(-0.18)
r_{t-3}	$\mathbf{0.041}^{***}$	0.029	$\mathbf{0.016}^{***}$	$\mathbf{-0.303}^{*}$
	(7.83)	(0.12)	**(5.35)**	$\mathbf{(-1.88)}$
r_{t-4}	$\mathbf{0.022}^{***}$	0.059	0.003	0.034
	(4.43)	(0.26)	(1.16)	(0.43)
$MRII_{t-1}(\mathrm{e}-2)$	-0.005	0.502^{***}	-0.001	0.373^{***}
	(-0.84)	(15.72)	(-0.11)	(18.78)
$MRII_{t-2}(\mathrm{e}-2)$	-0.002	-0.050^{*}	-0.002	-0.016
	(-0.30)	(-1.91)	(-0.31)	(-0.95)
$MRII_{t-3}(\mathrm{e}-2)$	-0.000	0.139^{***}	$\mathbf{-0.011}^{**}$	0.066^{***}
	(-0.07)	(5.32)	$\mathbf{(-1.95)}$	(4.29)
$MRII_{t-4}(\mathrm{e}-2)$	0.000	0.029	0.009	0.024
	(0.08)	(1.41)	(1.64)	(1.64)
N	42,918	42,918	502,580	502,580

t statistics in parentheses. * *p < 0.10,* ** *p < 0.05,* *** *p < 0.01*

Notes: This table is divided into two groups when Investor Attention (IA) is high in Panel A and low in Panel B.

Table 7: Vector autoregression about the return and MRSI according to investor attention groups

	Panel A: High IA		Panel B: Low IA	
	r_t	$MRSI_t$	r_t	$MRSI_t$
r_{t-1}	0.012^{*}	-0.443	$\mathbf{-0.030}^{***}$	$\mathbf{0.056}^{***}$
	(1.87)	(-1.23)	$\mathbf{(-9.77)}$	**(3.07)**
r_{t-2}	$\mathbf{0.014}^{**}$	$\mathbf{0.440}^{**}$	$\mathbf{-0.009}^{***}$	$\mathbf{0.092}^{***}$
	(2.55)	**(2.00)**	$\mathbf{(-3.56)}$	**(4.91)**
r_{t-3}	$\mathbf{0.042}^{***}$	-0.124	$\mathbf{0.019}^{***}$	$\mathbf{0.058}^{***}$
	(6.88)	(-0.55)	**(7.12)**	**(3.80)**
r_{t-4}	$\mathbf{0.021}^{***}$	0.065	0.004	0.020
	(3.80)	(0.31)	(1.46)	(1.39)
$MRSI_{t-1}(\mathrm{e}-2)$	-0.000	0.673^{***}	-0.111	0.813^{***}
	(-0.01)	(9.31)	(-0.96)	(18.51)
$MRSI_{t-2}(\mathrm{e}-2)$	0.007	-0.049	$\mathbf{-0.261}^{***}$	0.048^{*}
	(0.47)	(-1.04)	$\mathbf{(-3.64)}$	(1.84)
$MRSI_{t-3}(\mathrm{e}-2)$	-0.001	0.020	$\mathbf{-0.129}^{*}$	0.217^{***}
	(-0.06)	(0.41)	$\mathbf{(-1.70)}$	(7.80)
$MRSI_{t-4}(\mathrm{e}-2)$	$\mathbf{0.021}^{*}$	0.130^{***}	0.016	0.301^{***}
	(1.82)	(2.65)	(0.15)	(7.27)
N	42,918	42,918	502,580	502,580

t statistics in parentheses. * *p < 0.10,* ** *p < 0.05,* *** *p < 0.01*

Notes: This table is divided into two groups when Investor Attention (IA) is high in Panel A and low in Panel B.

In the regression of the rate of return in Table 7, for the high investor attention group, the fourth-order lagged variable of MRSI is significantly positive

(0.021>0.015), compared to the results in Table 5. This suggests that the degree to which the stock price overreacts to MRSI is more severe. However, for the low investor attention group, the second and third-order lagged variables of MRSI are significantly negative, indicating that investor attention can modulate the degree of overreaction of stock prices to the sentiment in the media. In the regression of MRSI, under the high investor group, the second-order lagged term of the return rate is significantly 0.440, which is greater than the sum of the return rate lagged terms under low investor attention (0.440>0.056+0.092+0.058). This demonstrates that the rate of return of stocks can also lead to an overreaction of media sentiment, and this overreaction is more pronounced when investor attention is high.

In summary, investor attention is an important avenue for regulating whether the stock price underreacts or overreacts to media reports. On one hand, the information in the media will be incorporated into the stock price more quickly due to investor attention, thereby reducing the degree of underreaction. On the other hand, the sentiment in the media can cause an overreaction in stock prices due to high investor attention.

4.3 Investor Sentiment Effect

Numerous studies have shown that investor decision-making is highly influenced by media sentiment (Henry, 2008; Ren et al., 2021; Fraiberger et al., 2021). Tetlock's (2007) findings indicate that media sentiment is accompanied by the regression of fundamentals, leading to the conclusion that media sentiment can serve as a representation of investor sentiment. The MRI we constructed not only includes information but also the sentiment of repetitive media coverage, therefore, from the channel of investor sentiment, we examine the impact of the information and sentiment in the MRI on stock prices.

We need to construct indicators that can directly reflect investor sentiment. Following Das et al. (2007), we adopt the sentiment of stock forum posts to construct investor sentiment. We measure negative investor sentiment (NIS) by subtracting the number of positive posts from the number of negative posts because MRI emphasizes negative media sentiment and accounts for the impact of repeated coverage. After building NIS, we test whether or not investor sentiment is one of the mediating factors that media affect stock prices through the regressions in equations (8)-(10).

$$r_{k,t} = \alpha + \beta_1 * MRII_{k,t} + \beta_2 * MRSI_{k,t} + Control + \varepsilon_{k,t}\ , \tag{8}$$

$$NIS_{k,t} = \alpha + \beta_1 * MRII_{k,t} + \beta_2 * MRSI_{k,t} + Control + \varepsilon_{k,t}\ , \tag{9}$$

$$r_{k,t} = \alpha + \beta_1 * MRII_{k,t} + \beta_2 * MRSI_{k,t} + \gamma * NIS_{k,t} + Control + \varepsilon_{k,t}\ . \tag{10}$$

Table 8 Testing the mediating effects of investor sentiment

	(1)	(2)	(3)
	r	NIS	r

MRII	**−0.038*****	0.772***	**−0.023*****
(e−2)	**(−17.38)**	(63.12)	**(−9.42)**
MRSI	**−0.053*****	0.284***	**−0.027*****
(e−2)	**(−20.78)**	(27.54)	**(−12.56)**
NIS			−0.038***
			(−165.77)
Control	Yes	Yes	Yes
Year FE	Yes	Yes	Yes
N	795,750	795,750	795,750
R^2	0.1875	0.0865	0.2164

t statistics in parentheses. * *p < 0.10,* ** *p < 0.05,* *** *p < 0.01*

Notes: This table reports the regression coefficients of equations (8)-(10). NIS refers to negative investor sentiment.

