On the prediction of stock price crash risk using textual sentiment of management statement

Xiao Yao
Business School, Central University of Finance and Economics, Beijing, China
Dongxiao Wu
School of Management, Beijing Union University, Beijing, China
Zhiyong Li
School of Finance, Southwestern University of Finance and Economics, Chengdu, China, and
Haoxiang Xu
School of Economics, Central University of Finance and Economics, Beijing, China

Received 3 December 2022
Revised 17 April 2023
6 August 2023
Accepted 28 August 2023

Abstract

Purpose – Since stock return and volatility matters to investors, this study proposes to incorporate the textual sentiment of annual reports in stock price crash risk prediction.

Design/methodology/approach – Specific sentences gathered from management discussions and their subsequent analyses are tokenized and transformed into numeric vectors using textual mining techniques, and then the Naïve Bayes method is applied to score the sentiment, which is used as an input variable for crash risk prediction. The results are compared between a collection of predictive models, including linear regression (LR) and machine learning techniques.

Findings – The experimental results find that those predictive models that incorporate textual sentiment significantly outperform the baseline models with only accounting and market variables included. These conclusions hold when crash risk is proxied by either the negative skewness of the return distribution or down-to-up volatility (DUVOL).

Research limitations/implications – It should be noted that the authors' study focuses on examining the predictive power of textual sentiment in crash risk prediction, while other dimensions of textual features such as readability and thematic contents are not considered. More analysis is needed to explore the predictive power of textual features from various dimensions, with the most recent sample data included in future studies.

Originality/value – The authors' study provides implications for the information value of textual data in financial analysis and risk management. It suggests that the soft information contained within annual reports may prove informative in crash risk prediction, and the incorporation of textual sentiment provides an incremental improvement in overall predictive performance.

Keywords Stock market prediction, Stock price crash risk, Textual sentiment, Machine learning techniques, Management statement

Paper type Research paper

1. Introduction

Since February 2020, the global stock markets have experienced a series of stock market crashes due to the coronavirus disease 2019 (COVID-19) pandemic. During March 2020, the USA market

The authors would like to thank for the support from the Fundamental Research Funds for the Central Universities, China [JBK2103013]. This work was also supported by Financial Innovation Centre (Project No. 2022A0011) at Southwestern University of Finance and Economics, China. The authors thank the editors and the anonymous reviewers for their comments to improve this paper.

Statement on ethics: This study involves no human subjects, human data or tissue, or animals.

Conflicts of interest: All authors declare that they have no conflict of interest.

China Finance Review International
© Emerald Publishing Limited
2044-1398
DOI 10.1108/CFRI-12-2022-0250

has triggered the circuit breaker for four times in response to the crashes of S&P (Standard and Poor's) 500 index [1]. Increasing attention has been paid to the stock price crash risk from the market practitioners and academic researchers (Jin *et al.*, 2016, 2021; Hu and Wang, 2018; Xu and Zou, 2019; Liang *et al.*, 2020). Crash risk is essentially a measure of the negative skewness of stock price return distributions, which represents the expected stock price extreme downside movement. Crash risk is considered to be originated from the release of hidden bad news leading to investor heterogeneity (Chen *et al.*, 2001; Hong and Stein, 2003). The negative information possessed by the bearish investors is not incorporated in the stock price due to the short-sale constraint, and the bad news accumulates until a tipping point. Following studies argue that stock market crashes can be explained by the conflict of interests between investors and managers, suggesting that managers have incentives to withhold bad news for their personal interests and crashes are triggered when the accumulated bad news reaches a tipping point and comes out to the market (Jin and Myers, 2006; Hutton *et al.*, 2009).

When the bad news is released, stock price crashes accordingly. The empirical evidence shows that bad news hoarding ubiquitously exists in the firms which is not only impacted by the firm financial performance disclosed in the annual report, but also affected by the disclosure quality such as reporting transparency (Jin and Myers, 2006). For example, on Jan 31st 2020, the short-selling firm Muddy Waters Research published an anonymous reporting claiming that the Chinese coffee brand Luckin Coffee listed on Nasdaq forged their accounting reports to inflate the firm performance. Luckin Coffee officially acknowledged that the firm has fabricated the sales by nearly 300mn USD until Apr 2nd 2020, and the stock price immediately dropped by 80% approximately [2]. As stock price tends to crash suddenly and unexpectedly, investors may have to bear huge losses when the crash occurs. Moreover, the investors' sentiment could be seriously damaged which would bring overwhelming panic and cause a wider range of market crashes. Therefore, an accurate prediction of stock price crash risk is not only crucial for the investors to avoid unexpected losses, but also provide an early warning signal to the regulators to ensure the market stability.

In recent years, the importance of soft information has been recognized in the studies of accounting and finance. The textual sentiment of newspapers, online social media and corporate disclosures has been explored and applied in various ways thanks to the rapid development of textual mining techniques (Maghyereh and Abdoh, 2022). Previous studies have shown that the textual information embedded in the corporate annual reports is informative of firm operational and market performance (Tetlock, 2007; Loughran and Mcdonald, 2011). And it is also predictive in consumer corporate credit scoring and bankruptcy prediction (Jiang *et al.*, 2017; Mai *et al.*, 2019; Matin *et al.*, 2019).

In annual reports, corporates are required to present the views and opinions of their senior managers regarding the firm performance in the past year as well as the outlook in the following year in the section of management discussions and analysis (MD&A). If managers intend to hide bad news for their own interest such as promotion or higher bonus, firm's stock price crash risk tends to increase according to the bad news hoarding theory (Jin and Myers, 2006). In addition to the hard numbers disclosed in the annual reports, more attention has been paid to the predictive power of soft information such as textual sentiment embedded in the MD&A section (Loughran and Mcdonald, 2011), which is expected to be positive if managers are optimistic on the firm's future performance. Thus, we believe textual sentiment provides another angle for the investors to identify the firm risk, which could be interpreted by investors differently and subsequently triggers the crash when the bad news comes out (Hong and Stein, 2003). Since 2012, Chinese listed firms are required to disclose the management review and outlook of firm performance in the MD&A sections of annual reports. Different from the developed markets where institutional investors play a major role, a key feature of the China stock market is that it is dominated by small and medium-sized investors. Because such investors have less expertise in accounting reports compared to the institutional investors, they are expected to be more sensitive to the management statements while paying less attention to the complex quantitative numbers in the annual reports. Once the pessimistic sentiment is sensed by the small investors from the disclosure, they would be more panic to sell their shares. Thus, the textual sentiment embedded in the MD&A section is expected to be predictive for the stock price crash risk in the China market. Previous studies have placed the emphasis on exploring informative determinants of crash risk from the perspectives of causal effect identification based on the econometrics models (Callen and Fang, 2013; Kim *et al.*, 2014; Fu *et al.*, 2021), however, there is lack of studies on the prediction of stock price crash risk using sophisticated predictive models such as machine learning techniques. In this study, we are motivated to predict stock price crash risk by applying machine learning techniques with the incorporation of textual sentiment from annual reports. Our study does not attempt to establish the identification of any causal effect with respect to crash risk. Our research questions include twofolds: (1) if the prediction of crash risk can be improved by the application of machine learning techniques and (2) if the incorporation of textual sentiment from annual report brings incremental improvement in crash risk prediction?

