Discrete Dynamics in Nature and Society

News Sentiment and the Risk of a Stock Price Crash Risk: based on Financial Dictionary combined BERT-DCA

Shuyi Li,[1] Junhao Kong,[2]

[1] School of Economics, Zhejiang University, Hangzhou 310030, China.
[2] School of Economics, Zhejiang University, Hangzhou 310030, China.

Correspondence should be addressed to Junhao Kong; 12101011@zju.edu.cn

Abstract

This study combines a financial knowledge dictionary and pretraining method based on BERT (Bidirectional Encoder Representation from Transformers) to construct a deep learning model for identifying stock news sentiment. The study then calculates the sentiment metrics of all stocks and analyzes the impact of news sentiment on the risk of a stock price crash and its heterogeneity. The results show that stocks with more positive sentiment metrics have a higher risk of crash in the following year. We also investigate the information intermediation and investor sentiment channels by which news sentiment affects the risk of a crash. The results show that more net insider sales, lower information transparency, and less analyst coverage amplify the impact of news sentiment on future crash risk, which is consistent with the information intermediation channel. Additionally, more retail investor positions, more active investor sentiment, and divergence between analysts' opinions and news amplify the impact of news sentiment on the risk of a future stock price crash, consistent with the investor sentiment channel.

1.Introduction

Stability is a necessary condition for the financial market to efficiently facilitate economic development. In developing countries with weak financial infrastructure and low market efficiency, the risk of firm-specific stock price crash can seriously threaten investors and undermine their confidence in the stock market. The regularly used method of portfolio diversification cannot fully mitigate it, leading researchers to focus on the factors inducing stock price crash. A different strand of literature investigates the effects of media coverage or overall media tone on stock price crash risk, to which the present study belonged. We, however, adopted a slightly different perspective—the asymmetric impact of sentiment heterogeneity (i.e., positive or negative).

Our study aims to examine the asymmetric impact of news sentiment on stock price crash risk attributable to the role of media as information intermediary that influences investors' sentiment. On the one hand, regardless of the bias of media coverage, positive news will inevitably lead to investors' optimism. The media report positive news about some firm, according to the theory of the spiral of silence [36], pessimistic investors may remain silent and optimistic investors may dominate the market. When short selling is not allowed, pessimism about that firm cannot be hedged normally. Subsequently, a stock price crash will happen when the optimistic sentiment dissipates and pessimistic investors become the marginal buyers [21]. On the other hand, the management has every intention of defending the firm's image and concealing negative information from the public [22, 29], which unintendedly leads to an overvaluation of stock prices, resulting in price bubbles. When the management can no longer delay or conceal negative news and has to release it to the market in a short period, the pessimism of investors will be amplified, inducing a stock price crash [5, 24, 45, 46].

We distinguish among the sentiments of the media coverage and study the asymmetric impacts of positive and negative news on stock price crash risk. Thus, we need a precise measurement of the sentiment for objectively evaluating the tone and extent of the news. In this regard, a burgeoning body of finance and accounting literature has used natural language processing (NLP) algorithms to extract financial texts sentiment [8, 10, 31, 35]. The emergence of modern deep learning algorithms, such as Elmo, GPT (Generative Pre-Training), and Google BERT, makes NLP move up a gear in the accuracy of sentiment analysis. Related studies have suggested that relative to traditional methods, modern deep learning algorithms that combine prior knowledge in certain field have better performance. In this study, we customize a state-of-the-art deep learning NLP algorithm(BERT) for financial texts and document its advantages over traditional approaches.

We collect a total of 1,132,856 initial media coverage articles between 2011 and 2020 from the China Stock Market and Accounting Research database [50]. We manually construct our own sentiment dictionary in the financial domain and use it as a corpus for sentiment classification. Moreover, the BERT-based pretraining model is designed to help machines understand the characteristics of human language and extract sentiment information effectively. Based on the model, we classify the related news of each stock. Finally, our sample consisted of 17,267 firm-year observations representing 2,277 individual firms.

To begin with, we examine whether media sentiment is associated with firm-specific future price crash risk. We measure media sentiment from three dimensions: mixed (on average), positive, and negative sentiment. We define the indicators of the average sentiment, and those of the positive and negative sentiment, respectively. We use three proxies of stock price crash risk: the binary variable (CRASH) that equals 1 for a firm-year that experiences one or more crash weeks during the fiscal year and 0 otherwise; the negative coefficient of skewness of firm-specific weekly returns (NCSKEW); and the down-to-up volatility (DUVOL) of firm-specific weekly returns [6, 22, 27]. The results show that firms with more positive media coverage tend to have a higher risk of future stock price crash. Meanwhile, negative media coverage shows a limited effect on the risk of future stock price crash and a significant negative relation with current stock price crash risk, implying that they can lead to a stock price crash in the short term.

We then address the natural question of identifying the channels through which media sentiment affects the stock price crash risk. We hypothesize two possible channels, namely, information intermediation and investor sentiment, and then design various settings to examine them. The stock market is significantly impacted by the media, as an important information intermediary between firm management and market participants. On the one hand, media outfits disseminate value-relevant information on firms' current and future earnings to outside investors, reduce market frictions, improve investor perceptions, and mitigate information asymmetry [4, 6, 33]. On the other hand, media are not the perfect messenger. Media coverage is not always objective and neutral, but rather offers ambiguous, out-of-date, and even exaggerated and biased contents [7, 20]. According to previous findings, more net insider sales, lower information transparency, and less analyst coverage amplify the impact of media sentiment on future crash risk, which is consistent with the "information intermediation" channel.

Investors, owing to their limited attention and overconfidence, can be expected to overreact to catchy, anecdotal, less relevant information, but underreact to abstract, statistically listed, relevant information [37]. Furthermore, they may exhibit confirmation bias, which is the tendency to seek and believe information that supports one's beliefs while ignoring later signals that are inconsistent with their prior beliefs after developing a favorable impression of a firm [40]. As such, media coverage of firms (particularly positive news), regardless of whether the content is outdated or not, can easily pique the interest of investors, causing them to overreact or overestimate a firm's prospects and bring about a short- or long-term increase above the fundamental value [1, 2, 11]. However, when actual operational problems are revealed or a firm fails to meet expectations, negative sentiment will emerge and the stock price will reverse, resulting in a crash; this process will be reinforced when media coverage is biased and exaggerated. More retail investor positions, more active investor sentiment, and divergence between analysts' opinions and media coverage amplify the impact of news sentiment on future crash risk, consistent with the "investor sentiment" channel.

We expect to contribute to the literature in the following ways. First, our work also adds to the growing literature on the determinants of firm-specific price crash risk. Numerous studies have established a link between media sentiment and stock prices [13, 14, 40, 44, 45]. Sentiment in news articles contain novel information on stock prices [45, 46], but few studies have paid attention to the relation between media sentiment's asymmetric effect on the future firm stock price crash risk. We attempt to fill this gap. We provide the formal piece of empirical evidence that positive news sentiment predicts higher firm-specific future price crash risk, whereas negative sentiment increases current crash risk. We provide a thorough examination of the impact of media sentiment from both perspectives of inspiration of investor sentiment of media tone and information economics, revealing new evidence that investors' irrational and excessive optimism could be a major cause of stock price bubbles and crashes in China, where retail investors predominate and short selling is restricted.

Second, our research combines advanced deep learning and dictionary methods, which take full advantage of the performance and intelligence of computer technology and greatly improve the identification accuracy and efficiency of massive sentiment information. The sentiment dictionary approach uses word and syntactic analyses of text to calculate sentiment values as the basis for determining text sentiment tendencies. However, individuals can add necessary semantic words, such as praise words, degree adverbs, and negative words, which play an important role in enhancing or weakening sentiment semantic words [37]. Sentiment dictionary classification methods, which ignore the characteristics of language, such as grammar, context, and subjective construction methods, are likely to have the problem of omission. We attempt to mitigate this problem through integrating a cutting-edge deep learning method. To the best of our knowledge, this study is the first to combine deep learning and dictionary methods to perform sentiment analysis in the financial field, thereby extending the application of sentiment analysis methods in the financial field.

Third, our study contributes to the literature on text analysis in the economic field [16, 29]. We introduce BERT, a deep learning pretraining model, and combine it with a relatively mature sentiment dictionary. Using the BERT pretraining model, researchers can take full advantage of the contextual information in the news. The vector expression of the same words is different between news and contexts, which was difficult to address in previous studies. In the pretraining model, by pretraining large text corpora as a language model, we create embeddings for the context associations (embedding) of each word in a sentence, which could then be entered into subsequent tasks, thereby enabling a full quantification of the information contained in the text.

2.Theory and Hypothesis

The media, as an important vehicle for information dissemination, plays a significant role in the risk of stock price crashes. Sentiments given media expresses matter. Pure positive or negative news, which may mask the firms' actual situation, can exacerbate information asymmetry between firms and outside investors. In addition, the problems of irrational sentiment, herding effects, and "chasing the upside and killing the downside" phenomenon are aggravated by biased news. Both channels can increase the risk of a stock price crash. Considering the preceding ideas, we formulate the following competing hypotheses:

Hypothesis *H1a*: Higher average media sentiment can exacerbate the future stock price crash risk.

Hypothesis *H1b*: Higher average media sentiment can alleviate the future stock price crash risk.

We also examine the heterogeneous effect of sentiment in news on future stock price crash risk. When the media exaggerate the positive parts of news, they send positive signals on the firms to outside investors. With information asymmetry, investors will overestimate the firms' value, which can lead to abnormal stock price increases. As short selling is not allowed in China, rational and pessimistic investors are unable to engage in the market and stock prices will continue to increase until investors realize that there is an overvaluation component in the news. According to Solomon [44], the media's whitewashing behavior of overusing positive terms to disclose the information of listed firms can lead to a sharp decline in stock prices. Therefore, intense positive news can increase stock price crash risk.

Alternatively, related research reveals the tendency of firm management to conceal negative information from the public [22, 28], which inevitably leads to an overestimate of firm value, resulting in higher stock prices. Simultaneously, retail investors are more sensitive to negative news [42]. When management has no choice but release negative information to the market, retail investors will sell off stock holdings, which increases the risk of a stock price crash [52]. Therefore, the coverage of negative news can increase the risk of stock price crash. Considering the preceding ideas, we hypothesize as follows:

Hypothesis *H2a*: News with positive sentiment can exacerbate the future stock price crash risk.

Hypothesis *H2b*: News with negative sentiment can exacerbate the future stock price crash risk.

Subsequently, we then proceed with identifying the underlying mechanisms. We hypothesize that the impact may come from the two channels of "information intermediation" and "investor sentiment."

According to Li and Stewart [32], when there is information asymmetry, the agency problem, such as management's rent-seeking and concealment of negative news, can affect the share price crash risk significantly. Insiders have information that is not yet publicly available, which can be used to assess the value of the company and predict future firm performance [39]. Insider sell-off behavior is positively associated with the risk of a stock price crash [19].

The insider’s choice to sell stocks sends negative signals to outside investors, thereby raising the probability of a future crash risk. Based on the explanation above, we hypothesize,

Hypothesis *H3a*: The impact of media sentiment on the future stock price crash risk is enhanced when insiders have more net sales of stock.

Information asymmetry can prevent investors from knowing the firm’s actual operation, and investors may be deceived by false public information. Especially, firms whose information transparency is low and management is more likely to hide bad news are more likely to experience a sharp stock price fall in future [22]. When financial opacity is high, investors cannot fully grasp the true state of a firm through public information. They will rely more on the media coverage to make investment decisions, which will amplify the role of media sentiment. Therefore, we hypothesized as follows.

Hypothesis *H3b*: The impact of media sentiment on the future stock price crash risk is enhanced when the firm has higher financial opacity.

Analysts serve as both information intermediaries and management monitors [43]. They acquire and process data about firms using public information, field research, and other sources and reduce information asymmetry between firms and investors [3, 30]. According to He et al. [19], analyst coverage reduces stock price crash risk via analysts’ role as information intermediaries and monitors. Nonetheless, when analysts cannot fully perform information intermediary role, it is more difficult for investors to learn the real situation of the firm, and they cannot accurately identify the noise in the media coverage, thus the impact of media sentiment on future stock price risk is more evident through driving investors' decision-making process. We thus hypothesize as follows:

Hypothesis *H3c*: The impact of media sentiment on the future stock price crash risk is enhanced when the firm has lower analyst coverage.

Next, we test the existence of the “investor sentiment” channel.

The sentiment of media affects investors differently. Institutional investors have more information and specialized ability; therefore, it is easier for them to judge the validity of the information contained in the news. Sentiments in news thus have limited influence on institutional investors. Conversely, retail investors do not have the information and skills possessed by institutional investors; thus, sentiments in news have a greater impact on retail investors. Therefore, we formulate the following hypothesis:

Hypothesis *H4a*: The impact of media sentiment on the future stock price crash risk is enhanced when the stock has a higher proportion of retail investors.

Investors may place too great faith in catchy, anecdotal, low-relevance information and overreact to it as a result of limited attention and overconfidence [37]. They may also exhibit confirmation bias, which is the tendency to seek and believe information that supports one’s beliefs while ignoring later signals that are inconsistent with prior beliefs after developing a favorable impression of the firm [40]. Thus, positive media coverage of firms attracts investors’ attention, thereby causing them to overreact or form over-expectations about the firm’s prospects, resulting in a stock price that is briefly or chronically above the underlying value. Therefore, we proposed the following hypothesis:

Hypothesis *H4b*: The impact of media sentiment on the future stock price crash risk is enhanced when investors are more optimism.