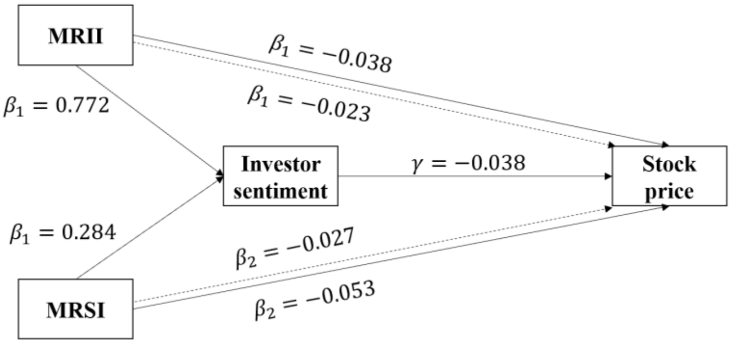


Figure 1 Mediating effect of investor sentiment

Notes: The regression results in Table 8 can be illustrated with a decomposition of investor sentiment, where the solid line represents the direct regression, and the dashed line represents the impact of MRII/MRSI on stock price after incorporating investor sentiment (NIS).

The results of Table 8 and Figure 1 indicate that investor sentiment constitutes a partial mediation effect of the media on stock prices. For the information part in the media, the coefficient of MRII changes from −0.038 to −0.023 in the regression with investor sentiment, and the t-value also changes from −7.38 to −9.24. For the sentiment part in the media, the coefficient of MRSI changes from −0.053 to −0.027 in the regression with investor sentiment, and the t-value also changes from −20.78 to −12.56. This indicates that the mediation role of investor sentiment in media sentiment is greater than the mediation role of media information.

We further examine the relationship between investor sentiment, stock prices, and media. Ren et al. (2021) propose two filtering mechanisms of social media on mass media reporting. The first mechanism is demand-driven media bias, where mass media news caters to reader preferences. The second mechanism is the mass media's correction of social media noise. The stock forum should also exhibit these two mechanisms as a form of social media. To investigate our conjecture, a vector auto-regression model from equations (11)-(13) is used to test the constructed NIS, MRSI, and return.

$$r_{k,t} = \alpha_1 + \beta_1 * L4(r_{k,t}) + \beta_2 * L4(MRSI_{k,t}) + \beta_3 * L4(NIS_{k,t})\,, \tag{11}$$

$$MRSI_{k,t} = \alpha_1 + \beta_1 * L4(r_{k,t}) + \beta_2 * L4(MRSI_{k,t}) + \beta_3 * L4(NIS_{k,t})\,, \tag{12}$$

$$NIS_{k,t} = \alpha_1 + \beta_1 * L4(r_{k,t}) + \beta_2 * L4(MRSI_{k,t}) + \beta_3 * L4(NIS_{k,t})\,. \tag{13}$$

Table 9: This table describes the vector autoregressive model of the four-day lagged return, Media Risk Sentiment Index (MRSI), and Negative Investor Sentiment (NIS)

	(1) r_t	(2) $MRSI_t$	(3) NIS_t
r_{t-1}	−0.007***	0.048	−1.079
	(−3.47)	(0.55)	(−0.99)
r_{t-2}	0.003*	0.187***	2.824***
	(1.89)	(3.94)	(2.99)
r_{t-3}	0.024***	0.046	2.375***
	(12.80)	(1.05)	(2.99)
r_{t-4}	0.012***	0.013	6.626***
	(6.64)	(0.31)	(8.91)
$MRSI_{t-1}(\mathrm{e}-2)$	**−0.038*****	0.711***	**0.651****
	(−3.78)	(13.13)	**(2.23)**
$MRSI_{t-2}(\mathrm{e}-2)$	−0.002	−0.099**	**−0.799****
	(−0.23)	(−2.49)	**(−2.59)**
$MRSI_{t-3}(\mathrm{e}-2)$	−0.017	0.059*	0.061
	(−1.57)	(1.69)	(0.27)
$MRSI_{t-4}(\mathrm{e}-2)$	**0.016*****	0.124***	−0.229
	(2.05)	(3.60)	(−1.39)
$NIS_{t-1}(\mathrm{e}-3)$	**−0.039*****	0.483	0.436***
	(−6.79)	(0.47)	(32.49)
$NIS_{t-2}(\mathrm{e}-3)$	**0.015*****	0.266	0.067***
	(2.78)	(0.43)	(4.95)
$NIS_{t-3}(\mathrm{e}-3)$	−0.009*	0.087	0.121***
	(1.80)	(0.14)	(10.07)
$NIS_{t-4}(\mathrm{e}-3)$	−0.006	0.206	0.094***
	(−1.42)	(0.37)	(10.21)
N	777,250	777,250	777,250

t statistics in parentheses. * *$p < 0.10$,* ** *$p < 0.05$,* *** *$p < 0.01$*

Notes: This table reports the regression coefficients of equations (11)-(13). NIS refers to negative investor sentiment. e−2 represents the reported numbers in the unit of 10^{-2}.

The regression results are shown in Table 9. In the stock return regression, the first lag of NIS is −0.039, and the second lag becomes 0.015. The first lag of MRSI is −0.038, and the fourth lag becomes 0.016. This indicates that both media sentiment and investor sentiment can cause an overreaction in stock prices, with investor sentiment reacting more quickly than media sentiment. In the regression of MRSI, none of the lag variables of NIS are significant. In the regression of NIS, the first lag variable of MRSI (0.651) and the second lag variable (−0.799) have opposite signs. This suggests that investor sentiment does not influence media sentiment, but media sentiment can cause an overreaction in investor sentiment. Moreover, the second lag coefficient of MRI displays an opposite sign of the autocorrelation coefficient of NIS, implying that media reports can partially rectify investors' irrational emotions. Additionally, our findings indicate that investor sentiment tends to overreact to changes in stock prices. Table 8 demonstrates a negative impact of the same-day stock prices on investor sentiment, whereas Table 9 reveals a positive influence of past stock price movements on investor sentiment. This observation suggests that when stock prices do not meet investor expectations, investor sentiment exhibits an exaggerated response to these price fluctuations.

We replace MRSI with MRII in regressions (11)-(13) to examine the impact of information in the media on investor sentiment. The regression results are shown in Table 10. In the regression of MRII, the first lag of NIS is significant at 1.603, indicating that media reports pay attention to the previous day's investor sentiment and incorporate it as information in their reporting. This aligns with Ren's proposed demand-driven media bias mechanism. In the regression of NIS, none of the lag variables of MRII are significant, suggesting that investor sentiment is not influenced by the information in the media but only reacts to media sentiment.