Different from the traditional hard information such as the accounting ratios which can be directly applied in quantitative modeling, the soft information in the MD&A sections is not straightforward for mathematical calculation. Textual mining techniques are needed to transform the textual information to the modeling variables. A simple way for sentiment extraction is to identify the sentimental words and count their frequencies. For example, the sentimental words are normally categorized into positive and negative words, and a higher proportion of positive words indicate the overall sentiment is inclined to be positive and vice versa. However, this method relies on a careful design of the financial sentiment dictionary (Loughran and Mcdonald, 2011). As there is lack of an authoritative and publicly available dictionary for the Chinese Mandarin language, we propose to apply the word embedding and machine learning techniques to extract the textual sentiment. First, the textual context of MD&A sections in the annual reports is tokenized into individual sentences and cleansed, which are then transformed into mathematical vectors by word2vec (Mikolov *et al.*, 2013). Next, we label each sentence with the flag such as 1, 0 and -1 representing the sentiment of positive, neural and negative with the support of research assistants. Last, we apply the Naïve Bayes method to train the model and generate the predicted score of each sentence based on the weighted predicted probabilities by sentiment flag of each sentence. The final sentiment score of annual report is given as the averaged predicted scores of all sentences in the corresponding MD&A section. We collect a sample of 7,267 firm-year observations listed in Shanghai and Shenzhen stock exchanges from 2012 to 2017 and then match the extracted textual sentiment of annual reports as shown above. The hard information includes market and accounting variables are also merged into the sample to predict stock price crash risk, which is proxied by the negative conditional skewness (*NCSKEW*) and down-to-up volatility (*DUVOL*) of stock returns following previous studies (Chen *et al.*, 2001). To explore the predictive power of MD&A sentiment, we take the models using accounting and market variables as the baseline models and compare them with the treatment models incorporating textual sentiment and other hard information. The experimental results demonstrate that the incremental improvement of crash risk prediction is significant in both out-of-time and out-of-sample prediction, suggesting that the textual sentiment of annual reports should be applied to the prediction of stock price crash risk.

Our study makes contributions to in multiple folds. First, we find the modeling performance of crash risk prediction can be significantly improved with the inclusion of MD&A textual sentiment. The experimental results of out-of-time and out-of-sample prediction have shown that treatment models with MD&A textual sentiment included significantly outperform the baseline models in terms of mean absolute error (MAE) and root mean square error (RMSE) and the evidence holds for both proxies of crash risk. Next, our study demonstrates the advantages of machine learning techniques in crash risk prediction. We compare five predictive techniques including linear regression (LR) and another four machine learning techniques and find neural networks (NN) is the most competitive technique among all algorithms followed by random forest (RF) and gradient boosting decision tree (GBDT). Least square support vector regression (LS-SVR) is less competitive but still exhibits better performance than LR. The advantage is more significant when the sentiment of management statement is added. Previous studies on stock price crash risk have focused on investigating the unknown determinants, and our study presents the evidence that machine learning techniques such as NN are promising in crash risk prediction. However, it should be noticed that the purpose of our study is not running a horse-racing to find the best modeling technique for crash risk prediction. The purpose of this study is to explore the advantage of machine learning techniques to predict stock price crash risk. Last, our study provides implications to the information value of textual data in financial analysis and risk management. Textual sentiment of media reports and corporate disclosures has been recognized as a predictive indicator in financial market movement prediction and credit risk modeling. Our study is presumably the first large-sample study that explores the prediction of crash risk using machine learning techniques with the incorporation of textual sentiment from annual reports. The most relevant study in literature is Meng *et al.* (2017), which also document the information value of MD&A section in crash risk prediction in the China market, however, it does not investigate the textual sentiment of MD&A section. Instead, a vector normalization method is applied to measure the information value which is shown to be predictive of crash risk based on the econometric model. Our study differs from Meng *et al.* (2017) by measuring the textual sentiment using a naïve Bayes method combined with the application of machine learning technique to improve the crash risk prediction. Considering the unique feature of China market which is dominated by the small investors, our study provides implications to the importance of corporate disclosure for both quantitative accounting ratios and qualitative statements.

The rest of this paper is organized as follows. Section 2 briefly reviews the relevant literature. Section 3 introduces the sample and variables, and an experimental setup is described in Section 4. The experimental results are presented and discussed in Section 5. Section 6 concludes the paper.

2. Literature review

Our study contributes to a growing body of studies on stock price crash risk. Crash risk represents the negative skewness of stock return distributions, which is found to be more common compared to the positive skewness (Chen *et al.*, 2001). To date, the literature related to crash risk has placed emphasis on the explanation of crash risk. Hong and Stein (2003) propose a model based on the assumption of investor heterogeneity to explain the stock market crashes. It argues that the “bullish” and ‘bearish” investors hold different opinions on the stock, and the hidden information held by the bearish investors does not account for the stock pricing, due to the constraints of short-sale. Crashes are triggered when the accumulated information comes out, having reached a tipping point. Jin and Myers (2006) further explain crash risk from the perspective of agency theory. They suggest that managers tend to conceal bad news because of their personal incentives, which are in direct conflict with the interests of investors. Similarly, market crashes occur when the bad news withheld by managers swells up to a point and is then released. Thus, reporting transparency is crucial to a firm’s crash risk because it reduces information asymmetry (Hutton *et al.*, 2009). The following studies have explored other determinants that affect corporate disclosure, such as the shareholdings of institutional investors (Callen and Fang, 2013) and the disclosure of corporate social responsibility (Kim *et al.*, 2014). Regarding the impact of soft information on a firm’s crash risk, Fu *et al.* (2021) find that even the disclosure tone of a conference call is negatively associated with a stock price crash risk, suggesting that the predictive power of textual sentiment is indeed reflected in corporate disclosures. To date, previous studies in this area have focused on exploring the unknown determinants of crash risk, and the research on crash risk prediction is absent from the field.