Heterogeneous beliefs among investors increase when analysts' opinions differ from the sentiment of media coverage. In the absence of a short selling mechanism, more optimistic investors are expected in the market in the short term. However, over time, pessimistic beliefs will eventually emerge, which will increase the likelihood of a future stock price crash. Therefore, we hypothesize as follows:

Hypothesis *H4c*: The future stock price crash risk increases when analysts' points disagree with the sentiment of media.

3.Sentiment extraction model design

Our study obtains financial news data for the period January 2017 to December 2020 from the China Stock Market and Accounting Research database. We preprocess the data by deleting special symbols and irrelevant information. We select 3,305 news items from the 1,132,856 news texts to label the news regarding the financial entities as positive or negative sentiments and added them to the financial sentiment dictionary manually. Using the BERT pretraining model and financial sentiment dictionary-based attention mechanism, we classify the sentiment of 1,132,856 news items. We then derive the average sentiment index of each stock in each year using the weighted average of all related news sentiment.

3.1Data preprocessing

We construct a microsentiment corpus in the financial field. To avoid interference with the hard data contained in the news, we eliminate the company announcements and then preprocess the text by removing special symbols and using regular matching to remove irrelevant information.

Data clean up: We label the financial entities: we identify the company, person, and brand names in the text. Entity names are marked based on the principle of long matching. We also identifiy the company and brand names with the help of "Tianyancha." For example, in the following text, "Runtu Shares: Ruan Jiachun (Chairman) plans to reduce no more than the total share capital of 1.28," "Runtu Shares" and "LeTV" are marked as entities.

Label news sentiment: The sentiment polarity of financial entities is grouped into three categories—neutral, negative, and positive. Each category is defined as follows and Table 1 shows the distribution,

Positive sentiment: The text is marked as positive if the fact favors the operation of the company, and there are some artificial positive comments. For example, "Southeast network frame won the bid of 357-million-yuan project."

Negative sentiments: If the information in the text is bad for the company's operation because it includes some facts that are bad for the operation of the company and artificial negative comments. For example, "Tianmaotui will be delisted from the Shenzhen Stock Exchange on July 20."

Neutral sentiment: Unlike positive and negative sentiments, the labeling of neutral sentiments is relatively complex. The text information is related to the operation of the company but cannot be judged as favorable or unfavorable, or it has both favorable and unfavorable facts. For example, “E-commerce is the direction of future development, all enterprises are making efforts. So does Huawei, but at present, the effectiveness needs to be tested.”

To construct a microsentiment analysis dataset in the financial field, we select 4,516 samples from the 1,132,856 news texts obtained for annotation. After the independent annotation, all annotators discussed the additional annotations noted in the case of objectionable or uncertain results until consensus was reached. The annotation data were artificially modified, and the annotation was completed.

Table 1: Financial entities

	Positive	Neutral	Negative
Nums.	4,179	3,202	1,627
Pct	46.39%	35.55%	18.06%

Note: The sentiment polarity of financial entities is grouped into three categories—neutral, negative, and positive. Table 1 shows the distribution.

Finally, 3,644 financial entities are sorted out. Each financial entity corresponds to one or more sentences. Each article has a total of 10,112 sentiment sentences. Based on prior financial knowledge, we construct a sentiment dictionary in the financial domain, which contains 2,842 positive words, 1,230 neutral words, and 2,043 negative words (Table 2).

Table 2: Financial dictionary

	Positive	Neutral	Negative
Nums.	2,842	1,230	2,043
Pct	46.48%	20.11%	33.41%

Note: We construct a sentiment dictionary in the financial domain, which contains 2,842 positive words, 1,230 neutral words, and 2,043 negative words.

3.2 BERT pretraining model

In 2018, Google proposed the natural language pretraining model, BERT, in the article “BERT: Pretraining of Deep Bidirectional Transformers for Language Understanding”. Training of the BERT model mainly includes the following two steps.

BERT pretraining: The pretraining of BERT helps it learn the characteristics of a character, a word, statement levels, and understatement relationships among massive text data through simultaneous two pretraining tasks—masked language model and next sentence prediction. During pretraining, the same corpus is inputted into the model multiple times, but each input is preprocessed in different forms, allowing the same corpus to be fully utilized. For users, the pretrained models and parameters can be downloaded from the Internet and can be directly fine-tuned without having to do pretraining themselves, which reflects the convenience of BERT.

Fine-tuning: On the basis of the pretrained model, an output layer is customized and added to specific downstream tasks, such as text sentiment classification and sequence annotation. Then, the data from downstream tasks are used to fine-tune the model to generate models with higher prediction accuracy for various NLP tasks.

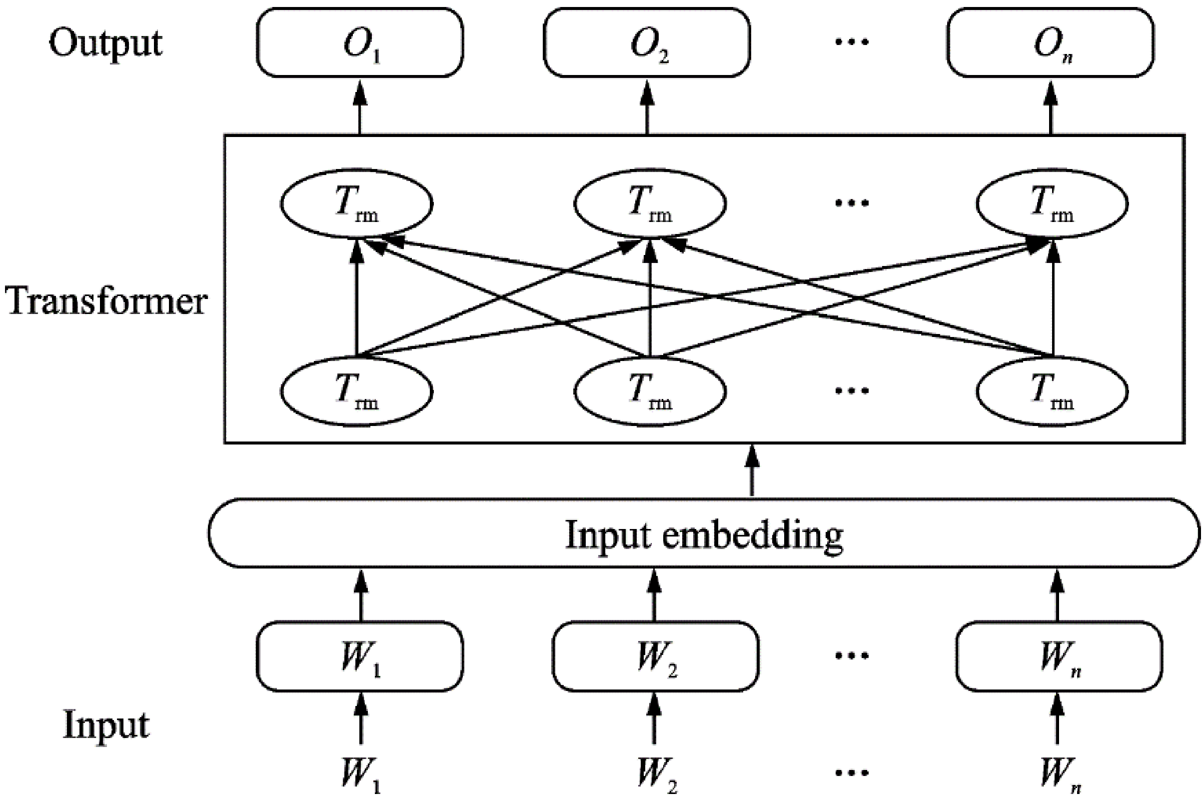

Figure 1: BERT model structure

BERT uses a more powerful bidirectional transformer encoder (Fig. 1) along with masked language model and next sentence prediction (NSP) as unsupervised goal, to enable the vector representation of each word and word output by the model to describe the overall information of the input text as comprehensively and accurately as possible. Thus, BERT provides better initial values of model parameters for subsequent fine-tuning tasks. Its input embedding is constructed by summing the token, segment, and position embedding of the corresponding word. It also contains more parameters, which give it a stronger word vector embedding ability.

3.3 Construction of BERT-DCA model

We construct a BERT-DCA model (Fig. 2) that combines the financial sentiment dictionary and attention for sentiment analysis. Two information processing channels—left semantic information attention channel (SAC) and right sentiment information attention channel (EAC)—are adopted in the structure. The SAC extracted semantic information, whereas the EAC allowed the model to pay attention to the particularity of different types of words to supplement the weights and obtain more information as a supplement to word-level information.

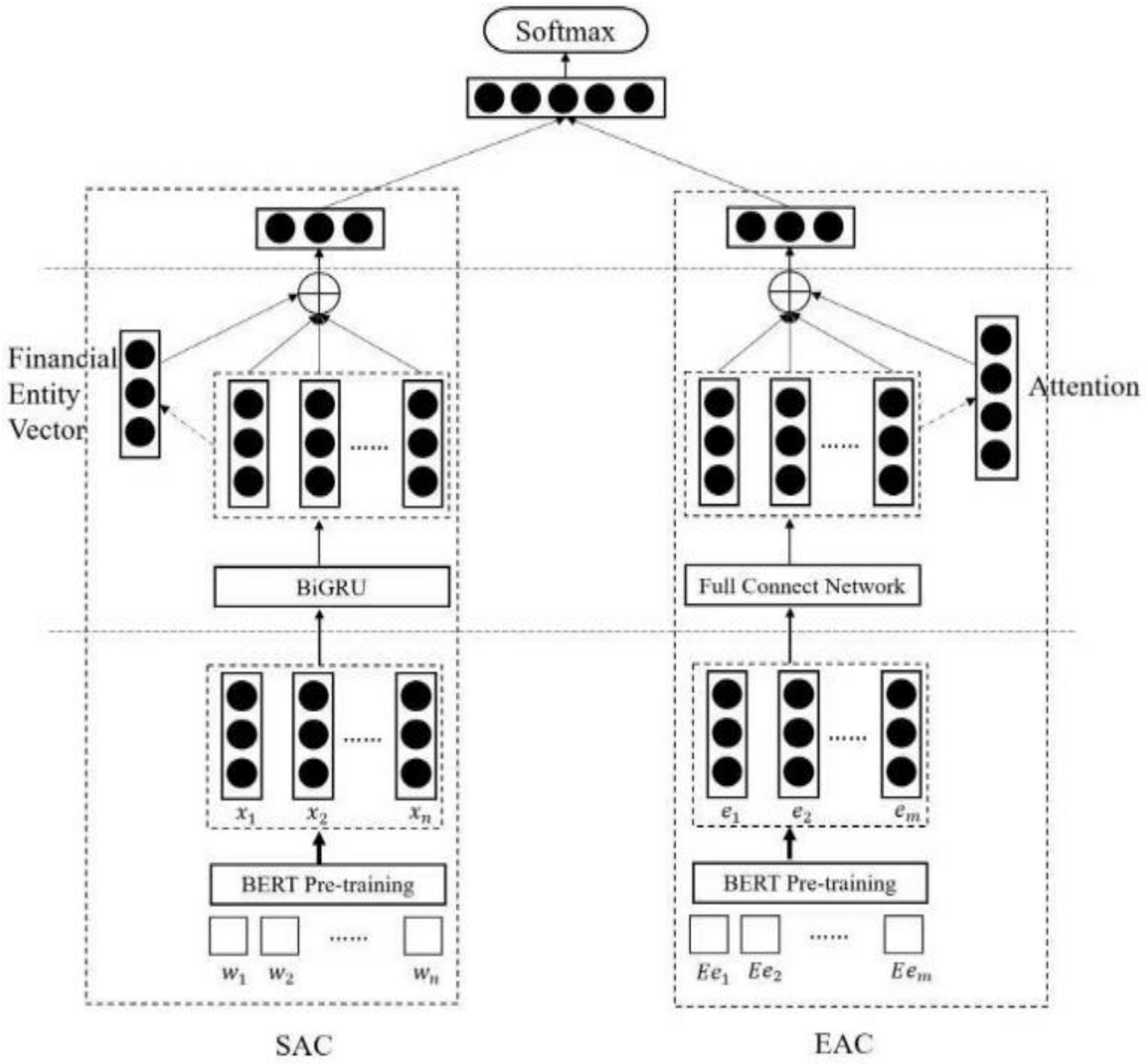

Figure 2 A model framework that combines sentiment dictionaries and attention

Input layer: For the text sentence sequence, after word partitioning, the word sequence $\{W1, W2, \cdots, Wn\}$ is used as the input for SAC. Based on the domain sentiment dictionary and financial entities, words are classified into the following five categories: Pos, Neg, Neu, Entity, and other, which denote positive, negative, neutral, financial entities, and others, respectively. They are from the sentimental dictionary discussed above. Then, we derive the sentimental information word collection $\{E1, E2, \cdots, Em\}$ as the input of EAC. We then use the pretraining model, BERT, to provide the word vector, which can achieve the dynamic adjustment of the word vector with the context, and train the real sentiment semantic embedding model to obtain the semantic information word vector matrix R_x and the sentiment information word vector matrix R_e.

$$R_x = x_1 \oplus x_2 \oplus \ldots . \oplus x_n \tag{1}$$

$$R_e = e_1 \oplus e_2 \oplus \ldots . \oplus e_m \tag{2}$$

where $\oplus$ is the row vector connection operator and the dimensions of R_x and R_e denote the number of words in the news and of annotated financial sentiment entities, respectively.