Table 10: This table describes the vector autoregressive model of the four-day lagged return, Media Risk Information Index (MRII), and Negative Investor Sentiment (NIS)

	(1) r_t	(2) $MRII_t$	(3) NIS_t
r_{t-1}	−0.007*** (−3.46)	−0.024 (-0.25)	−0.687 (−0.64)
r_{t-2}	0.003* (1.87)	−0.129 (−1.67)	2.956*** (3.14)
r_{t-3}	0.024*** (12.74)	−0.192** (−2.28)	2.56*** (3.23)
r_{t-4}	0.012*** (6.62)	0.076 (1.21)	6.610*** (8.82)
$MRII_{t-1}(\mathrm{e}-2)$	−0.006* (−1.74)	0.447*** (23.44)	0.066 (1.11)
$MRII_{t-2}(\mathrm{e}-2)$	0.002 (−0.73)	−0.014 (−1.11)	−0.043 (−0.63)
$MRII_{t-3}(\mathrm{e}-2)$	−0.002 (−0.47)	0.093*** (6.89)	0.045 (1.04)
$MRII_{t-4}(\mathrm{e}-2)$	0.004 (1.34)	0.048*** (4.33)	−0.002 (−0.05)
$NIS_{t-1}(\mathrm{e}-3)$	**−0.043***** **(−7.59)**	**1.602*** **(1.93)**	0.440*** (33.30)
$NIS_{t-2}(\mathrm{e}-3)$	**0.013**** **(2.47)**	−1.161 (−1.33)	0.063*** (4.79)
$NIS_{t-3}(\mathrm{e}-3)$	0.008 (1.49)	−0.282 (−0.41)	0.119*** (10.46)
$NIS_{t-4}(\mathrm{e}-3)$	-0.006 (-1.33)	−0.089 (−0.14)	0.092*** (10.28)
N	777,250	777,250	777,250

t statistics in parentheses. * *$p < 0.10$,* ** *$p < 0.05$,* *** *$p < 0.01$*

Notes: The model of this table is similar to formulas (11)-(13), the only difference is that MRSI is replaced by MRII. NIS refers to negative investor sentiment. e−2 represents the reported numbers in the unit of 10^{-2}.

In summary, we find that investor sentiment plays an intermediary role in the influence of media on stock prices. On one hand, the sentimental aspect of media coverage causes an overreaction in investor sentiment, resulting in an overreaction in stock prices. However, investor sentiment does not influence media sentiment. On the other hand, the informational aspect of media coverage caters to investor sentiment, but investor sentiment is not influenced by media information. Our results extend Tetlock's (2007) findings, which suggest that stock prices overreact to negative media reports. We further point out that this overreaction is caused by the interaction between media sentiment and investor sentiment, while media information causes an underreaction in stock prices due to investor attention.

We further validated two different mechanisms through which media influences investor sentiment and expanded on Ren et al.'s (2021) work. On one hand, media coverage incorporates past investor sentiment as information. On the other hand, media sentiment triggers an excessive reaction in investor sentiment, it also plays a role in correcting irrational investor sentiment. However, investor sentiment does not affect media sentiment. Notably, Ren et al. (2021) identify the evidence of the demand-driven media bias mechanism only, likely due to the fact that the users of Weibo in their study are not exclusively financial media audience members. In contrast, our study focuses on investors' exclusive social media platforms, specifically stock forum comments, thereby highlighting the corrective mechanism employed by media.

4.4 Implications

The previous analysis points out that MRI has two paths on stock price: investor attention effect and investor sentiment effect. In the following, we provide some implications for companies to manage information disclosure and market value and for investors to manage their investment.

4.4.1 Which companies are less affected by MRI?

Given the impact of MRI on stock price, how can companies greatly utilize the media as an intermediary of information and reduce the impact of media sentiment? In other words, which companies are less affected by media risks?

Peress and Fang (2009) conduct a related study in which they discover that less-publicized stocks significantly outperform those with high media attention in terms of future returns. They attribute this difference to Merton's investor cognition hypothesis (Merton, 1987), which suggests that in markets where information is incomplete, investors lack knowledge of all securities, leading to a need for higher returns from lesser-known stocks to compensate for imperfect portfolio diversification. In our hypothesis, investors with a higher level of cognition are more cable to recognize media hype sentiment. MRI will cause varying degrees of reactions to stock prices because of differing levels of investor cognition.

In Peress and Fang (2009), analyst coverage, individual ownership ratio, and idiosyncratic volatility are used as indicators of stock information level. Similarly, we also measure investor cognition by analyst attention, investor literacy, audit quality, and corporate financial transparency.

We conduct group regression analysis in equation (4) considering four perspectives: analyst coverage, investor literacy, audit quality, and transparency of financial reporting. Analysts can provide valuable information to investors through their research and reports, therefore companies that are followed by analysts are better understood by investors. We consider institutional investors as more rational investors in the market who can possess a deeper understanding of financial knowledge and have professional information processing capabilities. Hence, the percentage of institutional investors (POI) is used to represent investor literacy. Considering that the Big Four accounting firms supposedly have higher audit quality, we use whether the company is audited by international Big Four accounting firms (BIG4) to represent audit quality. In line with Hutton et al.'s (2009) methodology, we use the sum of controlled accrual items' absolute values over the previous three years to measure the transparency of financial reporting (Opaque) (see Appendix E for a detailed account of the construction process of Opaque). Considering space limitations, we primarily report the coefficients of the core variable MRI for different groups and quantify empirical results in Table 11.

The regression results validate that companies with a higher level of investor cognition are less affected by this media hype phenomenon on stock prices. This finding suggests that MRI encompasses not only information but also the role of sentiment, thereby providing evidence for the integration of media sentiment impact and information impact. To mitigate the influence of investor cognition on stock prices, we perform an additional interaction test (in Appendix F), confirming the hypothesis's consistency.

Table 11：Group regression from four perspectives, analyst coverage, investor literacy, audit quality, and transparency of financial reporting information transparency

	Analyst		RIO	
	Yes	No	High	Low
MRI(e-2)	-0.031^{***} (-9.26)	-0.098^{***} (-24.66)	-0.058^{***} (-16.00)	-0.066^{***} (-17.59)
Control	Yes	Yes	Yes	Yes
N	184,759	230,980	207,974	207,765
R^2	0.2221	0.1668	0.1935	0.1842
	BIG4		Opaque	
	Yes	No	High	No
MRI(e-2)	-0.025^{***} (-5.79)	-0.087^{***} (-26.19)	-0.101^{***} (-24.69)	-0.069^{***} (-22.53)
Control	Yes	Yes	Yes	Yes
N	19,818	354,163	208,114	316,360
R^2	0.1643	0.1989	0.1778	0.1790

t statistics in parentheses. * p < 0.10, ** p < 0.05, *** p < 0.01

Notes: The dependent variable is return, and the explanatory variable is MRI. Companies are divided into two groups based on whether there is analyst attention, the proportion of institutional investors, whether they are audited by the Big Four accounting firms, and the transparency of the company's financial reports. The coefficient sizes of the MRI for these two groups are then compared. e−2 represents the reported numbers in the unit of 10^{-2}.

This result provides guidance for listed companies' information disclosure and investor education. Companies can enhance investors' understanding by improving auditing quality and transparency in financial reports, thereby reducing the negative impact on stock prices from media's excessive hype and avoiding the phenomenon of media manipulating stock prices. Regulatory authorities can also monitor the market via MRI and improve investors’ financial literacy and analytical skills, thereby preventing price crashes caused by investor panic.

4.4.2 Portfolio analysis based on MRI

We have shown that the stock price underreacts to the MRI, which can be used to predict the next-day stock prices. To test the predictive power of the MRI, we construct a long-short portfolio based on the MRI and examine their effectiveness in asset pricing.

We sort the 250 companies from high to low based on their MRI in the previous day and divide them into five portfolios. Table 12 reports Jensen’s alpha and risk loadings for the MRI quintile portfolios under CAPM and Fama-French three-factor models. Notably, owing to MRI’s negative predictive effect on stock prices, we subtract the high group from the low group when constructing long-short portfolios, resulting in a high likelihood of achieving positive excess returns.