This study is also related to the literature on financial market movement prediction, which is an undoubtedly challenging topic in the field of finance and operational research. In general, there are two strands of literature related to this topic. One strand is focused on the power of predictive algorithms to improve forecast accuracy. Econometric models, including multivariate regression and time series analysis techniques, used to be the primary tools used to forecast the movements of stock price (Bezerra and Albuquerque, 2017; Zhang *et al.*, 2017) and crashes (Lleo and Ziemba, 2015). With the application of machine learning techniques, researchers and practitioners are now able to capture the non-stationary pattern of the financial market's movement. The most widely applied techniques used to predict the stock market movement direction and stock price return are NN (Kim and Han, 2000; Cao *et al.*, 2005; Atsalakis and Valavanis, 2009) and support vector machines (SVMs) (Kim, 2003; Yu *et al.*, 2009; Huang and Tsai, 2009). Other techniques applied in the literature include fuzzy logic (Wang, 2002, 2003; Chen *et al.*, 2014), K-nearest neighbors (Ballings *et al.*, 2015; Chen *et al.*, 2017) and RF (Kumar and Thenmozhi, 2014; Patel *et al.*, 2015). In recent years, the development of deep learning algorithms such as convolutional neural networks (CNN) and recurrent neural networks (RNN) have been proven to be an attractive method and have been trending recently because of theirs successful applications in computer vision and natural language learning and because they are recognized to be more powerful in handling the nonlinear patterns of big datasets. Kim and Won (2018) propose a hybrid long-short-term memory (LSTM) model (Hochreiter and Schmidhuber, 1997) which combines the LSTM with various GARCH (Generalized Autoregressive Conditional Heteroskedasticity) models to predict stock index volatility and find the proposed hybrid LSTM model outperforms other forecasting techniques. Chatzis *et al.* (2018) compare a range of techniques to forecast the global stock market crisis and show that the application of deep learning algorithms significantly improves the predictive accuracy in the building of early warning systems. In addition, deep learning algorithms have been applied to stock market index forecasting, which yields significant advantages, helping improve predictive accuracy (Baek and Kim, 2018; Cao *et al.*, 2019). A detailed review of the application of machine learning techniques in financial market prediction can be referred to in Henrique *et al.* (2019).

Another strand of studies on financial market prediction explores the application of soft information such as newspaper and media, search engines, social media and corporate disclosures to improve model predictive performance, of which the textual sentiment of news articles has been the main interest of the related literature (Schumaker *et al.*, 2012; Hagenau *et al.*, 2013; Nassirtoussi *et al.*, 2014). The explosion of social media has attracted more attention compared to more conventional media, which is expected to have a large impact on financial markets, due to the convenience of Internet accessibility. Yu *et al.* (2013) compare the impact of conventional media and social media on stock equity value and the empirical findings suggest that social media has a stronger impact on a firm's short-term performance compared to conventional media. A variety of techniques such as filter approach (Al Nasseri *et al.*, 2015; Oliveira *et al.*, 2017), SVM (Nguyen *et al.*, 2015) and CNN (Jin *et al.*, 2020) are applied to extract and compile the textual information as informative indicators to predict market dynamics. Klußmann *et al.* (2019) further examine the link between the signals derived from Twitter's textual data and financial market dynamics. They show that the platform's expert users are the main drivers behind such links and that the experts' sentiments are more predictive for negative market returns, which allows investors to achieve better risk-adjusted returns. Moreover, the combination of various sources of information from Twitter, message board, news and blogs has been shown to be more effective in stock market prediction (Weng *et al.*, 2017; Audrino *et al.*, 2020).

Our study also contributes to the analysis of textual information in MD&A section. Early studies have documented the evidence that MD&A section is informative of firm's future performance (Bryan, 1997; Clarkson *et al.*, 1999; Xue *et al.*, 2010; Brown and Tucker, 2011; Davis and Tama, 2012). Textual tone is the most widely used feature to capture the embedded information in textual context of corporate disclosure and media reports. However, the application of using textual tone from corporate disclosure to crash risk prediction is relatively limited. Recent studies have explored the link between textual tone and bankruptcy risk (Mayew *et al.*, 2015) firm inventory increase (Sun, 2010) and debt market performance (Wang, 2021). In the context of China market, Meng *et al.* (2017) apply the vector normalization method to measure the information value of MD&A section and find it is negatively related with stock price crash risk, but it does not discuss the application of textual tone. Du *et al.* (2022) propose a bespoke financial sentiment dictionary to reflect the linguistic features of Chinese Mandarin. Huang *et al.* (2020) provide a systematic review of the application of textual analysis in China's financial market study.

In summary, we find that the existing studies place emphasis on exploring the new determinants of stock price crash risk, but limited efforts have been made in crash risk prediction itself. As the unstructured textual data has been successfully applied for stock market movement prediction, our study aims to utilize the textual sentiment of annual reports and machine learning techniques to improve the prediction of stock price crash risk.

3. Data and sample

3.1 Textual sentiment

The textual sentiment contained within management statements reflects the opinions and attitude of the managers regarding the firm's performance and outlook. To extract the textual sentiment from MD&A sections, we collect the annual reports of listed firms, which are saved as "pdf" files. The MD&A sections are loaded into Python and tokenized into individual sentences. The data sample is cleansed by removing the headings, titles and bullet numbers. In total, there are more 600,000 sentences in the textual dataset.