Feature extraction: For semantic information texts, we used the BiGRU neural network (model structure shown in Figure 3) to handle both forward and reverse text sequences. We extracted the deep text information and then used the financial dictionary to guide attention mechanisms to assign corresponding weights to the extracted feature information. For sentiment information sets, affective information words were encoded using a fully connected network combined with attention mechanisms to obtain affective signals.

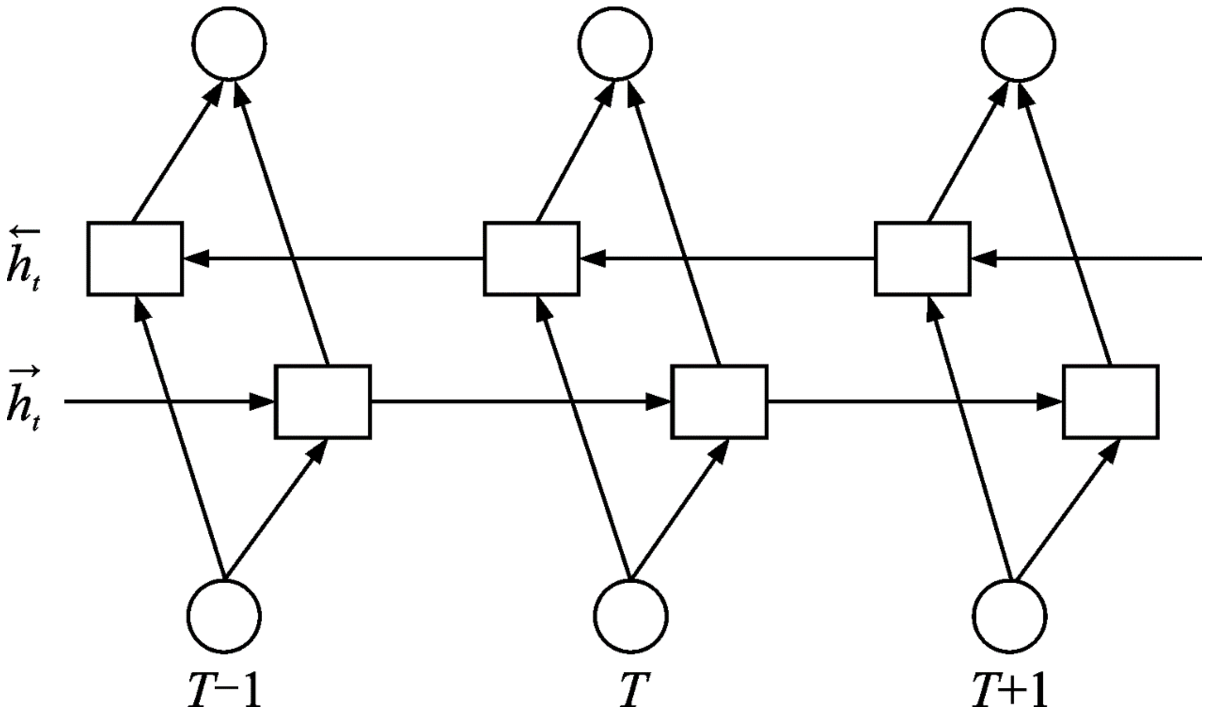

Figure 3 Structure of BiGRU

The output of the BiGRU information extraction model at time t is composed of the output of the forward and reverse GRU, calculated as follows:

$$x_t = W_e w_t, \;\; t \in [1, T] \tag{3}$$

$$\overrightarrow{h}_t = \overrightarrow{GRU}(x_t), \;\; t \in [1, T] \tag{4}$$

$$\overleftarrow{h}_t = \overleftarrow{GRU}(x_t), \;\; t \in [1, T] \tag{5}$$

$$s_t = [\overrightarrow{h}_t, \overleftarrow{h}_t], \;\; t \in [1, T] \tag{6}$$

By combining $\overrightarrow{h}_t$ and $\overleftarrow{h}_t$ to obtain the semantic representation s_t, the forward and reverse semantic information elements are considered in the same position.

Next, we use the financial dictionary to guide attention mechanisms. To improve the accuracy of the sentiment analysis of financial text, we model the relation between sentiment and each word, assigning different weights to the semantic characteristics of the clause using the attention mechanism. In this way, more important words get more attention. Based on the financial entities and BiGRU layer output $H^s = \{h_1^s, h_2^s, \dots, h_t^s\}$, we obtained vectorized representations as word-level attention. The weight is calculated as follows:

$$\alpha_{st} = \frac{\exp(\gamma(h_c^s, e^E))}{\sum_c \exp(\gamma(h_c^s, e^E))} \tag{7}$$

$$\gamma(h_c^s, e^E) = \tanh\left(h_c^s \cdot \omega_m^T \cdot e^{E^T} + b_a\right) \tag{8}$$

The output after BiGRU processing is expressed as $[h_1^s, h_2^s, \dots, h_c^s]$, where ω_m^T is the weight matrix, b_a is the offset, e^E is the word vector of the financial entity, and α_{st} is the attention weight of the word w_{st} relative to the financial entity e^E. The text features with attention-weighted sentences are represented as

$$o_s = \sum_t a_{st} h_{st} \tag{9}$$

where o_s is a semantic representation weighted by attention.

Feature fusion layer: The main task of the feature fusion layer is to combine the feature vector O^s generated in SAC and feature vector O^e generated in EAC, to construct the overall sentimental feature vector. To simplify the calculation of the model, we perform feature fusion by row connection, constructing a matrix $O^* = (r_s + r_e) * c$ to generate the feature vector, where r_s and r_e are the number of rows of O^s and O^e, respectively, and c gives the column numbers for O^s and O^e.

Output layer: We input the sentiment feature vector O^* generated by the feature fusion layer into the Softmax classifier to obtain the final sentiment classification result predicted by the model:

$$p = softmax(w_o O^* + b_o) \tag{10}$$

where w_o is the weight coefficient matrix, b_o is the bias matrix, and p is the predicted sentiment label.

Model training: To use the constructed financial sentiment dictionary that could correspond to the input sentence, we need to construct a sentiment word vector of the same length as the term after the segmentation: Vec_{Att}, initialized as 0. After traversing the words in the input financial text, we set the corresponding position to 1 in the sentiment word vector if they appear in the financial sentiment dictionary. For example, assuming that the input financial text is “Langma cloud business development uncertainty,” we first initialize the sentiment word vector [0,0,0,0,0]. After the input sentence, the word “uncertainty” appears in the financial sentiment dictionary, and it is a negative word. Then, we set the word “uncertainty” in the corresponding position of the sentiment word vector to 1, after which the sentiment word vector of the sentence becomes [0,0,0,0,1].

To employ the financial dictionary shown above as a guiding attention mechanism, we modify the loss function and add $\lambda(\alpha - Vec_{Att})^2$ after the cross-entropy loss. Here, λ is the hyperparameter that determines the importance of sentiment dictionary loss, α is the score of the attention mechanism, and Vec_{Att} is the sentiment dictionary vector. Thus, the attention mechanism score α can fit the financial sentiment word vector for the model to pay more attention to the input financial text—the financial sentiment words. The loss function is

$$L = -\sum_{i \in D} y_i log\, p_i + \lambda(\alpha_{norm} - Vec_{Att}) \tag{11}$$

where D is the collection of samples, y_i is the true label, and p_i is the prediction result of the model. λ is the hyperparameter that determines the importance of affective dictionary loss, and α_{norm} is the average attention score.

We thus use the predicted labels of 1,132,856 news items as the sentiment score in the empirical analysis.

4.Empirical model

4.1Sample selection and data sources

Our sample covers all A-share listed firms from 2011 to 2020. Owing to the lag phase of the study, the data time span is nine years (from 2012 to 2020).

We obtain the financial stock trading data from the WIND database. Among them, stock yield is given as weekly data, and the rest, as annual data. Following prior studies [30, 47], we process the original sample as follows: (1) We exclude financial and insurance listed firms; (2) We exclude listed firms with ST or * ST[1]; (3) We exclude listed firms with missing or abnormal data; (4) We exclude listed firms with less than 15 weekly yield data.

We obtain data regarding internet media news from the GuoTai'an (CASMAR) database. We perform positive and negative analyses of each report using sentiment analysis technology and assign sentiment scores. We then calculate the sum of the number of relevant news reports during the research period, average level of sentiment scores, and sentiment scores weighted by the number of news reports.

4.2Econometric model and variables

To study the relation between diversification and future stock price crash risk, we construct a multiple regression model as follows:

$$CrashRisk_{i,t} = \beta_0 Cons + \beta_1 NewsSentiment_{i,t-1} + \beta_2 Size_{i,t-1} + \beta_3 Level_{i,t-1} + \beta_4 ROA_{i,t-1} + \beta_5 Ret_{i,t-1} + \beta_6 Sigma_{i,t-1} + \sum Year + \sum Firm + \varepsilon_{i,t}. \quad (12)$$

In the model, the explained variable $CrashRisk$ represents the risk of a crash in individual stocks. $NewsSentiment$ is the core explanatory variable, indicating the calculated sentimental score for the individual stock news report. $Size$ denotes the size of the enterprise, $Level$ represents the financial leverage of the company, ROA is the return on equity of the company, Ret is the average of the enterprise-specific weekly rate, and $Sigma$ is the standard deviation of the enterprise-specific weekly rate. $Year$ and $Firm$ are time and firm fixed effects, respectively. Here, we focus on the coefficient β_1. If β_1 is significantly positive, then there is a positive relation between the news reporting sentiment and risk of a stock price crash. Conversely, if β_1 is significantly negative, then there is a negative relation between the news reporting sentiment and risk of a stock price crash. The variables are introduced in Table 3 and elaborated as follows:

[1] ST: the company has suffered losses for two consecutive years and is specially treated, ST*: the company has suffered losses for three consecutive years and warned with delisting.

Explained variables. Based on the methods of Li and Stewart [32] and Xu et al. [49], our study employ three approaches to measure the risk of a stock price crash. The specific algorithm is as follows:

The unexplained weekly yield of individual stocks in the market is calculated using the following model:

$$R_{i,t} = R_{m,t-2} + R_{m,t-1} + R_{m,t} + R_{m,t+1} + R_{m,t+2} + \varepsilon_{i,t} \tag{13}$$

where $R_{i,t}$ represents the weekly yield of stock i in week t. $R_{m,t}$ is the weighted average of the weekly yield of week t. $\varepsilon_{i,t}$ is the residual in Equation (2), which represents the weekly return of stocks not explained in the market. Because $\epsilon_{i,t}$ is highly biased, we use $W_{i,t} = \ln(1+\varepsilon_{i,t})$ to represent stock-specific weekly yields. Based on $W_{i,t}$, we measure the risk of a stock price crash using three indicators—($CRASH$), a negative return bias coefficient ($NSCKEW$), and the earnings fluctuation ratio ($DUVOL$).

$CRASH$ is calculated as follows:

$$CRASH_{i,t} = 1[\exists t, W_{i,t} \leq Average\left(W_{i,t}\right) - 3.09\sigma_{i,t}] \tag{14}$$

$CRASH_{i,t}$ equals 1 if a firm experiences one or more firm-specific weekly returns $W_{i,t}$ falling 3.09 standard deviations below the mean firm-specific weekly return, and 0 otherwise

$NSCKEW$ is calculated as follows:

$$NCSKEW_{i,t} = \frac{-[n(n-1)^{3/2}\sum W_{i,t}^{3}]}{[(n-1)(n-2)\left(\sum W_{i,t}^{2}\right)^{3/2}]} \tag{15}$$

where n represents the number of stock i in year t. The coefficient of the negative return bias is a positive measure of the risk of a stock price crash. Thus, the greater the coefficient is, the higher the possibility of a stock price crash.

$DUVOL$ is calculated as follows:

$$DUVOL_{i,t} = log\left\{\frac{[(n_u-1)\sum_{DOWN} W_{i,t}^{2}]}{[(n_d-1)\sum_{UP} W_{i,t}^{2}]}\right\} \tag{16}$$

The core explanatory variable in our study is a quantitatively weighted news reporting sentiment propensity, calculated as follows:

Explanatory variables. The core explanatory variable in this study is a quantitatively weighted news reporting sentimental propensity, which is calculated as follows:

$$newSenti_{i,T} = \frac{NewsCount_{i,t} * SentimentScore_{i,t}}{\sum NewsCount_{i,t}} \tag{17}$$

where $newSenti_{i,T}$ represents the media report sentiment tone of stock i in year T. $NewsCount_{i,t}$ represents the number of news items regarding stock i in year T.

$SentimentScore_{i,t}$ represents the average media reporting sentiment scores of trading day t of stock i in year T, each calculated by our BERT-DCA model. Regarding the number of news, the higher it is, the more likely the investors will read the news. Thus, the probability that the reported sentiments are transmitted to investors is also higher. To examine how different types of internet news sentiment work, we construct both positive and negative news coverage sentiment indicators:

$$newPos_{i,T} = \frac{PosNewsCount_{i,t} * PosSentimentScore_{i,t}}{\sum PosNewsCount_{i,t}} \tag{18}$$

$$newNeg_{i,T} = \frac{NegNewsCount_{i,t} * NegSentimentScore_{i,t}}{\sum NegNewsCount_{i,t}} \tag{19}$$

To study the impact of market differences on the risk of a stock price crash, we use analysts' rating data to calculate their sentiments. First, we grade analysts on five points: +2, +1, 0, −1, and −2, indicating buy, overweight, neutral, reduction, and sell, respectively. We then calculate the total score of the year, divide it by the rating number, and finally standardize it using ~N(0,1), given as Ana_senti.