Table 12: CAPM and Fama-French alphas and risk loadings

Rank	Panel A	Panel B: CAPM		Panel C: Fama-French			
	$\alpha(\%)$	$\alpha(\%)$	β_{mkt}	$\alpha(\%)$	β_{mkt}	β_{smb}	β_{hml}
High	−0.032	**−0.062*****	0.754***	**−0.071*****	0.692***	0.453***	−0.102**
	(−0.88)	(−2.83)	(29.80)	(−3.48)	(25.36)	(9.57)	(−2.41)
2	0.009	−0.017	0.674***	−0.028*	0.603***	0.522***	−0.113***
	(0.29)	(−0.90)	(27.67)	(−1.65)	(23.91)	(12.88)	(−3.00)
3	0.020	−0.007	0.661***	−0.018	0.594***	0.521***	−0.070**
	(0.63)	(−0.40)	(32.21)	(−1.20)	(27.71)	(13.28)	(−2.03)
4	0.027	−0.002	0.758***	−0.017	0.671***	0.682***	−0.091***
	(0.78)	(−0.12)	(25.61)	(−0.97)	(22.46)	(18.10)	(−2.57)
Low	0.025	−0.004	0.729***	−0.017	0.623***	0.721***	−0.222***
	(0.71)	(−0.18)	(25.66)	(−0.96)	(22.10)	(16.51)	(−5.81)
Low-High	**0.057*****	**0.058*****	0.026	**0.053*****	−0.069***	0.268***	−0.120***
	(3.26)	(3.33)	(1.51)	(3.23)	(−4.11)	(9.82)	(−4.58)

t statistics in parentheses. * *$p < 0.10$,* ** *$p < 0.05$,* *** *$p < 0.01$*

Notes: This table sorts all the stocks into five groups based on the MRI of the previous day, and examines the abnormal returns of these five stock portfolios. Panel A directly examines the abnormal returns of different stock portfolios and long-short portfolios, whereas Panel B and Panel C take into account the CAPM model and the Fama three-factor model, respectively.

The empirical results in Table 12 demonstrate the predictive power of MRI. Although only the high group in Panel B and Panel C exhibits significant alpha, all three regressions generate substantial positive alphas for the long-short portfolios. The daily excess return of the long-short portfolio reaches more than 0.05% per day, or 12.5% per year (assuming there are 250 trading days in a year). Descriptive statistics reveal that MRI emerges at the 50th percentile, indicating that MRI is zero for more than half of the time for the companies. Therefore, no significant and evident decreasing trend of alpha is observed in the 3rd, 4th, and low quintiles. However, a significant negative alpha can be seen in the high and 2nd quintiles, with a greater negative magnitude in the high quintile. These findings suggest that the common factors cannot fully explain the return differences for different MRI portfolios. Investors can construct corresponding investor strategies through MRI to obtain excess returns.

5. Conclusions

This paper constructs a media risk index (MRI) specifically for Internet media. To construct the MRI, we first define media sentiments and risk features and use the AHP to distribute weights to various risks. Then the TF-IDF and one-hot methods are used to express news text as feature vectors. By using the Naive Bayes and Random Forest algorithms, we train machine learning models to extract sentiments and risk features from the news. Finally, the final MRI is constructed based on the agenda-setting theory.

We split the information and sentiment in MRI through media reporting intensity and based on this, we explore how the media influences stock prices through investor attention and investor sentiment. First, we find that whether stock prices underreact or overreact to media reports depends on the balance between information channel and sentimental channel. When information predominates, owing to investors' limited attention, the transmission of low-attention stock information is slower and the degree of underreaction is greater. When sentiment predominates, the media's hyping of sentiment can trigger overreaction from investor sentiment, which also leads to overreaction in stock prices. Second, our findings suggest that MRI's influence is diminishing for the firms that attract greater analyst interest and possess superior investor literacy, high audit quality, and extensive financial transparency. This result implies that the higher the level of investor recognition, the lower the impact of media hype on stock prices. Finally, we design MRI-guided long-short portfolios and show that the portfolio has a potential to yield over 12.5% in annualized excess returns.

We also provide managerial implications for listed companies, regulators, and investors. Our study suggests that listed companies can manage their market value by constructing MRI. Therefore, companies can mitigate the negative impact of media risk on stock prices by enhancing transparency in their financial information disclosure, improving audit quality, and guiding the media to provide objective and factual reports. Next, regulatory agencies can use MRI to monitor and prevent stock price crashes caused by sensational media coverage. They can also oversee listed companies' information disclosure through the media and strengthen financial literacy education for small and medium investors. These actions can reduce mispricing and improve market efficiency in pricing and resource allocation. Finally, our findings suggest a new investment strategy perspective that can utilize MRI to achieve excess returns. In the face of negative media sentiment, investors should exercise emotional control, avoid excessive influence from media reports, focus on company's actual operating conditions and financial information, and make rational investment decisions.

Acknowledgements

This research is sponsored by the National Natural Science Foundation of China (U1811462 and 71671191), and the Key Research and Development Project of Guangdong Province, China (2020B010110004).

Appendix A. Using TF-IDF to Identify Keywords for Risk Types

Using the TF-IDF approach, the top 20 keywords for each kind of risk are obtained after manual annotation of risk categorization. The results are displayed in Table 13. The overall keywords of each type of risk can reflect the risks contained even though only a few risk keywords are repeated. Operational risk is best illustrated by the terms “product,” “project,” and “business,” which indicate that it primarily refers to issues with products and projects during a company’s operational process. Problems with the company’s financing and investment activities are indicated by the words “Honor”, “overdue”, “debtor”, and “investor”. In the case of accounting risk, “%” denotes the frequency with which media reports compare the current year to the previous one, including the growth of net profit, assets, and liabilities. Accounting terms and accounting indicators in financial statements, such as “asset-liability ratio”, “net profit”, and “goodwill” can directly explain the accounting risk of the company. Take note that the words “tax”, “deferred”, and “income tax” in accounting risk suggest that taxes are another significant financial issue that is prone to attracting attention. Media coverage of stock market risk focuses primarily on equity risks, such as “shareholder”, “overweight”, and “underweight”. The company’s mergers and acquisitions, such as “shell” listings and “acquisition” are also included. Media reports on “judgment” outcomes, “lawsuit”, “case”, and some dispute cases, which primarily include “illegal”, “contract dispute”, and bond debt “settlement” types, are the focus of legal policy risk. Finally, among the other risks, the effect of the COVID-19, also known as “Hubei Province”, “pneumonia”, and “coronavirus” is the most obvious type of risk.

Table 13: The top 20 keywords for every risk type.