Next, we randomly select 10% from the total sample of sentences for labeling and model training. To be specific, each sentence is flagged with a sentimental label such as 1, 0 and −1, representing the sentiment of positive, neutral and negative. In this project, three assistants are allocated to label the sentiment of each sentence individually, and the final label is determined by the majority voting. Should there is any ambiguity, the authors would make the final decisions. The labeled sample is further divided into a training and a testing set, based on the ratio of 80/20. Table 1 shows the sample distribution by sentimental categories. It finds that positive and neutral sentences dominate the sample, while the proportion of negative sentences accounts for around 10% of the learning sample. This is consistent with the expectation that managers tend to send more positive signals in the annual reports rather than express negative sentiment. To transform the textual information into numeric vectors, the word embedding technique Word2Vec (Mikolov *et al.*, 2013) is applied to predict the

Category	Train	Test	Total	Percentage
Positive	14,771	3,693	18,464	43%
Neutral	16,076	4,020	20,096	47%
Negative	3,603	901	4,504	10%
Total	34,450	8,614	43,064	100%

Source(s): Authors' own work

Table 1.
Description of textual sample

meaning of words using their neighboring words to reduce the dimension of input features. Word2Vec is essentially formulated as a shallow neural network model, and the trained network can be applied to predict the probabilities of each word based on their neighboring words. We apply the pre-trained model based on a dictionary of the Chinese language to transform each sentence into a $v \times d$ matrix for model training, where v denotes the size of sentence and d represents the dimension of the feature space. In this way, a one-hot coded sentence can be projected into a lower d-dimensional space by the weight matrix of the trained network for predictive learning.

Last, we adopt the Naïve Bayes method as the classification technique for the training sample and validate on the testing sample, following Li (2010). Naïve Bayes is a conditional independent probabilistic model which assumes that the probability of each input feature conditioning on the class label is independent. According to the Bayes Theorem, the posterior probability that each observation belongs to each class C_k is given as

$$P(C_k|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) = \frac{P(C_k) \cdot P(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n|C_k)}{P(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n).}. \tag{1}$$

As the term $P(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ is fixed, Equation (1) can be reformulated as

$$P(C_k|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) \propto P(C_k)P(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n|C_k) = P(C_k)\prod_{m=1}^{M} P(x_m|C_k). \tag{2}$$

In this study, we generate the predicted sentiment score using the weighted predicted probabilities by sentiment flag, such as

$$s = \sum_{K}^{k=1} C_k P(C_k) \tag{3}$$

The sentiment score of the annual report is then obtained as the averaged predicted value of all sentences in the corresponding MD&A section. Following the above steps, we apply the trained model on the whole set to generate the annual report sentiment and then match it with the accounting and market variables for modeling inputs, based on their unique identifier.

3.2 Stock price crash risk

Following previous studies (Chen *et al.*, 2001; Jin and Myers, 2006; Hutton *et al.*, 2009), we employ two proxies to measure the firm-level stock price crash risk. Firstly, we estimate firm-specific weekly returns by removing the impact of market returns as below:

$$r_{i,w} = \alpha_i + \beta_1 r_{m,w-2} + \beta_2 r_{m,w-1} + \beta_3 r_{m,w} + \beta_4 r_{m,w+1} + \beta_5 r_{m,w+2} + e_{i,w} \tag{4}$$

where $r_{i,w}$ is the return on stock i in week w and $r_{m,w}$ is the return on the value-weighted market index in week w. To account for nonsynchronous trading, the lead and lag terms of the market return $(r_{m,w-1}, r_{m,w-2}, r_{m,w+1}, r_{m,w+2})$ are also included.

We use the residual return $e_{i,w}$ in Equation (4) to construct the firm-specific weekly return $W_{i,w}$ such as $\log(1 + e_{i,w})$. The first measure of firm-specific crash risk is the *NCSKEW* of future returns, which is the negative of the third moment of each stock's firm-specific weekly returns divided by the cubed standard deviation:

$$NCSKEW_{i,t} = -\left[n(n-1)^{3/2}\sum W_{i,w}^3\right] \Big/ \left[(n-1)(n-2)\left(\sum W_{i,w}^2\right)^{3/2}\right] \tag{5}$$

where n is the number of firm-level weekly returns of firm i in a fiscal year t and $NCSKEW_{i,t}$ is a measure of the crash risk in year t. A larger value of $NCSKEW$ indicates more negatively skewed weekly returns and a greater level of crash risk.

The second proxy of stock price crash risk is the DUVOL of firm-specific weekly returns, *DUVOL*, which is defined as follows:

$$DUVOL_{i,t} = \log\left\{\left[(n_{Up}-1)\sum\nolimits_{Up} W_{i,w}^2\right] \Big/ \left[(n_{Down}-1)\sum\nolimits_{Down} W_{i,w}^2\right]\right\} \tag{6}$$

Where n_{Up} and n_{Down} are the number of weeks during the firm-level returns and are above and below the stock's annual average return over the fiscal year t, respectively. We define weeks with firm-specific weekly returns above (below) and the mean of the period as the "up" ("down") sample. *DUVOL* is the log ratio of the standard deviation of the "down" sample to the standard deviation of the "up" sample. A larger value of *DUVOL* indicates a higher left-skewed weekly returns and a greater level of crash risk.

3.3 Sample and variables

We source our sample from the publicly listed firms on the Shanghai and Shenzhen stock exchanges. The latest revised version of the guidelines on annual report information disclosure published by the China Securities Regulatory Commission (CSRC) was enforced from 2012 [3]. To ensure the coherence and completeness of sample data, we select the sample from 2012 to 2017. Following previous studies (Hutton *et al.*, 2009; Kim and Zhang, 2016; Callen and Fang, 2017), we select a collection of market and accounting variables as the modeling input variables. The accounting variables include the firm size (*SIZE*), measured by the log of market value of equity; book-to-market ratio (*BTM*), measured by the ratio of book value to market value; leverage ratio (*LEV*), measured by the book value of all liabilities divided by total assets; return on assets (*ROA*), defined as operating earnings divided by total assets; firm-level accrual manipulation (*ACCRM*), measured by the moving sum of the absolute value of annual performance-adjusted discretionary accruals for the past three years. We then merge the market variables, including the average weekly returns (*RET*), measured as the average of firm-specific weekly return in fiscal year multiplied by 100; weekly return volatility (*SIGMA*), calculated as standard deviation of firm-specific daily returns and detrended share turnover (*DTURN*), measured as the average monthly share turnover over the fiscal year minus the average monthly share turnover over the previous year. Lastly, the textual tone extracted from the MD&A sections and the proxies of crash risk are merged in the sample. All the stock market information and accounting variables are sourced from the China Stock Market & Account Research Database (CSMAR). To ensure the predictive purpose of the model, all input variables are lagged one year ahead of the dependent variable. To account for the auto-correlation of crash risk, the one year lagged crash risk proxy is also included as the input factor. To be specific, for an observation recorded at year t, the corresponding crash risk proxies are measured at year t+1, and the predictive model can be constructed as below:

$$Crash\,Risk_{i,t+1} = f\left(Crash\,Risk_{i,t}, Input\,Features_{i,t}\right.$$

where $f(\cdot)$ is the unknown function mapping the input variables with the target variable. The observations with missing values are removed, leaving us with a total of 7,267 instances in the final sample. Table 2 presents the distributions of crash risk by year, and Figure 1 plots the trend. Since the dependent variable is mapped with the observations one year lagged, the distribution of crash risk ranges from 2013 to 2018 accordingly. It shows that the general trends of NCSKEW and DUVOL are highly correlated with each other. Crash risk remains relatively high from 2013 and peaks in 2015 and 2016 and then decreases to a lower level between 2017 and 2018. The nationwide stock market disaster began from June 2015 and reached to another high level in the beginning of 2016 due to the introduction of the circuit breaker system, which further amplified the panic of market investors. This is consistent with the statistics reported in Table 2. Table 3 reports the summarized statistics of the variables in the forecasting models. The standard deviations of *NCSKEW* and *DUVOL* are 0.6843 and 0.4873 respectively, indicating that the variations of crash risk are quite significant over the years. The mean of the overall scores of MD&A (*MDA_SCORE*) is 0.2051, suggesting that management teams hold an optimistic view regarding their firm's performance, as the sentiment of management statements is positive on average. For the input variables, *BTM*, *LEV* and *ACCRM* present extremely large values compared to their mean and median values. These variables are winsorized at the 1 and 99% percentiles to remove the outliers.

4. Experiment design

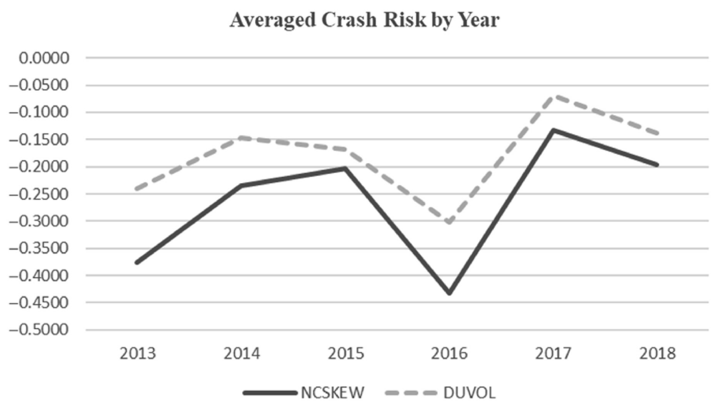
4.1 Predictive models

4.1.1 Linear regression. In this section we briefly introduce the predictive models used to forecast crash risk. We apply five predictive techniques in the experiment, including ordinary LR, NN, LS-SVR, RF and GBDT. LR is the most commonly used statistical regression model in predictive tasks. The model is formulated to predict the target variable y_i using the linear combination of the input variables such as

Year	N	*NCSKEW*	*DUVOL*
2013	968	−0.3759	−0.2394
2014	1,130	−0.2351	−0.1465
2015	1,203	−0.2036	−0.1674
2016	1,213	−0.4327	−0.3013
2017	1,368	−0.1321	−0.0695
2018	1,385	−0.1956	−0.1379

Source(s): Authors' own work

Table 2. Distributions of crash risk by year



Source(s): Authors' own work

Figure 1. Trend of stock price crash risk

Variables	N	Mean	Median	Std	Min	Max
NCSKEW	7,267	−0.2547	−0.2159	0.6843	−3.6860	3.5762
DUVOL	7,267	−0.1721	−0.1755	0.4873	−2.0754	4.0835
MDA_SCORE	7,267	0.2051	0.2049	0.1845	−0.5393	1.0000
SIZE	7,267	22.2042	22.0743	1.2824	15.5773	27.4673
BTM	7,267	0.9551	0.6288	1.0187	0.0029	12.1002
LEV	7,267	0.4644	0.4484	0.7889	0.0140	63.9712
ROA	7,267	0.0105	0.0068	0.0413	−1.9877	1.2068
ACCRM	7,267	0.2475	0.1433	2.1012	0.0000	122.4756
RET	7,267	−0.1238	−0.0913	0.1089	−1.3919	0.0000
SIGMA	7,267	0.0465	0.0431	0.0187	0.0000	0.1686
DTURN	7,267	−0.0429	−0.0356	0.3554	−2.8278	2.2349

Source(s): Authors' own work

Table 3.
Summarized statistics

$$y_i = \boldsymbol{w}^T \boldsymbol{x}_i + b + u_i, i = 1, \ldots, N \tag{7}$$

where $\boldsymbol{x}_i$ denotes the input vector of the i-th observation and u_i is the corresponding error term. N is the total number of observations in the training sample. The coefficients of LR can be estimated by the least squares method to derive the estimates of $\boldsymbol{w}$ and b and the prediction of y_i is obtained accordingly.

4.1.2 Neural networks. The NN algorithm is designed to mimic the mechanism of the human brain in the learning process. In this study we adopt a three-layer shallow network architecture, which is constructed by an input layer, a hidden layer and an output layer, with neurons connected by weights. Denote $\boldsymbol{\alpha}_j$ as the connectivity weights between the input and hidden layers and M as the number of hidden neurons. Now the output value of the hidden layer neuron z_j can be represented by the weighted linear combination of input variables, such as

$$z_j = \sigma\left(\boldsymbol{\alpha}_j^T \boldsymbol{x}_i + \alpha_0\right), j = 1, \ldots, M \tag{8}$$

Here $\sigma(\cdot)$ is the activation function, which can be chosen as the sigmoid or tanh function. The final output value is given as the aggregation of the hidden layer outputs, such that

$$y_i = \delta\left(\boldsymbol{\beta}_j^T z_j + \beta_0\right), i = 1, \ldots, N \tag{9}$$

where $\boldsymbol{\beta}_j$ is the weights connecting the hidden and output layers and $\delta(\cdot)$ is another activation function that transforms the linear combination of the outputs of the hidden layer to the final output. A common choice of $\delta(\cdot)$ is the identity function, such that $y_i = \boldsymbol{\beta}_j^T z_j + \beta_0$. The network weights $\boldsymbol{\alpha}_j$ and $\boldsymbol{\beta}_j$ can be estimated by the back-propagation learning algorithm, which minimizes the error functions and updates the weights backward, from the hidden layer to the input layer. The network architecture can be generalized easily, with more hidden layers added to accommodate higher data complexity.