Control variables. Following Li and Stewart [32], the control variables are defined as follows.

Enterprise size ($Size$) is expressed as the natural logarithm of the enterprise market value
Operating leverage ($Level$) is the enterprise asset–liability ratio
Compensation rate of corporate total assets (ROA) is an indicator used to measure corporate profitability
Previous value of negative return bias coefficient/fluctuation ratio is used to control the impact of the lag phase of the risk of a stock price crash
Stock-specific weekly earnings annual average (Ret) reflects the average level of stock yield
Weekly earnings volatility ($Sigma$) reflects the volatility levels of stock-specific weekly earnings

Table 3: Variables and definitions

	Indicator	Definition
Dependent variables	CRASH	Used to measure the risk of a crash: 1, 0 indicator
	NCSKEW	Used to measure the risk of a crash: negative return bias coefficient
	DUVOL	Used to measure the risk of a crash: the earnings fluctuation ratio
Independent variables	newSenti newPos newNeg	News coverage' sentiment
	anaSenti	Analysts' sentiment
Control variables	Size	The natural logarithm of the firm's market value
	Level	Firm's leverage, equal to the ratio of the firm's total liabilities to total assets
	ROA	The ratio of net profit to total assets

	Ret	The average value of the firm's annual weekly rate of return
	Sigma	The standard deviation of the firm's annual weekly rate of return
	Year	Yearly dummy variable
	Firm	Firm dummy variable

5.Empirical analysis

5.1 News sentiment and the risk of a stock price crash

Table 4 shows the analysis results for the relation between the media sentiment and future stock price crash risk. Columns 1, 3, and 5 in Table 4 are the effects of the sentiment weighted by the number of news items ($newSenti$) on the risk of a future stock price crash where control variables are not included. We found a significant positive effect of $CRASH$(0.087), $NCSKEW$ (0.395), and $DUVOL$ (0.203). Columns 2, 4, and 6 show the effect of sentiment ($newSenti$) on the future stock price crash risk after adding all control variables. Similarly, we find a significant positive effect on $CRASH$(0.064), $NCSKEW$ (0.190), and $DUVOL$ (0.103). Although the coefficient decreases after including control variables, they are still economically and statistically significant. Thus, the more positive the average media sentiment is, the higher the future stock price crash risk, or the more negative the current average news sentiment is, the lower the risk of a future stock price crash. These findings support $H1a$ but not $H1b$.

Table 4 Baseline regression

	(1) crash	(2) crash	(3) ncskew	(4) ncskew	(5) duvol	(6) duvol
L.newSenti	0.087***	0.064**	0.395***	0.190**	0.203***	0.103*
	(0.031)	(0.032)	(0.092)	(0.094)	(0.062)	(0.060)
L.crash		-0.147***				
		(0.009)				
L.ncskew				-0.096***		
				(0.009)		
L.duvol						-0.106***
						(0.008)
L.ret		0.957**		8.770***		6.232***
		(0.410)		(0.842)		(0.584)
L.sigma		0.594***		1.039**		1.192***
		(0.195)		(0.441)		(0.299)
L.roa		0.360		1.339***		0.860***
		(0.224)		(0.459)		(0.307)
L.level		-0.012		-0.130**		-0.113***
		(0.033)		(0.059)		(0.039)
L.size		0.000		0.000*		0.000
		(0.000)		(0.000)		(0.000)
Fixed effects:						
Year dummy	Yes	Yes	Yes	Yes	Yes	Yes
Firm dummy	Yes	Yes	Yes	Yes	Yes	Yes

N	18,775	17,267	18,775	17,267	18,775	17,267
R^2	0.000	0.035	0.001	0.060	0.001	0.065

This table reports baseline regression estimates of stock crash risk on the quantitatively weighted news coverage sentiment. The sample of CRASH is used for columns (1)-(2), the sample of NCSKEW is used for columns (3)-(4) and that of DUVOL is used columns (5)-(6). Columns (1), (3) and (5) only include the news coverage sentiment; columns (2), (4) and (6) control for lagged crash risk and firm characteristic; year and firm dummy variable are included; all control variables are included as lagged one year. Robust standard errors are reported in parentheses. The labels ***, **, and * indicate 1%, 5 %, and 10 % levels of significance, respectively.

Regarding the control variables, we observe a negative effect of firm financial leverage on the risk of a future stock price crash, implying that the latter risk is higher in firms with lower financial leverage—the smaller the size of the firm, the higher the risk. In addition, we find that a firm's stock return (ROA) is significantly and negatively related to the risk of a future stock price crash, implying that the better the performance of a firm, the less likely it is to have a stock price crash in future. The effects of the other control variables on future stock price crash risk are not robust.

According to the results of the baseline regression, media sentiment is positively related to the future stock price crash risk. We replace average media sentiment in the baseline regression with media coverage positive and negative sentiment indicators to examine hypothesis *H2*. The results are shown in Table 5.

507 Table 5 Positive sentiment and negative sentiments of news coverage

			Panel A: One lag period			
	(1)	(2)	(3)	(4)	(5)	(6)
	crash	ncskew	duvol	crash	ncskew	duvol
L.newPos	0.135***	0.516***	0.323***			
	(0.045)	(0.187)	(0.125)			
L.newNeg				0.033	0.087	0.015
				(0.073)	(0.162)	(0.110)
L.crash	-0.152***			-0.151***		
	(0.009)			(0.009)		
L.ncskew		-0.102***			-0.107***	
		(0.009)			(0.009)	
L.duvol			-0.111***			-0.116***
			(0.009)			(0.009)
L.ret	1.060**	9.014***	6.415***	1.016**	9.305***	6.560***
	(0.426)	(0.869)	(0.606)	(0.431)	(0.888)	(0.614)
L.sigma	0.618***	1.204***	1.261***	0.545***	1.059**	1.183***
	(0.204)	(0.455)	(0.310)	(0.203)	(0.459)	(0.312)
L.roa	0.371	1.388***	0.877***	0.425*	1.427***	0.812**
	(0.236)	(0.489)	(0.321)	(0.235)	(0.485)	(0.326)
L.level	-0.016	-0.111*	-0.108***	-0.004	-0.147**	-0.142***
	(0.035)	(0.059)	(0.040)	(0.037)	(0.063)	(0.042)
L.size	0.000	0.000*	0.000	0.000	0.000**	0.000*
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Fix effects:						
Year dummy	Yes	Yes	Yes	Yes	Yes	Yes
Firm dummy	Yes	Yes	Yes	Yes	Yes	Yes
N	16,267	16,267	16,267	15,966	15,966	15,966
R^2	0.036	0.060	0.067	0.036	0.062	0.068
			Panel B: Same period			
	(1)	(2)	(3)	(4)	(5)	(6)
	crash	ncskew	duvol	crash	ncskew	duvol
newPos	-0.068	-0.023	-0.010			
	(0.088)	(0.187)	(0.129)			
newNeg				-0.185**	-0.427***	-0.303***
				(0.077)	(0.159)	(0.111)
L.crash	-0.161***			-0.161***		
	(0.009)			(0.009)		
L.ncskew		-0.109***			-0.110***	
		(0.009)			(0.009)	
L.duvol			-0.113***			-0.116***
			(0.009)			(0.009)
L.ret	0.662	5.425***	3.671***	0.570	5.596***	3.864***
	(0.417)	(0.888)	(0.630)	(0.425)	(0.908)	(0.640)
L.sigma	0.566***	0.138	0.219	0.556***	0.105	0.311
	(0.202)	(0.466)	(0.322)	(0.201)	(0.475)	(0.326)
L.roa	0.647***	1.765***	1.094***	0.489**	1.272***	0.822**
	(0.221)	(0.461)	(0.316)	(0.214)	(0.467)	(0.321)
L.level	-0.048	-0.112*	-0.108**	-0.049	-0.116*	-0.115***
	(0.036)	(0.066)	(0.044)	(0.036)	(0.065)	(0.044)

L.size	-0.000	0.000	-0.000	-0.000	0.000	0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Fix effects:						
Year dummy	Yes	Yes	Yes	Yes	Yes	Yes
Firm dummy	Yes	Yes	Yes	Yes	Yes	Yes
N	16,267	16,267	16,267	15,966	15,966	15,966
R^2	0.041	0.066	0.072	0.037	0.064	0.070

This table reports regression estimates of stock crash risk on the news coverage positive and negative sentiment indicators. Panel A is one year lag; Panel B is the same year; year and firm dummy variable are included; all control variables are included as lagged one year. Robust standard errors are reported in parentheses. The labels ***, **, and * indicate 1%, 5 %, and 10 % levels of significance, respectively.

Columns 1, 2, and 3 in Panel A indicate that the effect of positive sentiment is significant at the 1% level, whereas Columns 4, 5, and 6 indicate that the effect of negative sentiment is insignificant. The regression results suggest that positive media sentiment plays a dominant role in China; the more positive the media sentiment, the higher the future stock price crash risk. Under information asymmetry, uninformed investors receive more positive information, and irrational investors develop an overvaluation of stock prices [6, 21]. Negative sentiment appears to curb future crash risk, but the effect is insignificant.

Indeed, investors react more strongly to negative news [42]. Negative news causes retail investors to sell and increases the risk of a future stock price crash [52]. However, our findings indicate that negative news in the previous one year does not increase the risk of a future stock price crash. A reason may be that investors are overly sensitive to negative news, particularly in markets with a high proportion of irrational investors, such as China. When negative news breaks, investors panic and quickly sell their stocks, resulting in a significant drop in the stock price below its fundamental value, which means the occurrence of the price crash in the current term rather than the future. Thus, negative news coverage can be hypothesized to increase the current stock crash risk, whereas positive media coverage decreases such a risk. The results are represented in Table 5, Panel B. It shows that the effects of negative news in the current period is significant, confirming our analysis that the more negative the news, the greater the risk of a crash in the current period. Meanwhile, the effects of positive media coverage are insignificant, indicating that positive news suppresses the risk of a crash in the current period. However, because investors are less sensitive to positive than negative news, their impact is not significant.

To investigate the robustness of the impact of positive and negative news, we considered quarterly level crash risk regressions that included current, prior one period to prior three period indicators for positive and negative news, respectively. The results are shown in Table 6. Overall, the quarterly regression results are consistent with the annual regression results. Positive news reduces the risk of a stock price crash in the current period (−0.035); however, the effect is not statistically significant. Columns 2–4 show that positive news significantly increases the risk of a stock price crash in the future period, consistent with the annual regression. Negative media coverage increases the current period crash risk (−0.164) and had a significant effect. Columns 6–8 reveal that negative media coverage reduces future stock price crash risk, but the suppression effect is insignificant except for the previous period. These results confirm *H2a* but not *H2b*.

Table 6 Quarterly crash risk for different future periods

	(1) crash	(2) crash	(3) crash	(4) crash	(5) crash	(6) crash	(7) crash	(8) crash
newPos	-0.035							
	(0.039)							
L.newPos		0.116***						
		(0.039)						
L2.newPos			0.082**					
			(0.039)					
L3.newPos				0.047				
				(0.039)				
newNeg					-0.164***			
					(0.036)			
L.newNeg						0.046		
						(0.036)		
L2.newNeg							0.042	
							(0.036)	
L3.newNeg								-0.048
								(0.036)
L.crash	-0.052***	-0.053***	-0.055***	-0.052***	-0.052***	-0.053***	-0.052***	-0.053***
	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)
L.ret	1.523***	1.824***	2.339***	2.999***	1.709***	1.917***	2.418***	2.833***
	(0.479)	(0.484)	(0.540)	(0.565)	(0.476)	(0.478)	(0.533)	(0.540)
L.sigma	1.734***	1.633***	1.807***	2.052***	1.698***	1.617***	2.066***	1.916***
	(0.253)	(0.260)	(0.282)	(0.283)	(0.260)	(0.256)	(0.276)	(0.280)
L.roa	-0.070	-0.132***	-0.194***	-0.141***	-0.095*	-0.143***	-0.134***	-0.155***
	(0.052)	(0.051)	(0.053)	(0.052)	(0.052)	(0.051)	(0.052)	(0.053)
L.level	-0.038**	-0.037**	-0.038*	-0.034*	-0.029	-0.030	-0.038*	-0.009
	(0.019)	(0.018)	(0.020)	(0.019)	(0.020)	(0.020)	(0.020)	(0.020)
L.size	0.000	-0.000	0.000	0.000	0.000	0.000	0.000	0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Fix effects								
Quarter dummy	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Firm dummy	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	51,110	51,110	51,110	51,110	51,919	51,919	51,919	51,919
R^2	0.040	0.039	0.039	0.038	0.037	0.038	0.039	0.038

This table reports regression estimates of quarterly stock crash risk on the news coverage positive and negative sentiment indicators. Columns (1) to (4) are positive indicators; columns (5) to (8) are negative indicators; quarter and firm dummy variable are included; all control variables are included as lagged one year. Robust standard errors are reported in parentheses. The labels ***, **, and * indicate 1%, 5 %, and 10 % levels of significance, respectively.