Risk type	Keywords (Chinese/English)
Operational risk	[‘产品’, ‘项目’, ‘金融’, ‘保理’, ‘投资者’, ‘供应链’, ‘兑付’, ‘资产’, ‘底层’, ‘问题’, ‘收购’, ‘业务’, ‘逾期’, ‘债务人’, ‘规模’, ‘风险’, ‘投资人’, ‘股权’, ‘系列产品’, ‘资金’]
	[‘product’, ‘project’, ‘finance’, ‘factoring’, ‘investors’,’ supply chain’, ‘honour’, ‘assets’,’ bottom’, ‘problem’, ‘acquisition’, ‘business’, ‘overdue’, ‘debtors’,’ size ‘, ‘risk’, ‘investors’,’ equity’, ‘series products, ‘money’]
Accounting	[‘%’, ‘负债’, ‘评级’, ‘资产负债率’, ‘总额’, ‘资产’, ‘融资’, ‘纳税’, ‘资金’, ‘递延’,

risk	‘所得税’, ‘半年报’, ‘税款’, ‘税局’, ‘债务’, ‘发行’, ‘披露’, ‘净利润’, ‘商誉’, ‘营收’] [‘%’, ‘debt’, ‘ratings’,’ asset-liability ratio’, ‘total’, ‘assets’,’ finance’, ‘tax’, ‘money’, ‘deferred’, ‘income tax’, ‘semiyearly report’, ‘tax’, ‘tax bureau’, ‘debt’, ‘issue’, ‘disclosure’, ‘profit’, ‘goodwill’, ‘revenue’]
Stock market risk	[‘增持’, ‘减持’, ‘股份’, ‘股东’, ‘股权’, ‘回购’, ‘转让’, ‘持有’, ‘公告’, ‘减值’, ‘上市公司’, ‘壳’, ‘业绩’, ‘质押’, ‘总股本’, ‘收购’, ‘控制权’, ‘股票’, ‘变动’, ‘重组’] [‘overweight’, ‘underweight’, ‘shares’, ‘shareholders’, ‘equity’, ‘buy back’, ‘transfer’, ‘hold’, ‘announcements’, ‘decrease in value’, ‘listed companies’, ‘shell’, ‘performance’, ‘pledge’, ‘total equity’, ‘acquisition’, ‘control’, ‘stocks’, ‘change’, ‘restructuring’]
Legal policy risk	[‘判决’, ‘担保’, ‘违规’, ‘诉讼’, ‘案件’, ‘承担’, ‘法院’, ‘责任’, ‘纠纷’, ‘合同’, ‘民法’, ‘签订’, ‘依约’, ‘事项’, ‘上诉’, ‘清偿’, ‘合同纠纷’, ‘被告’, ‘债权人’, ‘开庭’] [‘judgment’, ‘guarantee’, ‘illegal’, ‘lawsuit’, ‘case’, ‘undertake’, ‘court’, ‘responsibility’, ‘dispute’, ‘contract’, ‘civil law’, ‘sign’, ‘by appointment’, ‘items’, ‘appeal ‘, ‘settlement’, ‘contract dispute, ‘the accused’, ‘creditors’, ‘hold a court’]
Other risks	[‘慈善’, ‘捐赠’, ‘企业’, ‘中国’, ‘信托业’, ‘参与’, ‘供应链’, ‘物资’, ‘发展’, ‘湖北省’, ‘防控’, ‘肺炎’, ‘国际’, ‘金融’, ‘募集’, ‘产业’, ‘管理’, ‘地块’, ‘经济’, ‘冠状病毒’] [‘charity’, ‘donation’, ‘enterprises’, ‘China’, ‘trust business’, ‘participation’, ‘supply chain’, ‘goods’, ‘development’, ‘Hubei province’, ‘prevention and control’, ‘pneumonia’, ‘international’, ‘finance’, ‘raised’, ‘industry’, ‘management’, ‘site’, ‘economy’, ‘Coronavirus ‘]

Notes: The text news data in this table comes from all the news of a listed company, as detailed in Appendix B. We manually tag news of different risk types and calculate the top 20 keywords in all news of the same risk type using the TF-IDF method.

Appendix B. Train Machine Learning Model

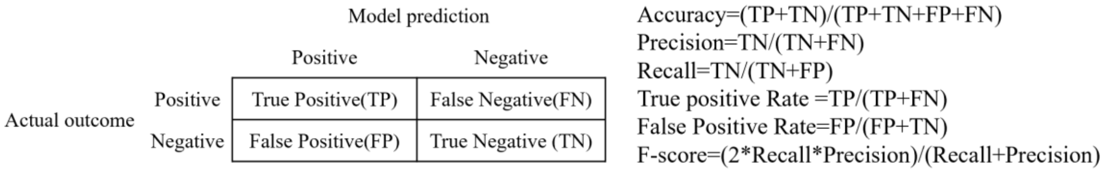
We take a listed company in Guangdong Province, Cedar Holdings, as an example and manually annotate all its related news. Bulk commodities, the chemical sector, industrial investment, financial trust, and other industries are all part of Cedar Holdings’ business. However, Cedar Holdings frequently receives negative reports about its operations, which sometimes results in the suspension of listed companies for correction. As a result, we choose Cedar Holdings as a representative example. The descriptive statistics for the labeled data are displayed in Table 14.

Table 14: Manually annotated statistical results

Sentiment classification statistics	
Negative:1572	Positive:790
	Total:2542
Risk classification statistics	
Operational risk:561	Accounting risk:947
Stock market risk:197	Legal policy risk:382
Other risks:455	Total:2542

Model Performance

The performance of the model should be assessed on the test set once it has been trained using machine learning on the training set. By selecting various thresholds, we can produce various prediction outputs given a probability value between 0 and 1. The expected and actual data can serve as the foundation for building the confusion matrix depicted in Figure 2. In addition, the following metrics are calculated: F-score, Accuracy, Precision, Recall, True Positive Rate, and False Positive Rate. Accuracy is an assessment of overall predictability. To assess the regional effects, we also use other indicators. The F-score is the harmonic mean of accuracy and recall. To get different True positive rates and False positive rates, different thresholds can be set. By plotting the True Positive Rates against the False Positive Rates on a coordinate axis, we obtain the ROC curve. The AUC score, which is located below the ROC curve, can also be used to assess the impact of the model.



		Model prediction	
		Positive	Negative
Actual outcome	Positive	True Positive(TP)	False Negative(FN)
	Negative	False Positive(FP)	True Negative (TN)

Accuracy=(TP+TN)/(TP+TN+FP+FN)
Precision=TN/(TN+FN)
Recall=TN/(TN+FP)
True positive Rate =TP/(TP+FN)
False Positive Rate=FP/(FP+TN)
F-score=(2*Recall*Precision)/(Recall+Precision)

Figure 2: Performance statistics. Sample confusion matrix and definition of our performance statistics.

Sentiment Classification

For the classification modeling of positive and negative sentiments, we use the titles of media reports. We choose our keywords using the TF-IDF method and create a dictionary of all documents using the word bag model. In order to obtain the vector representation of each document in the same spatial dimension, we use one-hot coding to mark each document. There are 41,461 title words in the labeled text, according to word segmentation statistics. To further extract the important details of each title, the top 10 keywords are extracted using TF-IDF. Then, a word bag model is built, and one-hot encoding is carried out. 1800 words are used to create the word bag. That is, a vector with 1800 dimensions is used to represent each document.

We use a 3:1 division to separate the labeled data into test and training sets. Cross-validation is done using the retention method to ensure the results are reliable. We mark negative sentiment as 1 and positive sentiment as 0 to better determine the specific judgment score. The outcome of the prediction is a continuous score between 0 and 1, which represents the likelihood of a negative prediction.

The prediction is made on the test set after the training is finished on the training set. As a start, 0.5 is chosen as the classification threshold value. Prediction results are therefore categorized as positive or negative depending on whether they are above or below 0.5. Second, we can create a confusion matrix by comparing it to the actual label. Through the use of the confusion matrix, the Accuracy, Precision, Recall, and F1-score are all calculated. Third, in accordance with the various thresholds chosen, we compute the True Positive and False Positive rates and measure AUC values. Table 15 presents the outcomes of machine learning.