4.1.3 Least square support vector regression. SVMs can separate samples in high-dimensional space by maximizing the margin of the parallel separating hyperplanes. SVM techniques are featured in the application of kernel functions, which map data instances in low-dimensional space into a high-dimensional space in order that they be linearly separable. A variety of SVM models have been proposed to solve the classification and regression models. This study applies the LS-SVR proposed by Suykens and Vandewalle (1999) for crash risk prediction. Following the notations used in Equation (7), the LS-SVR model is formulated in the following form:

$$\min J(\boldsymbol{w}, b; u_i) = \frac{1}{2}\|\boldsymbol{w}\|^2 + \frac{C}{2}\sum_{i=1}^{N} u_i^2$$
$$s.t.\, y_i = \boldsymbol{w}^T\varphi(\boldsymbol{x}_i) + b + u_i, i = 1, \ldots, N \tag{10}$$

where C denotes the regularized parameter to control the trade-off between error terms and margin maximization. Equation (10) can be solved by deriving the Lagrangian function, such as

$$L(\boldsymbol{w}, b, u_i; \alpha_i) = J(\boldsymbol{w}, u_i) - \sum_{i=1}^{N} \alpha_i\left(\boldsymbol{w}^T\varphi(\boldsymbol{x}_i) + b + u_i - y_i\right) \tag{11}$$

where α_i is the Lagrangian multiplier. According to the KKT (Karush-Kuhn-Tucker) conditions, the solution of the dual form is equivalent to solving the following linear equation systems

$$\begin{pmatrix} 0 & \boldsymbol{e}^T \\ \boldsymbol{e} & \overline{\mathbf{K}} \end{pmatrix}\begin{pmatrix} b \\ \boldsymbol{\alpha} \end{pmatrix} = \begin{pmatrix} 0 \\ y \end{pmatrix} \tag{12}$$

Where $\boldsymbol{e} = (1, \ldots, 1)^T_{1\times N}, \boldsymbol{y} = (y_1, \ldots, y_N)^T, \boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_N)^T, \overline{\boldsymbol{K}} = \boldsymbol{K} + (1/C)\boldsymbol{I}$. Here $\boldsymbol{K}$ is the kernel matrix and $\boldsymbol{I}$ is the identity matrix. The closed form solution is obtained as

$$\left\{ \boldsymbol{\alpha}^* = \overline{\boldsymbol{K}}^{-1}(\boldsymbol{y} - b^*\boldsymbol{e})b^* = \boldsymbol{e}^T\overline{\boldsymbol{K}}^{-1}\boldsymbol{y}\Big/\boldsymbol{e}^T\overline{\boldsymbol{K}}^{-1}\boldsymbol{e} \right. \tag{13}$$

Now the final predicted output is given as

$$\widehat{g}(\boldsymbol{x}) = \sum_i \alpha_i^* \boldsymbol{K}(\boldsymbol{x}_i, \boldsymbol{x}) + b^* \tag{14}$$

4.1.4 Random forest. RF is essentially an ensemble learning algorithm which aggregates the prediction of a group of individual decision trees based on the bagging method (Breiman, 2001). The training process of RF is not finished until all the decision trees are established. During each iteration, a bootstrap sample of N_{sub} observations is drawn from the training sample with N instances. The sample is trained based on a random subset with k variables from the total of K input variables. Each decision tree grows until all the nodes are split as leaves. The final predicted output is given by taking the average voting of all individual decision trees.

4.1.5 Gradient boosting decision tree. Gradient boosting is another type of ensemble learning algorithm which aggregates the prediction of individual base learner based on the boosting method, such that

$$f_J(\boldsymbol{x}) = \sum_{j=1}^{J} T(\boldsymbol{x}; \theta) \tag{15}$$

where $T(\boldsymbol{x}; \theta)$ is the j-th base learner and θ represents the corresponding parameter. In this study the base learner is chosen to be the decision tree, and the model is formulated as a GBDT. The idea of GBDT is to optimize the decision trees to minimize the loss function

(Friedman, 2001). To achieve such a target, each decision tree is trained and updated based on the gradient descent direction of the incumbent loss function, denoted by $L(y, f(\boldsymbol{x}); \gamma)$, and the negative gradient descent is given as

$$q_j = -\left[\frac{\partial L(y, f(\boldsymbol{x}))}{\partial f(\boldsymbol{x})}\right] \Big/ f(\boldsymbol{x}) = f_{j-1}(\boldsymbol{x}) \tag{16}$$

and the updated learner is given as

$$f_j(\boldsymbol{x}) = f_{j-1}(\boldsymbol{x}) + \gamma_j q_j \tag{17}$$

GBDT has been shown to be a powerful predictive technique for classification and regression problems. Moreover, it is capable of generating feature importance for variable selection. Thus, it has been chosen as another predictive model in our experimental analysis.

4.2 Performance metrics

To explore the predictive power of textual sentiment in crash risk forecasting, we conduct out-of-time and out-of-sample analyses in the experiment. For the out-of-time prediction, we divide the whole sample into a training set and a test set by year. To be specific, the observations from 2012 to 2016 are allocated into the training set, and the remaining observations in 2017 fall into the testing set. To account for the sampling randomness in model prediction, we also perform out-of-sample analyses by adopting the $N_C \times 2$ cross-validation strategy, following Dietterich (1998). First, the whole sample is randomly split into training and test sets by half and half, and then the training and test sets are swapped to once again generate the out-of-sample prediction. When the above process is repeated by N_C times, a total of $N_C \times 2$ out-of-sample predictive outputs are obtained. Here we select N_C to be 5 in the following analysis. The hyperparameters of machine learning techniques including NN, LS-SVR, RF and GBDT are tuned by ten-fold cross-validation in the training set. To examine if the model performance is improved with the textual sentiment, we apply a paired *t*-test to compare the outputs of all predictive models. Performance metrics including MAE and RMSE are applied for model evaluation presented in Equation (18)

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_i \left(y_i - \widehat{y}_i\right)^2}$$

$$\mathrm{MAE} = \frac{1}{n}\sum_i \left|y_i - \widehat{y}_i\right| \tag{18}$$

5. Results and discussions

This section reports the experimental results and discussions of both out-of-time and out-of-sample predictions, with the crash risk proxied by *NCSKEW* and *DUVOL* respectively. The predictive models applied in the experiment are introduced in Section 4.1.