5.2 Endogeneity

The relation between media coverage and stock returns is endogenous. The reverse causality, as media coverage is more likely to focus on stocks with higher returns, also there may be control variables that we are unaware of, resulting in the omitted variables problem.

Following Xu, Li, Yuan, and Chan [48] and Ertugrul, Lei, Qiu, and Wan [12], we select industry-level news sentiment means ($newSentiInd$) but exclude the company and province levels ($newSentiPro$) as the instrument variables for firms' media sentiment. We presume that other publicly traded firms in the same industry or province would face similar industry characteristics and external environments; thus, their media coverage may have a certain correlation. Furthermore, there is no evidence that media coverage of other publicly traded firms in the same industry or province will influence a firm's stock trading behavior, which satisfies the exclusion restriction to some extent.

Table 7 Two stages OLS regression

	First stage	Second stage		
	(1)	(2)	(3)	(4)

	newsenti	crash	ncskew	duvol
L.newSenti		0.074***	0.510**	0.313**
		(0.028)	(0.253)	(0.157)
L.newSentiInd	0.873***			
	(0.022)			
L.newSentiPro	0.747***			
	(0.041)			
Controls	Yes	Yes	Yes	Yes
Cragg-Donald Wald F	185.966***			
Sargan chi(p)		0.257(0.612)	0.571(0.450)	1.466(0.226)
N	16,845	16,845	16,845	16,845
R^2	0.212	0.035	0.059	0.064

This table reports regression estimates of stock crash risk on the news coverage sentiment indicators two stage OLS. Columns (1) is first stage; columns (2) to (4) are second stage; year and firm dummy variable are included; all control variables are included as lagged one year. Robust standard errors are reported in parentheses. The labels ***, **, and * indicate 1%, 5 %, and 10 % levels of significance, respectively.

Table 7 shows the regression results. The coefficients of the $newSentiInd$ and $newSentiPro$ variables are significantly positive in Column 1, indicating that the higher the media sentiment of listed firms in the industry and province, the higher the mean value of the sentiment of the listed firms. The Cragg−Donald F statistic equals 185.966, which is much larger than the critical value, and this statistic rejects the hypothesis that the instrumental variables are weak at the 1% level. The results of the second stage regression in Columns 2, 3, and 4 show that none of the values of the Sargan statistic reject the original hypothesis of instrumental variable exogeneity. The results of $newSenti$ continue to be significantly positive, in line with the results of the main regression.

5.3 Channels between news coverage and crash risk

Media sentiment affects the risk of a future stock price crash via two mechanisms. The first is through investor sentiment, which, in turn, affects stock crash risk. The second is that media coverage serves as an information intermediary, conveying true or false information on listed firms; investors influence the crash risk by interpreting the information they receive.

5.3.1 Information intermediation

Insiders have information that is not yet publicly available, which can be used to judge the value of the firm and predict future firm performance [39]. Insider sell-off behavior is positively associated with stock price crash risk [19]. The insider's choice to sell stocks sends negative signals to outside investors, thereby raising the probability of a future crash risk. Furthermore, the more overpriced a stock is, the greater the chance of a crash.

We thus divide the total sample into two subsamples, lower and higher groups, based on 30% and 70% quartiles of insiders' net stock sales, respectively. The regression results for various groups are shown in Table 8. The coefficients of the high net selling subgroups are significant, whereas those coefficients of the low net selling subgroups are insignificant. Thus, insiders sell more stocks, thereby amplifying the impact of media sentiment on the risk of a future stock price crash, confirming *H3a*.

Table 8 Regression results for corporate insider trading

	Low			High		
	(1)	(2)	(3)	(4)	(5)	(6)
	crash	ncskew	duvol	crash	ncskew	duvol
L.newSenti	0.030	-0.394	-0.154	0.103***	0.336***	0.119**
	(0.168)	(0.333)	(0.224)	(0.037)	(0.114)	(0.059)
L.crash	-0.144***			-0.185***		
	(0.010)			(0.033)		
L.ncskew		-0.088***			-0.096***	
		(0.034)			(0.009)	
L.duvol			-0.072**			-0.109***
			(0.032)			(0.009)
L.ret	1.570	10.925***	7.660***	0.933**	8.500***	6.077***
	(1.382)	(2.656)	(1.897)	(0.454)	(0.954)	(0.661)
L.sigma	1.427*	3.266**	2.124*	0.541**	0.707	1.049***
	(0.744)	(1.595)	(1.135)	(0.214)	(0.485)	(0.326)
L.roa	-0.717	0.484	0.712	0.453*	1.598***	1.016***
	(0.880)	(1.754)	(1.204)	(0.235)	(0.498)	(0.336)
L.level	-0.036	0.014	-0.035	0.020	-0.120*	-0.104**
	(0.112)	(0.214)	(0.150)	(0.036)	(0.064)	(0.041)
L.size	-0.000	-0.000	-0.000*	0.000	0.000**	0.000***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Fixed effects:						
Year dummy	Yes	Yes	Yes	Yes	Yes	Yes
Firm dummy	Yes	Yes	Yes	Yes	Yes	Yes
N	5,488	5,488	5,488	5,884	5,884	5,884
R^2	0.035	0.064	0.069	0.043	0.036	0.055

This table reports panel estimates of stock crash risk on the news coverage sentiment by using the subsamples of high insider trading groups and low insider trading groups. The total sample is divided into two subsamples: lower and higher groups, based on 30% and 70% quartiles of insiders' net stock sales respectively. Column (1) to (3) are low subgroup, the dependent variables are CRASH, NCSKEW and DUVOL respectively; Column (4) to (6) are high subgroup, the dependent variables are CRASH, NCSKEW and DUVOL respectively. All control variables are included as lagged one year. Robust standard errors are reported in parentheses. The labels ***, **, and * indicate 1%, 5 %, and 10 % levels of significance, respectively.

We examine the quality of firm disclosure to determine if it would mitigate the bubble created by media and reduce the risk of a stock price crash. We followed the method of Kim and Verrecchia [26] (KV index) to measure the quality of information disclosure. The higher the KV index, the lower the quality of the information disclosure of listed firms.

We again divide the total sample into two subsamples based on the 30% and 70% quartiles of the KV index into lower and higher groups. The regression results for the different groups are presented in Table 9. The coefficients of the lower disclosure group are significant, whereas those coefficients of the higher disclosure group are insignificant. Thus, the effect of news coverage on the risk of a future stock price crash is significantly enhanced in firms with poor disclosure. Meanwhile, the effect of media sentiment was significantly weaker when the quality of firm disclosure is high. Thus, *H3b* is supported.

Table 9 Regression results for information disclosure quality

	Low			High		
	(1)	(2)	(3)	(4)	(5)	(6)
	crash	ncskew	duvol	crash	ncskew	duvol

L.newSenti	0.051	0.026	-0.093	0.114***	0.384**	0.196**
	(0.106)	(0.222)	(0.144)	(0.043)	(0.179)	(0.098)
L.crash	-0.114***			-0.162***		
	(0.019)			(0.019)		
L.ncskew		-0.098***			-0.111***	
		(0.021)			(0.015)	
L.duvol			-0.101***			-0.124***
			(0.020)			(0.016)
L.ret	2.489***	11.948***	8.111***	0.494	5.299***	4.534***
	(0.939)	(2.231)	(1.424)	(0.853)	(1.451)	(1.117)
L.sigma	1.725***	1.739*	1.502**	0.239	1.582**	1.326**
	(0.411)	(1.045)	(0.662)	(0.437)	(0.744)	(0.579)
L.roa	-0.430	-0.536	0.197	0.768*	1.134*	0.586
	(0.544)	(1.349)	(0.864)	(0.405)	(0.665)	(0.487)
L.level	-0.141***	0.037	0.040	-0.025	-0.218**	-0.205***
	(0.042)	(0.112)	(0.069)	(0.059)	(0.102)	(0.076)
L.size	0.000	0.000**	0.000	0.000	0.000	0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Fixed effects:						
Year dummy	Yes	Yes	Yes	Yes	Yes	Yes
Firm dummy	Yes	Yes	Yes	Yes	Yes	Yes
N	5,488	5,488	5,488	5,884	5,884	5,884
R^2	0.038	0.064	0.070	0.044	0.080	0.079

This table reports panel estimates of stock crash risk on the news coverage sentiment by using the subsamples of high information disclosure quality groups and low information disclosure quality groups. The total sample is divided into two subsamples: lower and higher groups, based on 30% and 70% quartiles of KV index respectively. Column (1) to (3) are low subgroup, the dependent variables are CRASH, NCSKEW and DUVOL respectively; Column (4) to (6) are high subgroup, the dependent variables are CRASH, NCSKEW and DUVOL respectively. All control variables are included as lagged one year. Robust standard errors are reported in parentheses. The labels ***, **, and * indicate 1%, 5 %, and 10 % levels of significance, respectively.

Next, we examine the Hypothesis H3c. Follow He et al. [19], we calculate the analysts coverage as the number of analysts forecast over past three-year. Then, we divide the total sample into two subsamples: lower coverage and higher coverage groups, according to 30% and 70% quantiles of analysts' coverage to firms. Table 10 presents the regression results.

Table 10 Regression results for analysts' coverage

	Low			High		
	(1)	(2)	(3)	(4)	(5)	(6)
	crash	ncskew	duvol	crash	ncskew	duvol
L.newSenti	0.129**	0.543**	0.413**	0.029	0.132	0.147
	(0.058)	(0.221)	(0.167)	(0.078)	(0.220)	(0.146)
L.crash	-0.200***			-0.179***		
	(0.029)			(0.012)		
L.ncskew		-0.163***			-0.102***	
		(0.018)			(0.022)	
L.duvol			-0.156***			-0.121***
			(0.019)			(0.021)
L.ret	2.276	6.842***	6.149***	0.841	3.826*	2.842*
	(1.427)	(1.432)	(1.104)	(0.531)	(2.198)	(1.504)
L.sigma	1.788***	0.335	0.675	0.417	0.063	0.920
	(0.600)	(0.839)	(0.641)	(0.274)	(1.060)	(0.721)
L.roa	0.159	2.087***	0.600	0.480*	0.598	0.891
	(0.783)	(0.760)	(0.538)	(0.284)	(1.204)	(0.798)
L.level	0.029	-0.158	-0.131	-0.030	-0.051	-0.031
	(0.093)	(0.134)	(0.103)	(0.041)	(0.088)	(0.060)
L.size	0.000*	0.000	0.000**	0.000	0.000	-0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Fixed effects:						
Year dummy	Yes	Yes	Yes	Yes	Yes	Yes
Firm dummy	Yes	Yes	Yes	Yes	Yes	Yes
N	4618	4618	4618	4569	4569	4569
R^2	0.066	0.073	0.069	0.078	0.081	0.086

This table reports panel estimates of stock crash risk on the news coverage sentiment by using the subsamples of low analysts' attention groups and high analysts' attention groups. We used the total number of analysts coverage of firms to measure analysts' attention. We used the 30% and 70% quartiles of analysts' attention as the cut-off, and the firms were divided into low and high attention groups. Column (1) to (3) are low subgroup, the dependent variables are CRASH, NCSKEW and DUVOL respectively; Column (4) to (6) are high subgroup, the dependent variables are CRASH, NCSKEW and DUVOL respectively. All control variables are included as lagged one year. Robust standard errors are reported in parentheses. The labels ***, **, and * indicate 1%, 5 %, and 10 % levels of significance, respectively.

The coefficients of the low analyst coverage groups are significant, indicating that positive media coverage in the previous year increases the future stock price crash risk of firms with lower analyst coverage. The coefficients of the high analyst coverage groups are insignificant, and the impact of news coverage on stock price crash risk is attenuated. Thus, news coverage sentiment has a stronger impact on stock price crash risk when analysts pay less attention to a firm, supporting *H3c*.

5.3.2 Investor sentiment

In exploring the investor sentiment channel, we investigate whether an increase in number of retail investors could increase the emotional impact of the media. Table 11 shows the regression results for different groups, divided into low and high groups based on 30% and 70% quartiles of institutional holding.

Table 11 Regression results for different institutional holding group

	Low	High

	(1) crash	(2) ncskew	(3) duvol	(4) crash	(5) ncskew	(6) duvol
L.newSenti	0.092**	0.481***	0.358***	0.018	0.037	-0.027
	(0.044)	(0.176)	(0.124)	(0.094)	(0.220)	(0.144)
L.crash	-0.172***			-0.204***		
	(0.016)			(0.022)		
L.ncskew		-0.131***			-0.161***	
		(0.015)			(0.020)	
L.duvol			-0.170***			-0.182***
			(0.014)			(0.019)
L.ret	0.933	6.633***	5.416***	0.316	2.351	2.020
	(0.720)	(1.339)	(0.959)	(1.001)	(2.192)	(1.487)
L.sigma	0.697**	0.236	0.744	0.632	0.272	0.755
	(0.343)	(0.726)	(0.527)	(0.447)	(1.023)	(0.678)
L.roa	0.183	0.638	0.310	0.423	0.861	0.850
	(0.376)	(0.706)	(0.537)	(0.538)	(1.256)	(0.784)
L.level	-0.062	-0.072	-0.087*	-0.100	-0.346**	-0.181*
	(0.049)	(0.070)	(0.049)	(0.070)	(0.171)	(0.109)
L.size	0.000	0.000***	0.000***	-0.000	0.000	0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Fixed effects:						
Year dummy	Yes	Yes	Yes	Yes	Yes	Yes
Firm dummy	Yes	Yes	Yes	Yes	Yes	Yes
N	4,734	4,734	4,734	6,513	6,513	6,513
R^2	0.046	0.071	0.087	0.057	0.080	0.092

This table reports panel estimates of stock crash risk on the news coverage sentiment by using the subsamples of low institutional shareholding groups and high institutional shareholding groups. We used the 30% and 70% quartiles of the institutional shareholding as the cut-off point, and the firms were divided into low and high shareholding groups. Column (1) to (3) are low subgroup, the dependent variables are CRASH, NCSKEW and DUVOL respectively; Column (4) to (6) are high subgroup, the dependent variables are CRASH, NCSKEW and DUVOL respectively. All control variables are included as lagged one year. Robust standard errors are reported in parentheses. The labels ***, **, and * indicate 1%, 5 %, and 10 % levels of significance, respectively.