The results demonstrate the usefulness of the Naive Bayes model in text analysis, with all indexes over 90%, an accuracy rate of 92.73%, and an AUC value of 97.5%, making it the best machine learning algorithm overall (Jegadeesh & Wu, 2013; Antweiler & Frank, 2004; Das & Chen, 2007; Li, 2010). With only 78.78% accuracy, the SVM model performs the worst. SVM is ineffective for text classification of sparse vectors, which may be the cause. With an accuracy of 88.21%, Decision Tree, the most fundamental machine learning algorithm, provides incredibly reliable results. The results of Random Forest, a traditional ensemble learning algorithm based on bagging, are superior to those of the Decision Tree, with an accuracy of 90.37%. Adaboost, which belongs to the classical ensemble learning algorithm boosting class, performs less accurately than Decision Tree, with an accuracy of only 83.89%. Adaboost's excessive focus on a training set accuracy and overfitting in the test set are the causes of this.

Table 15: Results of using different machine learning models to judge the positive and negative sentiment of media reports.

	Accuracy	Precision	Recall	F1-score	AUC
NB	**0.9273**	**0.9073**	**0.9278**	**0.9164**	**0.9750**
DT	0.8821	0.8880	0.8282	0.8502	0.9313
RF	0.9037	0.8993	0.8691	0.8821	0.9575
Adaboost	0.8389	0.8106	0.8479	0.8220	0.9244
SVM	0.7878	0.8647	0.6552	0.6714	0.9275

Notes: The machine learning models in the table include Naive Bayes (NB), Decision Tree (DT), Random Forest (RF), Adaboost and Support Vector Machine (SVM).

Risk Classification

We use the content of articles as a data source to categorize media risk. The labeled article contains 2,983,803 words in total after we segment it all. We use the same technique as sentiment prediction to create the space vector for each document. Each article contains more than 500 words, so in order to more effectively sum up what each one is about, we use the TF-IDF method to extract the top 50 keywords from each one and combine them all into a word bag model. Our word bag, which has 7,006 keywords, is the final result. Each article's space vector is represented using one-hot encoding. To ensure the robustness of the model, we continue to divide all labeled data into training and test sets in a 3:1 ratio and perform cross-validation using the leave-one-out method. The test set is put to the test using the trained model from the training set. The outcomes are displayed in Table 16.

Table 16: Results of using different machine learning models to judge risk type of media reports.

	Accuracy	Precision	Recall	F1-score
NB	0.8527	0.8180	0.8482	0.8256
DT	0.8409	0.8017	0.8205	0.8090
RF	**0.8802**	**0.8624**	**0.8745**	**0.8675**
Adaboost	0.7466	0.7862	0.7081	0.7153
SVM	0.6189	0.8264	0.4806	0.5315

Notes: The machine learning models in the table include Naive Bayes (NB), Decision Tree (DT), Random Forest (RF), Adaboost and Support Vector Machine (SVM).

In multi-class classification models, the Random Forest (RF) model outperforms the Naive Bayes model and Decision Tree model, with all evaluation indexes over 85%. Therefore, we choose to use the RF model for the subsequent expansion of risk prediction. The SVM model and Adaboost model performed poorly, especially the SVM model with a recall of only 48.06%, indicating the presence of many false positive samples.

Appendix C. Variable Definition and Descriptive Statistics

Table 17 gives the names, definitions and sources of all the variables involved in the paper.

Table 17: Variable description

Variable	Definition	Source
r(%)	The daily return of a listed company’s stock	CSMAR
MRI	The daily Media Risk Index of a listed company. See section 3 for detailed construction steps.	ChinaScope
MRII	The daily Media Risk Information Index of a listed company. See section 3 for detailed construction steps.	ChinaScope
MRSI	The daily Media Risk Sentiment Index of a listed company. See section 3 for detailed construction steps.	ChinaScope
Operational risk	Daily operational risk values in the company’s news. See Table 1 for detailed construction steps.	ChinaScope
Accounting risk	Daily accounting risk values in the company’s news. See Table 1 for detailed construction steps.	ChinaScope
Stock market risk	Daily stock market risk values in the company’s news. See Table 1 for detailed construction steps.	ChinaScope
Legal policy risk	Daily legal policy risk values in the company’s news. See Table 1 for detailed construction steps.	ChinaScope
Other risks	Daily other risks values in the company’s news. See Table 1 for detailed construction steps.	ChinaScope
MKT_RF(%)	Market risk factor.	CSMAR
SMB(%)	Size risk factor.	CSMAR
HML(%)	Market-to-value risk factor.	CSMAR
BIG4	Dummy variable is 1 if the company is audited by a Big Four accounting firm in the year, and 0 otherwise.	CSMAR
SOE	Dummy variable is 1 if the company is a state-owned enterprise in the year, and 0 otherwise.	CSMAR
TOP5(%)	The proportion of the top five shareholders of the company in the year.	CSMAR
POI(%)	The proportion of institutional investors in the company during the year.	CSMAR
SIZE	The size of the company for that year is represented by the logarithm of total assets.	CSMAR
BM	The book-to-market ratio of the company for the year.	CSMAR
ROA	The return of asset ratio of the company for the year.	CSMAR

LEV	The leverage ratio of the company for the year.	CSMAR
Tpostnum	The daily post volume of the company’s stock bar.	CNRDS
Pospostnum	The daily positive post volume of the company’s stock bar.	CNRDS
Negpostnum	The daily negative post volume of the company’s stock bar.	CNRDS

Table 18 shows the descriptive statistics of the data sample used in our study. The minimum value of the company’s stock return is −80.732% and the maximum value is 47.058% which indicates that the listed companies that have been penalized are more likely to experience extreme situations. The maximum value of MRI is 204.789, and its 25th and 50th percentile are both 0, which indicates that negative news has the characteristics of concentrated outbreaks and rapid dissemination. Since the MRI is composed of MRII and MRSI, and MRII is the residual term of MRI regressed on the intensity of media reporting, the average value of MRII is 0 and the mean of MRSI is equal to the mean of MRI. Each of the five media risks in the 25th, 50th, and 75th percentile is 0, indicating that each type of media risk presents a relatively sparse feature. The mean of BIG4 is 0.048, indicating that 4.8% of companies are audited by a Big Four accounting firm. The mean of SOE is 0.254, indicating that about 25% of companies are state-owned enterprises. The mean ROA is −0.069, suggesting that the penalized companies have relatively poor profitability.