5.1 NCSKEW

Table 4 shows the results of out-of-time prediction on the testing set for both baseline and treatment models. It finds that the modeling performance, including MAE and RMSE, has been improved when the textual sentiment is included in the model inputs, suggesting that the MD&A sentiment is indeed predictive of stock price crash risk. Specifically, the performance improvement of NN is more substantial than the other techniques. The MAE and RMSE are improved from 0.5642 to 0.7554 to 0.5422 and 0.7318 respectively and also outperforms the other learning algorithms. GBDT presents the second best predictive performance in terms of the MAE and RMSE, followed by LS-SVR and RF. LR is shown to be the least competitive forecasting technique, and the improvement of applying MD&A sentiment is marginal, indicating that machine learning techniques are more advantageous in the prediction of stock price crash risk.

Table 5 presents the outputs of out-of-sample predictions, based on the $N_C \times 2$ cross-validation analysis, where N_C is specified as 5. Therefore, we report the mean and standard deviation of each performance metric for the ten predictive outputs on the testing set. Panel A in Table 5 shows the results of MAE, which finds that the MD&A sentiment provides a slight improvement in terms of MAE. According to Panel B in Table 5, the improvement of modeling performance is more noticeable in terms of RMSE. Specifically, NN is shown to be the most competitive technique in the out-of-sample prediction. When the MD&A sentiment is used in the model inputs, MAE and RMSE are improved, to 0.4865 and 0.6320, according to both Panel A and Panel B in Table 5. The MAE of GBDT and LS-SVR is improved to be less than 0.5, with the inclusion of textual sentiment, based on Panel A in Table 5, and the corresponding RMSE is also improved accordingly. The improvement of MAE is marginal for RF, which demonstrates very little advantage to LR. In contrast, the RMSE of RF is significantly better than LR, although it is still outperformed by the other machine learning algorithms. It should be noted that NN suffers from the largest standard deviation for either performance metric, which suggests that the stability of NN is relatively lower compared to the other algorithms and that the inclusion of textual sentiment cannot improve the variance of performance metrics. Figure 2 presents the comparisons of modeling outputs for out-of-time and out-of-sample predictions based on the negative skewness of distribution. It shows that the improvement of treatment models in out-of-sample prediction is remarkable for all predictive techniques, but the improvement in the out-of-time prediction is relatively marginal for LR.

Table 6 reports the p values of the paired t-test, to investigate if the improvement of modeling performance is statistically significant with the incorporation of management

	Baseline		Treatment	
Model	MAE	RMSE	MAE	RMSE
LR	0.6023	0.8010	0.6014	0.7993
NN	0.5642	0.7554	0.5422	0.7318
LS-SVR	0.5788	0.7731	0.5645	0.7694
RF	0.5724	0.7798	0.5648	0.7639
GBDT	0.5647	0.7624	0.5540	0.7522

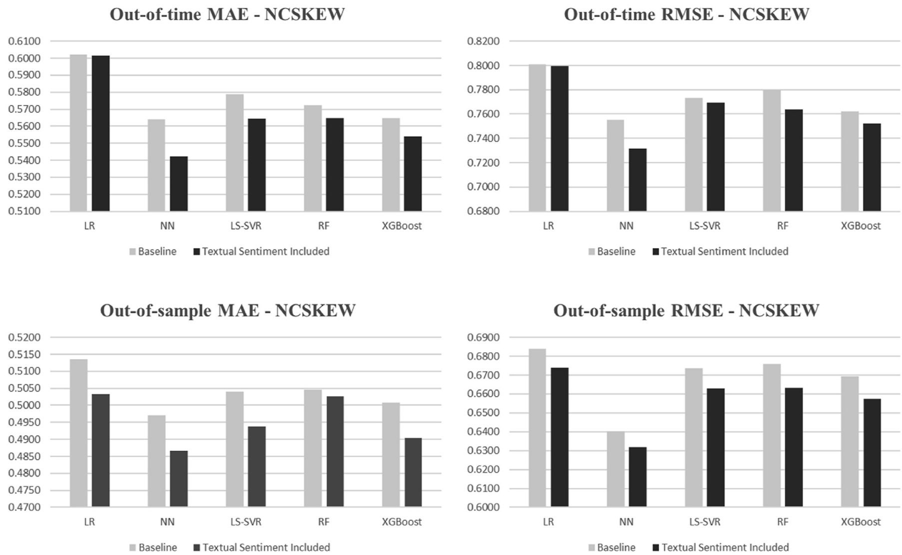
Source(s): Authors' own work

Table 4. Out-of-time prediction - NCSKEW

	Panel A. MAE				Panel B. RMSE			
	Baseline		Treatment		Baseline		Treatment	
Model	Mean	Std	Mean	Std	Mean	Std	Mean	Std
LR	0.5135	0.0035	0.5034	0.0036	0.6840	0.0067	0.6739	0.0067
NN	0.4971	0.0103	0.4865	0.0176	0.6404	0.0210	0.6320	0.0139
LS-SVR	0.5041	0.0029	0.4937	0.0030	0.6736	0.0061	0.6629	0.0065
RF	0.5045	0.0043	0.5027	0.0049	0.6759	0.0074	0.6631	0.0086
GBDT	0.5008	0.0037	0.4904	0.0033	0.6694	0.0068	0.6575	0.0066

Source(s): Authors' own work

Table 5. Out-of-sample predictions – NCSKEW



Source(s): Authors' own work

Figure 2. Comparisons of predictive performance – NCSKEW

Table 6. Paired *t*-test - NCSKEW

Prediction	MAE	RMSE
Out of time	0.0336 **	0.0526 *
Out of sample	0.0071 ***	0.0001 ***

Source(s): Authors' own work

sentiment. We calculate the difference between the treatment and the baseline for each model and test if the mean difference is significant. For the out-of-time prediction, the improvement of modeling performance is significant at the 5 and 10% levels for MAE and RMSE, with the p values reported as 0.0336 and 0.0526 respectively. Moreover, the improvement of out-of-sample prediction is significant at the 1% level for both MAE and RMSE. The t-test results indicate that textual sentiment is able to improve the crash risk proxied by *NCSKEW* significantly for the predictive algorithms used in our study.