The coefficients for the low institutional holding subgroups are significant, whereas those coefficients for the high institutional holding subgroups are insignificant. The findings suggest that as retail investors increase their holdings, they tend to behave more irrationally, thereby amplifying the emotional tendency of media sentiment. Our results support *H4a*.

We consider direct proxy variables for investor sentiment and construct investor sentiment indicators according to Rhodes-Kropf, Robinson, and Viswanathan [41], dividing the total sample into pessimistic and optimistic groups based on 30% and 70% quartiles of investor sentiment. The regression results for different groups are shown in Table 12. The coefficients for the pessimistic subgroups are insignificant, whereas those coefficients for the optimistic subgroups are significant. The findings suggest that optimistic investor sentiment, which increases the likelihood that the current stock price is overvalued, increases the impact of news coverage on future crash risk, supporting *H4b*.

Table 12 Regression results for different investor sentiment

	Low			High		
	(1)	(2)	(3)	(4)	(5)	(6)
	crash	ncskew	duvol	crash	ncskew	duvol
L.newSenti	-0.011	-0.063	-0.097	0.176**	0.441**	0.254**
	(0.109)	(0.231)	(0.151)	(0.073)	(0.196)	(0.128)
L.crash	-0.178***			-0.173***		
	(0.019)			(0.016)		
L.ncskew		-0.144***			-0.104***	
		(0.019)			(0.018)	
L.duvol			-0.133***			-0.126***
			(0.019)			(0.017)
L.ret	2.540***	7.085***	4.014***	2.426***	6.721***	4.536***
	(0.677)	(1.893)	(1.318)	(0.759)	(1.778)	(1.205)
L.sigma	0.665	1.373	0.977	0.789**	0.826	0.019
	(0.437)	(0.929)	(0.606)	(0.371)	(0.925)	(0.634)
L.roa	0.755	1.529	1.205*	0.554	2.048**	1.666**
	(0.605)	(0.981)	(0.617)	(0.409)	(0.995)	(0.699)
L.level	0.036	-0.196*	-0.102	-0.003	-0.089	-0.087
	(0.067)	(0.106)	(0.066)	(0.055)	(0.128)	(0.088)
L.size	0.000	0.000*	0.000	0.000	0.000	0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Fixed effects:						
Year dummy	Yes	Yes	Yes	Yes	Yes	Yes
Firm dummy	Yes	Yes	Yes	Yes	Yes	Yes
N	4,881	4,881	4,881	5,863	5,863	5,863
R^2	0.047	0.078	0.085	0.050	0.071	0.083

This table reports panel estimates of stock crash risk on the news coverage sentiment by using the subsamples of pessimistic investor groups, and optimistic investor groups. We used the 30% and 70% quartiles of the investor sentiment as the cut-off point, and the firms were divided into low and high shareholding groups. Column (1) to (3) are low subgroup, the dependent variables are CRASH, NCSKEW and DUVOL respectively; Column (4) to (6) are high subgroup, the dependent variables are CRASH, NCSKEW and DUVOL respectively. All control variables are included as lagged one year. Robust standard errors are reported in parentheses. The labels ***, **, and * indicate 1%, 5 %, and 10 % levels of significance, respectively.

To investigate whether disagreement between professional analysts and retail investors also enhances the risk of a stock price crash, we add the cross-term of analysts' sentiment and media sentiment into the regression (Table 13).

The proxy variable used in this study to represent consistency and disagreement is the cross-term of analysts and news sentiment. When there are significant differences between the two opinions, often one is less than 0 and the other is greater than 0, so the cross-term is negative and represents the differences in opinions. Conversely, the two types of views have the same symbol and positive multiplication, implying that they are consistent.

Table 13 Regression of adding analyst and media coverage sentiment cross terms

	(1)	(2)	(3)
	crash	ncskew	duvol
L.newSenti	0.186*	0.360**	0.266
	(0.109)	(0.171)	(0.158)
L.anaSenti	0.039***	0.057**	0.049***
	(0.011)	(0.024)	(0.016)

L.newSenti#L.anaSenti	-0.247*	-0.172*	-0.156*
	(0.145)	(0.100)	(0.092)
L.crash	-0.151***		
	(0.009)		
L.ncskew		-0.102***	
		(0.009)	
L.duvol			-0.111***
			(0.009)
L.ret	0.942**	8.677***	6.188***
	(0.426)	(0.867)	(0.603)
L.sigma	0.524**	0.874*	1.120***
	(0.206)	(0.460)	(0.315)
L.roa	0.285	1.286***	0.813**
	(0.231)	(0.484)	(0.325)
L.level	0.006	-0.141**	-0.109***
	(0.036)	(0.062)	(0.041)
L.size	0.000	0.000*	0.000
	(0.000)	(0.000)	(0.000)
Fixed effects:			
Year dummy	Yes	Yes	Yes
Firm dummy	Yes	Yes	Yes
N	15,753	15,753	15,753
R^2	0.037	0.064	0.069

This table reports panel estimates of stock crash risk on the interaction between the news coverage sentiment and the analysis report sentiment. The dependent variables in column (1), (2) and (3) are CRASH, NCSKEW and DUVOL respectively. All control variables are included as lagged one year. Robust standard errors are reported in parentheses. The labels ***, **, and * indicate 1%, 5 %, and 10 % levels of significance, respectively.

The results of the regression are reported in Table 13. The results demonstrate the impact of analyst-rated sentiment on future stock price crash risk. The coefficient of the cross-term is negative, consistent with our theoretical hypothesis that the future stock price crash risk decreases when media sentiment and analyst sentiments are consistent, and increases otherwise. The regression results show that divergence of opinions in the market could increase media sentiment tendencies, thereby supporting *H4c*.

5.4 Robust test

To test the robustness of the indicators, we replace the main explanatory variable of the previous regression ($newSenti$), which is the weighted average of news sentiment, with an equally weighted average ($newSenti2$) and repeat the previous regressions. The results are shown in Tables 14–23 and are consistent with those in the previous contents.

To test the robustness of the indicators, we replace the main explanatory variable of the previous regression ($newSenti$), which is the weighted average of news sentiment, with an equally weighted average ($newSenti2$) and repeat the previous regressions. The results are shown in Tables 14–23 and are consistent with those in the previous contents.

Table 14 Robust test: Baseline regression

	(1)	(2)	(3)	(4)	(5)	(6)

	Crash	Crash	Ncskew	Ncskew	Duvol	Duvol
L.newSenti2	0.100***	0.078**	0.561***	0.221**	0.317***	0.151**
	(0.037)	(0.038)	(0.091)	(0.106)	(0.060)	(0.071)
L.crash		-0.145***		-0.092***		
		(0.008)		(0.008)		
L.ncskew						-0.101***
						(0.008)
L.duvol				8.451***		6.011***
				(0.814)		(0.565)
L.ret		0.948**		0.794*		1.015***
		(0.391)		(0.427)		(0.288)
L.sigma		-0.589***		1.303***		0.734**
		(0.188)		(0.431)		(0.292)
L.roa		0.435**		-0.142**		-0.114***
		(0.210)		(0.057)		(0.037)
L.level		0.019		0.000*		0.000
		(0.031)		(0.000)		(0.000)
L.size		0.000		0.000*		0.000
		(0.000)		(0.000)		(0.000)
Fixed effects:						
Year dummy	Yes	Yes	Yes	Yes	Yes	Yes
Firm dummy	Yes	Yes	Yes	Yes	Yes	Yes
N	20,107	18,456	20,107	18,456	20,107	18,456
R^2	0.000	0.037	0.002	0.062	0.002	0.068

This table reports robust baseline regression estimates of stock crash risk on the quantitatively weighted news coverage sentiment. The sample of CRASH is used for columns (1)-(2), the sample of NCSKEW is used for columns (3)-(4) and that of DUVOL is used columns (5)-(6). Columns (1), (3) and (5) only include the news coverage sentiment; columns (2), (4) and (6) control for lagged crash risk and firm characteristic; year and firm dummy variable are included; all control variables are included as lagged one year. Robust standard errors are reported in parentheses. The labels ***, **, and * indicate 1%, 5 %, and 10 % levels of significance, respectively.

Table 15 Robust test: Positive sentiment and negative sentiments of news coverage

			Panel A: One lag period			
	(1)	(2)	(3)	(4)	(5)	(6)
	crash	ncskew	duvol	crash	ncskew	duvol
L.newPos2	0.161***	0.331**	0.228**			
	(0.061)	(0.164)	(0.113)			
L.newNeg2				0.066	0.180	0.045
				(0.068)	(0.154)	(0.105)
L.crash	-0.147***			-0.147***		
	(0.008)			(0.009)		
L.ncskew		-0.095***			-0.099***	
		(0.008)			(0.009)	
L.duvol			-0.104***			-0.107***
			(0.008)			(0.008)
L.ret	1.058***	8.719***	6.152***	0.896**	8.770***	6.304***
	(0.400)	(0.830)	(0.577)	(0.404)	(0.844)	(0.585)
L.sigma	-0.627***	0.937**	1.058***	-0.550***	0.966**	1.160***
	(0.194)	(0.439)	(0.298)	(0.192)	(0.440)	(0.298)
L.roa	0.417*	1.291***	0.773**	0.462**	1.487***	0.845***
	(0.220)	(0.454)	(0.301)	(0.218)	(0.452)	(0.307)
L.level	0.018	-0.116**	-0.103***	0.010	-0.163***	-0.141***
	(0.032)	(0.057)	(0.038)	(0.034)	(0.060)	(0.040)
L.size	0.000	0.000*	0.000	0.000	0.000**	0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Fix effects:						
Year dummy	Yes	Yes	Yes	Yes	Yes	Yes
Firm dummy	Yes	Yes	Yes	Yes	Yes	Yes
N	17,363	17,363	17,363	16,276	16,276	16,276
R^2	0.036	0.060	0.067	0.036	0.062	0.068
			Panel B: Same period			
	(1)	(2)	(3)	(4)	(5)	(6)
	crash	ncskew	duvol	crash	ncskew	duvol
newPos2	-0.065	0.006	0.042			
	(0.085)	(0.180)	(0.123)			
newNeg2				-0.303***	-0.739***	-0.531***
				(0.080)	(0.154)	(0.108)
L.crash	-0.153***			-0.154***		
	(0.009)			(0.009)		
L.ncskew		-0.102***			-0.105***	
		(0.009)			(0.009)	
L.duvol			-0.108***			-0.108***
			(0.009)			(0.009)
L.ret	0.363	4.913***	3.543***	0.423	5.466***	3.870***
	(0.395)	(0.844)	(0.594)	(0.407)	(0.863)	(0.610)
L.sigma	-0.508***	0.146	0.299	-0.519***	0.104	0.297
	(0.190)	(0.446)	(0.307)	(0.192)	(0.448)	(0.310)
L.roa	0.661***	1.581***	0.894***	0.524**	1.468***	0.919***
	(0.207)	(0.428)	(0.296)	(0.207)	(0.439)	(0.304)
L.level	0.053	-0.110*	-0.104**	0.070**	-0.083	-0.083**
	(0.033)	(0.060)	(0.041)	(0.032)	(0.059)	(0.040)

L.size	-0.000	0.000	0.000	-0.000	0.000	0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Fix effects:						
Year dummy	Yes	Yes	Yes	Yes	Yes	Yes
Firm dummy	Yes	Yes	Yes	Yes	Yes	Yes
N	17,363	17,363	17,363	16,276	16,276	16,276
R^2	0.041	0.066	0.072	0.037	0.064	0.070

This table reports robust regression estimates of stock crash risk on the news coverage positive and negative sentiment indicators. Panel A is one year lag; Panel B is the same year; year and firm dummy variable are included; all control variables are included as lagged one year. Robust standard errors are reported in parentheses. The labels ***, **, and * indicate 1%, 5 %, and 10 % levels of significance, respectively.