Table 18: Summary statistics

	Obs.	Mean	Std. Dev.	min	max	25%	50%	75%
r(%)	795,750	0.025	0.037	−80.732	47.058	−1.615	0.040	1.567
MRI	795,750	0.294	1.784	0	204.789	0	0	0.014
MRII	795,750	0.000	1.362	−123.806	202.366	−0.156	−0.044	0
MRSI	795,750	0.294	1.151	−2.705	180.978	0.016	0.076	0.232
Operational risk	795,750	0.030	0.152	0	19.040	0	0	0
Accounting risk	795,750	0.075	0.341	0	64.473	0	0	0
Stock market risk	795,750	0.116	0.425	0	48.667	0	0	0
Legal policy risk	795,750	0.125	0.476	0	57.128	0	0	0
Other risks	795,750	0.041	0.194	0	23.14	0	0	0
MKT_RF(%)	2,148	0.040	0.016	−9.478	9.165	−0.626	0.118	0.802
SMB(%)	2,148	0.028	0.008	−5.848	4.051	−0.365	0.091	0.482
HML(%)	2,148	0.011	0.007	−4.267	3.615	−0.461	−0.019	0.420
BIG4	1,972	0.048	0.214	0	0	0	0	0
SOE	1,972	0.254	0.436	0	1	0	0	1
TOP5(%)	1,972	49.006	15.892	6.907	97.456	37.519	48.633	59.369
POI(%)	1,972	40.846	23.034	0.001	98.125	22.685	40.098	57.745
SIZE	1,972	22.104	1.533	16.649	29.218	21.183	21.880	22.783
BM	1,972	0.391	0.646	−2.582	14.022	0.159	0.295	0.498
ROA	1,972	−0.069	1.302	−48.316	1.408	−0.012	0.017	0.045
LEV	1,972	2.155	9.116	−56.445	275.340	0.386	0.883	2.020
Analyst	1,972	0.951	1.147	0	4	0	0	2
Tpostnum	795,750	42.837	109.922	1	18,855	6	17	43

Pospostnum	795,750	10.989	26.532	0	3919	2	5	11
Negpostnum	795,750	9.596	27.210	0	4839	1	3	10

Notes: Among all statistics, daily degree panel data include: r, MRI, operational risk, accounting risk, stock market risk, legal policy risk, and other risks, Tpostnum, Pospostnum, Negpostnum. Annual panel data include: BIG4, SOE, TOP, POI, SIZE, BM, ROA, LEV and Analyst. Daily time series data include: MKT_RF, SMB and HML.

Appendix D. A Supplementary Analysis on Underreaction

We examine the sustained relationship between MRI and the company's stock return. We report the regression coefficients of the stock price returns for the next ten trading days with MRI in Table 19, showing that the regression coefficients of MRI for the next 1-10 trading days are negative and significant, providing evidence of investors underreacting to negative media sentiment. The prediction influence of MRI on future returns is -0.039 for the first trading day, which is lower than the impact of MRI on the stock price on the day in Table 3 (-0.064). As time goes by, this negative impact gradually decreases and becomes less significant, indicating that the MRI causes an underreaction of stock prices.

Table 19: Overreaction and underreaction test: regression of the stock price returns for the next ten trading days with MRI.

	$r_{i,t+1}$	$r_{i,t+2}$	$r_{i,t+3}$	$r_{i,t+4}$	$r_{i,t+5}$
MRI(e−2)	-0.039^{***} (-13.57)	-0.025^{***} (-8.73)	-0.022^{***} (-7.44)	-0.017^{***} (-5.91)	-0.016^{***} (-5.71)
Control	Yes	Yes	Yes	Yes	Yes
N	415,489	415,239	414,989	414,739	414,489
R^2	0.0030	0.0019	0.0020	0.0019	0.0020
	$r_{i,t+6}$	$r_{i,t+7}$	$r_{i,t+8}$	$r_{i,t+9}$	$r_{i,t+10}$
MRI(e−2)	-0.015^{***} (-5.46)	-0.014^{***} (-4.89)	-0.012^{***} (-4.12)	-0.007^{***} (-2.46)	-0.012^{*} (-1.62)
Control	Yes	Yes	Yes	Yes	Yes
N	414,239	413,989	413,739	413,489	413,239
R^2	0.0017	0.0026	0.0019	0.0023	0.0024

t statistics in parentheses. * p < 0.10, ** p < 0.05, *** p < 0.01

Notes: The explained variable is the stock return rate in the next ten trading days, and the explanatory variable is the MRI of that day. $e-2$ represents *the reported numbers in the unit of* 10^{-2}.

Appendix E. Construction of Financial Report Transparency

Based on the methodology of Hutton et al. (2009), we utilize the sum of absolute values of a company's operating accruals over the past three years ($Opaque$) to measure the transparency of financial information. A higher Opaque indicates lower transparency in financial information of the company. The manipulation of accruals ($DisAcc$) is estimated by the modified Jones model. The formulas for the respective variables are expressed as follows,

$$Opaque = Abs(DisAcc_{t-1}) + Abs(DisAcc_{t-2}) + Abs(DisAcc_{t-3})$$

$$\frac{TA_{k,t}}{Asset_{k,t-1}} = \alpha_1 \frac{1}{Asset_{k,t-1}} + \alpha_2 \frac{\Delta REV_{k,t}}{Asset_{k,t-1}} + \alpha_3 \frac{PPE_{k,t}}{Asset_{k,t-1}}$$

$$DisAcc_{k,t} = \frac{TA_{k,t}}{Asset_{k,t-1}} - (\alpha_1 \frac{1}{Asset_{k,t-1}} + \alpha_2 \frac{\Delta REV_{k,t} - \Delta REC_{k,t}}{Asset_{k,t-1}} + \alpha_3 \frac{PPE_{k,t}}{Asset_{k,t-1}})$$

TA represents the total accruals, which is equal to operating income minus cash flow from operations. $Asset$ represents the total assets, ΔREV represents the growth in sales revenue, ΔREC represents the growth in accounts receivable, and PPE represents the original cost of fixed assets.

Appendix F. Interaction Term Test of Investor Cognition and MRI

In the interaction test, variables including investor cognition ($Cognition$), MRI, and the interaction term between MRI and investor cognition are simultaneously included in the regression equation, as shown in the following equation. Since MRI has a negative impact on stock prices, if the coefficient of the interaction term between MRI and investor cognition is positive, it indicates that investor cognition does indeed weaken the negative media effect. Similarly, using analyst attention (Analyst), investor literacy (RIO), audit quality (BIG4), and corporate financial transparency (Opaque) as proxies for investor cognition level, the regression results, as shown in Table 20, provide additional evidence for Hypothesis 6,

$$r_{k,t} = \alpha + \beta_1 * MRI_{k,t} + \beta_2 * Cognition_{k,t} + \beta_3 * Cognition_{k,t} * MRI_{k,t} + control$$

Table 20：Interaction term test of investor cognition and MRI

r	(1) Analyst	(2) RIO	(3) BIG4	(4) Opaque
MRI	-0.102^{***}	-0.074^{***}	-0.069^{***}	-0.064^{***}
(e−2)	(−27.40)	(−17.18)	(−24.60)	(−23.83)
Analyst	-0.044^{***}			
(e−2)	(−6.55)			
Analyst* MRI	**0.0237*****			
(e−2)	(13.97)			
RI-		−0.004		
(e−3)		(−1.18)		
RIO * MRI		**0.003*****		
(e−3)		(2.71)		
BIG4			1.663^{***}	
(e−2)			(14.89)	
BIG4 * MRI			**0.032*****	
(e−2)			(4.16)	
Opaque				−0.002
(e−2)				(−1.28)
Opaque * MRI				0.000
(e−2)				(−0.74)

Control	Yes	Yes	Yes	Yes
N	415,739	415,739	415,739	415,739
R^2	0.1881	0.1878	0.1878	0.1877

t statistics in parentheses. * *p < 0.10,* ** *p < 0.05,* *** *p < 0.01*

Notes: The explained variable is return, and the explanatory variable is the proxy variable representing the degree of investor cognition and its interaction term with MRI. We use analyst attention (Analyst), investor literacy (RIO), audit quality (BIG4), and corporate financial transparency (Opaque) as proxies for investor cognition level. $e{-}2$ *represents the reported numbers in the unit of* 10^{-2}.

References

Antweiler, W., & Frank, M. Z. (2004). Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards. *The Journal of Finance* (No.3), 1259-1294.

Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology* (No.6), 1173-1182

Biktimirov, E. N., Sokolyk, T., & Ayanso, A. (2021). Sentiment and hype of business media topics and stock market returns during the COVID-19 pandemic. *Journal of Behavioral and Experimental Finance*, 31, 100542.

Birru, J., & Young, T. (2022). Sentiment and uncertainty. *Journal of Financial Economics*, 146(3), 1148-1169.

Chan, W. S. (2003). Stock price reaction to news and no-news: drift and reversal after headlines. *Journal of Financial Economics* (No.2), 223-260.

Cao, R. Q., Schniederjans, D. G., & Gu, V. C. (2021). Stakeholder sentiment in service supply chains: big data meets agenda-setting theory. *Service Business*, 15(1), 151-175.

Cohen, B. C. (1963). *The Press and Foreign Policy*. Princeton: Princeton University Press.

Costola, M., Hinz, O., Nofer, M., & Pelizzon, L. (2023). Machine learning sentiment analysis, COVID-19 news and stock market reactions. *Research in International Business and Finance*, 64, 101881.

Das, S. R. D. O., & Chen, M. Y. L. L. (2007). Yahoo! for Amazon: Sentiment Extraction from Small Talk on the Web. *Management Science* (No.9), 1375-1388.

Davis, A. K., Piger, J. M., & Sedor, L. M. (2012). Beyond the Numbers: Measuring the Information Content of Earnings Press Release Language. *Contemporary Accounting Research* (No.3), 845-868.

Delong, J. B., Shleifer, A., Summers, L. H., & Waldmann, R. (1990). Noise Trader Risk in Financial Markets. *Journal of Political Economy* (No.4), 703-738.

Dong, D., Wu, K., Fang, J., Gozgor, G., & Yan, C. (2022). Investor attention factors and stock returns: Evidence from China. *Journal of International Financial Markets, Institutions and Money*, 77, 101499.

Fama, E. F., & French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics* (No.1), 3-56.

Fama, E. F., & MacBeth, J. D. (1973). Risk, Return, and Equilibrium: Empirical Tests. *Journal of Political Economy*, 81(3), 607-636

Fraiberger, S. P., Lee, D., Puy, D., & Ranciere, R. (2021). Media sentiment and international asset prices. *Journal of International Economics*, 133, 103526.

Frank, M. Z., & Sanati, A. (2018). How does the stock market absorb shocks? *Journal of Financial Economics* (No.1), 136-153.

Funkhouser, G. R. (1973). The Issues of the Sixties: An Exploratory Study in the Dynamics of Public Opinion. *Public Opinion Quarterly* (No.1), 62-75.

Garcia, D. (2013). Sentiment during Recessions. *Journal of Finance*, 68(3), 1267-1300.

Guégan, D., & Renault, T. (2021). Does investor sentiment on social media provide robust information for Bitcoin returns predictability? *Finance Research Letters*, 38, 101494.

Henry, E. (2008). Are investors influenced by how earnings press releases are written? *Journal of Business Communication* (No.4), 363-407.

Hillert, A., Jacobs, H., & Müller, S. (2014). Media Makes Momentum. *The Review of Financial Studies* (No.12), 3467-3501.

Hutton, A. P., Marcus, A. J., & Tehranian, H. (2009). Opaque financial reports, R^2, and crash risk. *Journal of Financial Economics*, 94(1), 67-86.

Jegadeesh, N., & Wu, D. (2013). Word power: A new approach for content analysis. *Journal of Financial Economics* (No.3), 712-729.

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. Cognitive Psychology, 3(3), 430-454.

Kim, J. J., Dong, H., Choi, J., & Chang, S. R. (2022). Sentiment change and negative herding: Evidence from microblogging and news. *Journal of Business Research*, 142, 364-376.

Liang, C., Tang, L., Li, Y., & Wei, Y. (2020). Which sentiment index is more informative to forecast stock market volatility? Evidence from China. *International Review of Financial Analysis*, 71, 101552.

Leo, M., Sharma, S., & Maddulety, K. (2019). Machine Learning in Banking Risk Management: A Literature Review. *Risks* (No.1), 1-22.

Li, F. (2010). The information content of forward-looking statements in corporate filings-A naïve bayesian machine learning approach. *Journal of Accounting Research* (No.5), 1049-1102.

Lippmann, W. (1922). *Public Opinion*. New York: Macmillan.

Loughran, T., & Mcdonald, B. (2011). When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks. *Journal of Finance* (No.1), 35-65.

Maghyereh, A., & Abdoh, H. (2022). Can news-based economic sentiment predict bubbles in precious metal markets? *Financial Innovation*, 8(1), 35.

Manela, A. A., & Moreira, A. B. (2017). News implied volatility and disaster concerns. *Journal of Financial Economics* (No.1), 137-162.

Merton, R. C. (1987). A Simple Model of Capital Market Equilibrium with Incomplete Information. *Journal of Finance* (No.3), 483-510.

Pathak, A. R., Pandey, M., & Rautaray, S. (2021). Topic-level sentiment analysis of social media data using deep learning. *Applied Soft Computing*, 108, 107440.

Peress, J. (2014). The Media and the Diffusion of Information in Financial Markets: Evidence from Newspaper Strikes. *The Journal of Finance* (No.5), 2007-2043.

Peress, J., & Fang, L. (2009). Media Coverage and the Cross-section of Stock Returns. *The Journal of Finance* (No.5), 2023-2052.

Ren, J., Dong, H., Padmanabhan, B., & Nickerson, J. V. (2021). How does social media sentiment impact mass media sentiment? A study of news in the financial markets. *Journal of the Association for Information Science and Technology*, 72(9), 1183-1197.

Savor, P. G. (2012). Stock returns after major price shocks: The impact of information. *Journal of Financial Economics*, 106(3), 635-659.

Schumaker, R. P., Zhang, Y., Huang, C., & Chen, H. (2012). Evaluating sentiment in financial news articles. *Decision Support Systems*, 53(3), 458-464.

Tetlock, P. C. (2007). Giving Content to Investor Sentiment: The Role of Media in the Stock Market. *The Journal* *of Finance* (No.3), 1139-1168.

Tetlock, P. C. (2011). All the News That's Fit to Reprint: Do Investors React to Stale Information? *The Review of Financial Studies*, 24(5), 1481-1512.

Tetlock, P. C., & Macskassy, M. S. A. S. (2008). More than Words: Quantifying Language to Measure Firms' Fundamentals. *The Journal of Finance*(No.3), 1437-1467.

Umar, Z., Adekoya, O. B., Oliyide, J. A., & Gubareva, M. (2021). Media sentiment and short stocks performance during a systemic crisis. *International Review of Financial Analysis*, 78, 101896.

Vaidya, O. S., & Kumar, S. (2006). Analytic hierarchy process: An overview of applications. *European Journal of Operational Research*, 169(1), 1-29.

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, 359(6380), 1146-1151.

Zhang, J. L., Härdle, W. K., Chen, C. Y., & Bommes, E. (2016). Distillation of News Flow into Analysis of Stock Reactions. *Journal of Business & Economic Statistics*, 34(4), 547-563.

Zou, L., Cao, K. D., & Wang, Y. (2019). Media Coverage and the Cross-Section of Stock Returns: The Chinese Evidence. *International Review of Finance*, 19(4), 707-729.