Table 16 Robust test: Quarterly crash risk for different future periods

	(1) crash	(2) crash	(3) crash	(4) crash	(5) crash	(6) crash	(7) crash	(8) crash
newPos2	-0.049							
	(0.038)							
L.newPos2		0.096**						
		(0.038)						
L2.newPos2			0.084**					
			(0.039)					
L3.newPos2				0.025				
				(0.038)				
newNeg2					-0.248***			
					(0.035)			
L.newNeg2						0.034		
						(0.033)		
L2.newNeg2							-0.015	
							(0.035)	
L3.newNeg2								0.027
								(0.033)
L.crash	-0.048***	-0.049***	-0.047***	-0.047***	-0.049***	-0.049***	-0.048***	-0.045***
	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)
L.ret	1.921***	1.909***	2.445***	3.096***	1.837***	1.773***	2.562***	3.184***
	(0.447)	(0.450)	(0.502)	(0.508)	(0.457)	(0.458)	(0.505)	(0.525)
L.sigma	-1.686***	-1.464***	-2.182***	-2.174***	-1.780***	-1.411***	-1.931***	-2.287***
	(0.237)	(0.234)	(0.259)	(0.262)	(0.239)	(0.239)	(0.263)	(0.264)
L.roa	-0.091**	-0.123***	-0.122***	-0.121***	-0.062	-0.124***	-0.141***	-0.118**
	(0.046)	(0.046)	(0.047)	(0.046)	(0.047)	(0.046)	(0.049)	(0.047)
L.level	-0.037**	-0.034*	-0.029*	-0.018	-0.036**	-0.045***	-0.038**	-0.029
	(0.016)	(0.018)	(0.017)	(0.019)	(0.017)	(0.017)	(0.017)	(0.018)
L.size	0.000	-0.000	0.000	0.000	0.000	0.000	0.000	0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Fix effects								
Quarter dummy	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Firm dummy	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	58,530	58,530	58,530	58,530	59,748	59,748	59,748	59,748
R^2	0.038	0.037	0.039	0.038	0.040	0.039	0.039	0.038

This table reports robust regression estimates of quarterly stock crash risk on the news coverage positive and negative sentiment indicators. Columns (1) to (4) are positive indicators; columns (5) to (8) are negative indicators; quarter and firm dummy variable are included; all control variables are included as lagged one year. Robust standard errors are reported in parentheses. The labels ***, **, and * indicate 1%, 5 %, and 10 % levels of significance, respectively.

Table 17 Robust test: Two stages OLS regression

	First stage	Second stage		
	(1)	(2)	(3)	(4)
	newsenti	crash	ncskew	duvol
L.newSenti2		0.092**	0.994*	0.419*
		(0.045)	(0.456)	(0.246)
L.newSentiInd	0.494***			
	(0.020)			
L.newSentiPro	0.476***			
	(0.037)			
Controls	Yes	Yes	Yes	Yes
Cragg-Donald Wald F	185.966***			
Sargan chi(p)		0.121(0.788)	0.004(0.947)	0.254(0.614)
N	16,845	16,845	16,845	16,845
R^2	0.212	0.035	0.059	0.064

This table reports robust regression estimates of stock crash risk on the news coverage sentiment indicators two stage OLS. Columns (1) is first stage; columns (2) to (4) are second stage; year and firm dummy variable are included; all control variables are included as lagged one year. Robust standard errors are reported in parentheses. The labels ***, **, and * indicate 1%, 5 %, and 10 % levels of significance, respectively.

Table 18 Robust test: Regression results for corporate insider trading

	Low			High		
	(1)	(2)	(3)	(4)	(5)	(6)
	crash	ncskew	duvol	crash	ncskew	duvol
L.newSenti2	0.011	-0.394	-0.154	0.097**	0.336***	0.119**
	(0.052)	(0.333)	(0.224)	(0.048)	(0.114)	(0.059)
L.crash	-0.143***			-0.174***		
	(0.009)			(0.031)		
L.ncskew		-0.091***			-0.094***	
		(0.032)			(0.009)	
L.duvol			-0.084***			-0.104***
			(0.030)			(0.009)
L.ret	1.834	10.422***	7.613***	1.020**	8.353***	5.892***
	(1.346)	(2.642)	(1.858)	(0.429)	(0.917)	(0.638)
L.sigma	-1.305*	2.975*	2.061*	-0.586***	0.550	0.891***
	(0.721)	(1.559)	(1.090)	(0.205)	(0.467)	(0.313)
L.roa	-0.316	1.063	0.783	0.516**	1.509***	0.876***
	(0.801)	(1.615)	(1.111)	(0.218)	(0.465)	(0.317)
L.level	-0.027	-0.058	-0.020	0.028	-0.138**	-0.108***
	(0.108)	(0.208)	(0.144)	(0.033)	(0.062)	(0.040)
L.size	-0.000	-0.000	-0.000**	0.000	0.000**	0.000***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Fixed effects:						
Year dummy	Yes	Yes	Yes	Yes	Yes	Yes
Firm dummy	Yes	Yes	Yes	Yes	Yes	Yes
N	5,488	5,488	5,488	5,884	5,884	5,884
R^2	0.035	0.065	0.071	0.043	0.047	0.067

This table reports robust panel estimates of stock crash risk on the news coverage sentiment by using the subsamples of high insider trading groups and low insider trading groups. The total sample is divided into two subsamples: lower and higher groups, based on 30% and 70% quartiles of insiders' net stock sales respectively. Column (1) to (3) are low subgroup, the dependent variables are CRASH, NCSKEW and DUVOL respectively; Column (4) to (6) are high subgroup, the dependent variables are CRASH, NCSKEW and DUVOL respectively. All control variables are included as lagged one year. Robust standard errors are reported in parentheses. The labels ***, **, and * indicate 1%, 5 %, and 10 % levels of significance, respectively.

Table 19 Robust test: Regression results for information disclosure quality

		Low			High	
	(1)	(2)	(3)	(4)	(5)	(6)
	crash	ncskew	duvol	crash	ncskew	duvol
L.newSenti2	0.011	0.026	-0.093	0.122***	0.384**	0.196**
	(0.098)	(0.222)	(0.144)	(0.039)	(0.179)	(0.098)
L.crash	-0.120***			-0.157***		
	(0.017)			(0.018)		
L.ncskew		-0.094***			-0.101***	
		(0.020)			(0.014)	
L.duvol			-0.100***			-0.117***
			(0.018)			(0.015)
L.ret	2.396***	10.914***	7.517***	0.510	5.451***	4.603***
	(0.884)	(2.111)	(1.351)	(0.787)	(1.366)	(1.061)
L.sigma	-1.702***	1.233	1.136*	-0.165	1.549**	1.347**
	(0.384)	(0.995)	(0.631)	(0.410)	(0.712)	(0.548)
L.roa	-0.277	-0.455	-0.010	0.665*	1.002	0.514
	(0.495)	(1.223)	(0.789)	(0.374)	(0.632)	(0.468)
L.level	0.140***	0.007	0.033	-0.007	-0.211**	-0.205***
	(0.040)	(0.112)	(0.067)	(0.055)	(0.097)	(0.072)
L.size	0.000	0.000**	0.000	0.000	0.000	0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Fixed effects:						
Year dummy	Yes	Yes	Yes	Yes	Yes	Yes
Firm dummy	Yes	Yes	Yes	Yes	Yes	Yes
N	5,906	5,906	5,906	6,243	6,243	6,243
R^2	0.040	0.064	0.071	0.047	0.080	0.080

This table reports robust panel estimates of stock crash risk on the news coverage sentiment by using the subsamples of high information disclosure quality groups and low information disclosure quality groups. The total sample is divided into two subsamples: lower and higher groups, based on 30% and 70% quartiles of KV index respectively. Column (1) to (3) are low subgroup, the dependent variables are CRASH, NCSKEW and DUVOL respectively; Column (4) to (6) are high subgroup, the dependent variables are CRASH, NCSKEW and DUVOL respectively. All control variables are included as lagged one year. Robust standard errors are reported in parentheses. The labels ***, **, and * indicate 1%, 5 %, and 10 % levels of significance, respectively.

Table 20 Robust test: Regression results for different analysts coverage

	Low			High		
	(1)	(2)	(3)	(4)	(5)	(6)
	crash	ncskew	duvol	crash	ncskew	duvol
L.newSenti2	0.074	0.515***	0.345***	0.029	-0.027	-0.073
	(0.076)	(0.154)	(0.107)	(0.113)	(0.267)	(0.190)
L.crash	-0.178***			-0.203***		
	(0.012)			(0.027)		
L.ncskew		-0.136***			-0.182***	
		(0.012)			(0.025)	
L.duvol			-0.139***			-0.192***
			(0.012)			(0.024)
L.ret	0.713	7.801***	5.910***	2.251*	6.580**	4.741**
	(0.507)	(1.019)	(0.729)	(1.342)	(3.080)	(2.039)
L.sigma	-0.351	-0.211	-0.666	-1.794***	-1.440	-1.033
	(0.265)	(0.594)	(0.421)	(0.566)	(1.363)	(0.868)
L.roa	0.569**	1.054*	0.414	0.179	0.860	0.371
	(0.277)	(0.556)	(0.380)	(0.742)	(1.585)	(1.046)
L.level	-0.010	-0.106	-0.110*	0.020	-0.181	-0.090
	(0.040)	(0.087)	(0.062)	(0.086)	(0.206)	(0.128)
L.size	0.000	0.000**	0.000***	0.000	0.000*	-0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Fixed effects:						
Year dummy	Yes	Yes	Yes	Yes	Yes	Yes
Firm dummy	Yes	Yes	Yes	Yes	Yes	Yes
N	5,028	5,028	5,028	4,809	4,809	4,809
R^2	0.067	0.069	0.068	0.071	0.118	0.119

This table reports robust panel estimates of stock crash risk on the news coverage sentiment by using the subsamples of low analysts' attention groups and high analysts' attention groups. We used the total number of analysts coverage of firms to measure analysts' attention. We used the 30% and 70% quartiles of analysts' attention as the cut-off, and the firms were divided into low and high attention groups. Column (1) to (3) are low subgroup, the dependent variables are CRASH, NCSKEW and DUVOL respectively; Column (4) to (6) are high subgroup, the dependent variables are CRASH, NCSKEW and DUVOL respectively. All control variables are included as lagged one year. Robust standard errors are reported in parentheses. The labels ***, **, and * indicate 1%, 5 %, and 10 % levels of significance, respectively.

Table 21 Robust test: Regression results for different institutional holding group

	Low			High		
	(1)	(2)	(3)	(4)	(5)	(6)
	crash	ncskew	duvol	crash	ncskew	duvol
L.newSenti2	0.102**	0.481***	0.358***	0.028	0.037	-0.027
	(0.044)	(0.176)	(0.124)	(0.106)	(0.220)	(0.144)
L.crash	-0.198***			-0.174***		
	(0.020)			(0.015)		
L.ncskew		-0.131***			-0.161***	
		(0.015)			(0.020)	
L.duvol			-0.170***			-0.182***
			(0.014)			(0.019)
L.ret	0.698	6.795***	5.479***	0.933	2.429	1.849
	(0.949)	(1.287)	(0.932)	(0.670)	(2.091)	(1.426)
L.sigma	-0.777*	0.043	0.806	-0.621*	0.018	0.540
	(0.423)	(0.696)	(0.503)	(0.335)	(0.970)	(0.642)
L.roa	0.464	0.860	0.321	0.277	0.723	0.699
	(0.509)	(0.655)	(0.505)	(0.362)	(1.178)	(0.737)
L.level	-0.094	-0.074	-0.085*	0.064	-0.367**	-0.156
	(0.065)	(0.068)	(0.048)	(0.048)	(0.159)	(0.104)
L.size	-0.000	0.000***	0.000***	0.000	0.000	0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Fixed effects:						
Year dummy	Yes	Yes	Yes	Yes	Yes	Yes
Firm dummy	Yes	Yes	Yes	Yes	Yes	Yes
N	5,120	5,120	5,120	6,959	6,959	6,959
R^2	0.046	0.071	0.086	0.067	0.084	0.095

This table reports robust panel estimates of stock crash risk on the news coverage sentiment by using the subsamples of low institutional shareholding groups and high institutional shareholding groups. We used the 30% and 70% quartiles of the institutional shareholding as the cut-off point, and the firms were divided into low and high shareholding groups. Column (1) to (3) are low subgroup, the dependent variables are CRASH, NCSKEW and DUVOL respectively; Column (4) to (6) are high subgroup, the dependent variables are CRASH, NCSKEW and DUVOL respectively. All control variables are included as lagged one year. Robust standard errors are reported in parentheses. The labels ***, **, and * indicate 1%, 5 %, and 10 % levels of significance, respectively.

726 Table 22 Robust test: Regression results for different investor sentiment

	Low			High		
	(1)	(2)	(3)	(4)	(5)	(6)
	crash	ncskew	duvol	crash	ncskew	duvol
L.newSenti2	-0.039	0.022	-0.017	0.173***	0.544***	0.348**
	(0.101)	(0.216)	(0.147)	(0.063)	(0.210)	(0.137)
L.crash	-0.177***			-0.165***		
	(0.018)			(0.015)		
L.ncskew		-0.141***			-0.103***	
		(0.018)			(0.017)	
L.duvol			-0.132***			-0.125***
			(0.018)			(0.016)
L.ret	2.871***	6.671***	3.952***	2.329***	6.885***	4.676***
	(0.947)	(1.836)	(1.283)	(0.710)	(1.703)	(1.149)
L.sigma	-0.537	0.860	0.689	0.784**	0.694	0.072
	(0.422)	(0.901)	(0.592)	(0.359)	(0.865)	(0.590)
L.roa	0.751	1.287	0.873	0.676*	2.094**	1.581**
	(0.574)	(0.945)	(0.612)	(0.390)	(0.875)	(0.616)
L.level	0.035	-0.193*	-0.109*	0.023	-0.084	-0.078
	(0.064)	(0.101)	(0.064)	(0.051)	(0.120)	(0.082)
L.size	0.000	0.000	0.000	0.000	0.000	0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Fixed effects:						
Year dummy	Yes	Yes	Yes	Yes	Yes	Yes
Firm dummy	Yes	Yes	Yes	Yes	Yes	Yes
N	5,213	5,213	5,213	6,258	6,258	6,258
R^2	0.047	0.077	0.085	0.051	0.082	0.085

This table reports robust panel estimates of stock crash risk on the news coverage sentiment by using the subsamples of pessimistic investor groups, and optimistic investor groups. We used the 30% and 70% quartiles of the investor sentiment as the cut-off point, and the firms were divided into low and high shareholding groups. Column (1) to (3) are low subgroup, the dependent variables are CRASH, NCSKEW and DUVOL respectively; Column (4) to (6) are high subgroup, the dependent variables are CRASH, NCSKEW and DUVOL respectively. All control variables are included as lagged one year. Robust standard errors are reported in parentheses. The labels ***, **, and * indicate 1%, 5 %, and 10 % levels of significance, respectively.

727

Table 23 Robust test: Regression of adding analyst and media coverage sentiment cross terms

	(1) crash	(2) ncskew	(3) duvol
L.newSenti2	0.238**	0.618**	0.336*
	(0.117)	(0.269)	(0.185)
L.anaSenti	0.035***	0.061***	0.050***
	(0.011)	(0.023)	(0.016)
L.newSenti2#L.anaSenti	-0.121**	-0.424*	-0.186*
	(0.057)	(0.235)	(0.105)
L.crash	-0.150***		
	(0.009)		
L.ncskew		-0.099***	
		(0.009)	
L.duvol			-0.108***
			(0.009)
L.ret	0.933**	8.405***	5.992***
	(0.406)	(0.839)	(0.584)
L.sigma	0.528***	0.664***	0.953***
	(0.198)	(0.144)	(0.303)
L.roa	0.377*	1.260***	0.688**
	(0.217)	(0.453)	(0.308)
L.level	0.014	-0.146**	-0.112***
	(0.034)	(0.060)	(0.039)
L.size	0.000	0.000**	0.000
	(0.000)	(0.000)	(0.000)
Fixed effects:			
Year dummy	Yes	Yes	Yes
Firm dummy	Yes	Yes	Yes
N	16,829	16,829	16,829
R^2	0.039	0.066	0.072

This table reports robust panel estimates of stock crash risk on the interaction between the news coverage sentiment and the analysis report sentiment. The dependent variables in column (1), (2) and (3) are CRASH, NCSKEW and DUVOL respectively. All control variables are included as lagged one year. Robust standard errors are reported in parentheses. The labels ***, **, and * indicate 1%, 5 %, and 10 % levels of significance, respectively.

6.Conclusions

We construct a deep learning model of stock news sentiment recognition based on the advanced approach of financial knowledge dictionary and NLP (BERT-based pretraining) technology. We use this model to calculate the sentiment indicators of all stocks from 2011 to 2020. Subsequently, we analyze the impact of media sentiment on future stock price crash risk and its heterogeneity.

We find that average media sentiment exacerbates the risk of future stock price crashes. The heterogeneity results indicate that positive coverage significantly increases future stock price crash risk, whereas negative coverage has limited effect. However, negative coverage is highly correlated with current stock price crash risk. We also investigate the information intermediation and investor sentiment channels by which media sentiment affects the risk of a crash. The results show that more net insider sales, lower information transparency, and less analyst coverage amplify the impact of media sentiment on future crash risk, which is consistent with the information intermediation channel. Additionally, more retail investor positions, more active investor sentiment, and divergence between analysts' opinions and news amplify the impact of news sentiment on the risk of a future stock price crash, consistent with the investor sentiment channel.

Our finding that positive media sentiment can lead to an extreme outcome in the stock market is useful for both regulators and investors. Our examination of the impact of media sentiment on the future stock price crash risk adopted both behavioral finance and information economics perspectives, and revealed that investors' irrational and excessive optimism could be a major cause of stock price bubbles and crashes in China's stock market, which is dominated by retail investors who are restricted from short selling.

In terms of research methodology, our study combined advanced deep learning and dictionary methods, fully utilizing computer performance and intelligence to significantly improve the recognition accuracy and efficiency of massive amounts of sentiment data. To the best of our knowledge, we are the first to combine deep learning and dictionary methods for sentiment analysis in finance, thereby broadening the scope of sentiment analysis methods in finance.

Data Availability

The data copyright belongs to the GuoTai'an(CSMAR) and Wind, disclosing is not allowed.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

References

[1] Ahern K. R., Sosyura D. (2015) Rumor has it: sensational is min financial media. Review of Financial Studies, 28, 2050-2093.
[2] Ahern K. R., Sosyura D. (2014) Who Writes the News? Corporate Press Releases during Merger Negotiations. Journal of Finance, 69, 241-291.
[3] Andrade S., Bian J., and Burch T. (2013) Analyst Coverage, Information, and 620 Bubbles. Journal of Financial and Quantitative Analysis, 48, 1573-1605.
[4] Bushee B. J., Core J. E., Guay W., et al. (2010) The Role of the Business Press as an Information Intermediary. Journal of Accounting Research, 48, 1-19.
[5] Chan W. S. (2003) Stock price reaction to news and no-news: drift and reversal after headlines. Journal of Financial Economics, 70, 223-260.
[6] Chen, J., Hong H., Stein J. C. (2001) Forecasting Crashes: Trading Volume, Past Returns and Conditional Skewness in Stock Price. Journal of Financial Economics, 61, 345-381.
[7] Cui X., Sensoy A., Nguyen D. K., Yao S. and Wu Y. (2022) Positive information shocks, investor behavior and stock price crash risk. Journal of Economic Behavior and Organization, 197, 493-518.
[8] Das S. R. (2014) Text and content: language analytics in finance. Foundations and Trend in Finance, 8, 145-261.
[9] Drake M. S., Guest N. M., Twedt B. J. (2014) The Media and Mispricing:The Role of the Business Press in the Pricing of Accounting Information. Accounting Review, 89, 1673-1701.

[10] El-Haj M., Rayson P., Walker M., Young S. and Simaki V. (2019) In search of meaning: lesson, resources and next steps for computational analysis of financial disclosure. Journal of Business Finance & Accounting, 46, 235-306.

[11] Engelberg J., Sasseville C., Williams J. (2012) Market Madness? The Case of Mad Money. Management Science, 58, 351-364 .

[12] Ertugrul M., Lei J., Qiu J. and Wan C. (2017) Annual Report Readability, Tone Ambiguity, and the Cost of Borrowing. Journal of Financial and Quantitative Analysis, 52, 811-836

[13] Fraiberger S. P., Lee D., Puy D. (2021) Media sentiment and international asset prices. Journal of International Economics, 133, 103526.

[14] Gan B., Alexeev V., Bird R. (2020) Sensitivity to sentiment: News vs social media. International Review of Financial Analysis, 67, 101390.

[15] Harvey C. R., Siddique A. (2000) Conditional Skewness in Asset Pricing Tests. Journal of Finance, 55, 1263-1295.

[16] Hassan T., Hollander S., van Lent L., Tahoun A. (2019) Firm-Level Political Risk: Measurement and Effects. The Quarterly Journal of Economics, 134, 2135–2202.

[17] Hatzilygeroudis I., Perikos I. (2016) A Classifier Ensemble Approach to Detect Emotions Polarity in Social Media. WEBIST.

[18] He G., Bai L. and Ren H. M. (2019) Analyst Coverage and Future Stock Price Crash Risk. Journal of Applied Accounting Research, 20, 63-77.

[19] He G., Ren M. H., Taffler R., (2021) Do corporate insiders trade of future stock price crash risk? Review of Quantitative Finance & Accounting, 56, 1561–1591

[20] Hillert A., Jacobs H., Möller S. (2014) Media Makes Momentum. Review of Financial Studies, 27(12), 3467-3501.

[21] Hong, H. and J. C. Stein (2003). Differences of Opinion, Short-Sales Constraints and Market Crashes. Review of Financial Studies, 16, 487-525.

[22] Hutton, A. P., Marcus A. J. and Tehranian H. (2009). Opaque financial reports, R2, and crash risk. Journal of Financial Economics, 94, 67-86.

[23] Jiang, X, Y, Xu, N, Kam C. Chan, and Yi Z. (2013) Analyst Coverage, Optimism, and Stock Price Crash Risk: Evidence from China, Pacific- Basin Finance Journal, 25, 217-239.

[24] Joe J. R., Louis H. and Robinson D. (2009) Managers' and investors' responses to media exposure of board ineffectiveness. Journal of Financial & Quantitative Analysis, 44, 579-605

[25] Keshavarz H., Abadeh M. S. (2017) ALGA: Adaptive lexicon learning using genetic algorithm for sentiment analysis of microblogs. Knowledge-Based Systems, 122, 1-16.

[26] Kim O., and Verrecchia R. E. (2001) The Relation Among Disclosure, Returns and Trading. The Accounting Review, 76, 633-654.

[27] Kim J. B., Li Y. and Zhang L. (2011) CFOs versus CEOs: equity incentives and crashes. Journal of Financial Economics, 101, 713-730.

[28] Kothari S. P., Susan S. and Peter W. (2009) Do Managers Withhold Bad News? Journal of Accounting Research, 47, 241-276.

[29] Larcker D. F., Zakolyukina, A. A. (2012) Detecting deceptive discussions in conference calls. Journal of Accounting Research, 50, 495-540.

[30] Li C. T., Song F. M. and Zhang X. (2014) Analyst Following and Corporate Earnings Management: Evidence from China. Journal of Financial Research, 7, 124-139.

[31] Li F. (2010) The information content of forward-looking statements in corporate filing-a naïve bayesian machine learning approach. Journal of Accounting Research, 48, 1049-1102.

[32] Li J., Stewart C. M. (2006) R2 around the world: New theory and new tests. Journal of Financial Economics, 79(2), 257-292.
[33] Li Q., Wang J., Bao L. (2022) Media tone, bias, and stock price crash risk: evidence from China. Asia-Pacific Journal of Accounting and Economics, 29, 1-35.
[34] Liu G. R., Yan M. Y. and Jin Y. H. (2017) A test of the relationship between investor sentiment and stock market crash risk. Finance and Accounting Monthly, 5, 39-46.
[35] Loughran T. I. M. and Mcdonald B. (2016) Textual analysis in accounting and finance: a survey. Journal of Accounting Research. 54, 1187-1230.
[36] Noelle-Neumann E. (1974) The Spiral of Silence a Theory of Public Opinion. Journal of Communication, 24: 43-51.
[37] Odean T. (1998) Volume, Volatility, Price, and Profit. When All Traders Are Above Average. Journal of Finance, 53, 1887-1934.
[38] Peng Z, Zhang D. (2019) Bidirectional-GRU based on attention mechanism for aspect-level sentiment analysis. Proceedings of the 2019 11th International Conference on Machine Learning and Computing. New York, USA: ACM.
[39] Piotroski J. D., Roulstone D. T. (2005) Do Insider Trades Reflect Both Contrarian Beliefs and Superior Knowledge about Future Cash Flow Realizations? Journal of Accounting and Economics, 39, 55-81.
[40] Pouget S., Sauvagnat J. and Villeneuve S. (2017) A Mind Is a Terrible Thing to Change: Confirmatory Bias in Financial Markets. Review of Financial Studies, 30, 2066-2109.
[41] Rhodes-Kropf M., Robinson D., Viswanathan S. (2005) Valuation Waves and Merger Activity: The Empirical Evidence. Journal of Financial Economics, 77, 561-603.
[42] Rozin P., Royzman E. B. (2001) Negativity bias, negativity dominance, and contagion. Personality and Social Psychology Review, 5, 296-320.
[43] Schipper K. (1991) Analysts' Forecasts. Accounting Horizons, 5, 105-121.
[44] Solomon D. H. (2012). Selective Publicity and Stock Prices. Journal of Finance, 67, 599-638.
[45] Tetlock P. C. (2007) Giving content to investor sentiment: the role of media in the stock market. Journal of Finance, 62, 1139–68.
[46] Tetlock P. C., Saar-Tsechansky M. and Macskassy S. (2008) More than words: quantifying language to measure firms’ fundamentals. Journal of Finance, 63, 1437-1467.
[47] Tetlock P. C. (2010) Does Public Financial News Resolve Asymmetric Information? The Review of Financial Studies, 23, 3520–3557.
[48] Xu N., Li X., Yuan Q. and Chan K. C. (2014) Excess Perks and Stock Price Crash Risk: Evidence from China. Journal of Corporate Finance, 25, 419-434.
[49] Xu N., Jiang X., Yi Z. and Xu X. (2012) Conflicts of Interest, Analyst Optimism and Stock Price Crash Risk. Economic Research Journal, 7, 127-140.
[50] You J., Zhang B. and Zhang L. (2017) Who captures the power of the pen? The Review of Financial Studies, 31, 43-96.
[51] Zhang L, Zhou Y, Duan X. (2017) A hierarchical multi-input and output bi-GRU model for sentiment analysis on customer reviews. Proceedings of IOP Conference Series: Materials Science and Engineering. Berlin, German: Springer.
[52] Zhang L., Wang C., Yao H. (2021) Information contagion and stock price crash risk. Mathematical Problems in Engineering.