5.2 DUVOL

This section continues with a discussion of the experimental results with the crash risk proxied by *DUVOL*. The modeling outputs of out-of-time prediction are presented in Table 7. It shows that NN demonstrates a remarkable advantage over the other predictive techniques, with the MAE and RMSE reported as 0.3977 and 0.5341 for the baseline model, which are improved to become 0.3788 and 0.5260, with the MD&A sentiment added to the treatment model. The improvement is less substantial for GBDT, which is close to RF with the inclusion of textual sentiment. LS-SVR and LR rank at the bottom among all the predictive algorithms. It may be that the improvement of LR is too trivial to be noticed, which is consistent with the evidence reported in Table 4.

Similar to the above, Table 8 exhibits the outputs of out-of-sample prediction with crash risk proxied by *DUVOL*. According to Panel A of Table 8, there is no significant difference in terms of MAE for machine learning techniques, although the advantage of NN is more noticeable when the MD&A sentiment is not included. We also find that the improvement of treatment models is less substantial compared to the results reported in Table 5. Based on Panel A and Panel B in Table 8, the MAE of NN is reduced from 0.3694 to 0.3617 and RMSE is improved from 0.4668 to 0.4602. In summary, the reduction of MAE and RMSE for NN is less than 0.01. For the other forecasting techniques, the improvement is also relatively marginal. The performance of RF and GBDT are comparable, followed by LS-SVR and LR. Moreover, NN still demonstrates the largest standard deviation of MAE and RMSE, which is consistent with Table 5.

Following our previous analysis, we compare the predictive performance between the baseline and treatment models based on the paired *t*-test and report the *p*-values in Table 9. It shows that the improvement in out-of-time prediction is significant at the 5% level for MAE and RMSE. Although the magnitude of improvement in Table 7 is not remarkable, it still confirms the effectiveness of using MD&A sentiment in crash risk prediction. Regarding the modeling outputs of out-of-sample prediction, the improvement made by the treatment models remains significant at the 5% level. The evidence in Table 9 is consistent with Table 6, suggesting that the predictive models with the inclusion of MD&A sentiment is more advantageous in crash risk prediction. Figure 3 summarizes the modeling comparisons based on the DUVOL. It finds that the improvement based on *DUVOL* is more noticeable compared to the modeling comparisons based on *NCSKEW* presented in Figure 2. In general, the improvement of LR is marginal and is in fact outperformed by the other machine learning techniques.

5.3 Robustness tests

In this section, we examine the robustness of predictive power of textual sentiment using alternative measure of crash risk. We construct a flag variable *CRASH* to indicate if a firm experiences a crash in a given year following Hutton *et al.* (2009). *CRASH* is defined as one if

	Baseline		Treatment	
Model	MAE	RMSE	MAE	RMSE
LR	0.4302	0.5565	0.4301	0.5546
NN	0.3977	0.5341	0.3778	0.5260
LS-SVR	0.4124	0.5364	0.4071	0.5293
RF	0.4152	0.5424	0.3964	0.5364
GBDT	0.4073	0.5343	0.3978	0.5325

Source(s): Authors' own work

Table 7. Out-of-time predictions - DUVOL

	Panel A. MAE				Panel B. RMSE			
	Baseline		Treatment		Baseline		Treatment	
Model	Mean	Std	Mean	Std	Mean	Std	Mean	Std
LR	0.3754	0.0016	0.3750	0.0016	0.4788	0.0045	0.4781	0.0025
NN	0.3694	0.0154	0.3617	0.0080	0.4668	0.0104	0.4602	0.0110
LS-SVR	0.3768	0.0014	0.3651	0.0029	0.4802	0.0042	0.4780	0.0038
RF	0.3748	0.0031	0.3652	0.0028	0.4790	0.0059	0.4679	0.0051
GBDT	0.3736	0.0018	0.3682	0.0029	0.4775	0.0040	0.4650	0.0040

Source(s): Authors' own work

Table 8. Out-of-sample prediction - DUVOL

the firm experiences at least one crash week during a given year and zero otherwise. It is considered to be a crash week If the firm-specific weekly return falls 3.09 or more standard deviations below the mean weekly returns over year.

We first test the link between textual sentiment and crash risk using a panel regression model with year and industry fixed effects as well as other hard information controlled. The regression outputs are presented in Table 10 and columns (1) to (3) exhibit the results with each measure of crash risk specified as the dependent variable. For *NCSKEW* and *DUVOL*, the lagged term is also included as the control variable respectively, which is omitted for *CRASH*. The regression model is specified in the logit form when regressing on *CRASH*. It finds that textual sentiment is negatively associated with *NCSKEW* and *DUVOL* at the 5% significance level, which remains significant at the 10% level for *CRASH*. The lagged terms of *NCSKEW* and *DUVOL* are correlated with the dependent variables at the 1% level, indicating strong auto-correlation effect of crash risk. It is also noticed the market driven variables such as *SIGMA* and *DTURN* are both strongly related with crash risk, although *RET* is not predictive for *NCSKEW* and *DUVOL* as expected.

Next, we investigate if the inclusion of textual sentiment improves the prediction of *CRASH* by repeating the above experiment. Since *CRASH* is a binary indicator, we report the classification results in terms of AUC (Area Under the ROC Curve) in Table 11. It shows that the inclusion of textual sentiment improves the modeling outputs for both out-of-time and out-of-sample tests. Moreover, machine learning techniques are evidently more advantageous than logistic regression, and NN still presents better performance compared

Table 9.
Paired *t*-test - DUVOL

Prediction	MAE	RMSE
Out-of-time	0.0491 **	0.0195 **
Out of sample	0.0230 **	0.0463 **

Source(s): Authors' own work

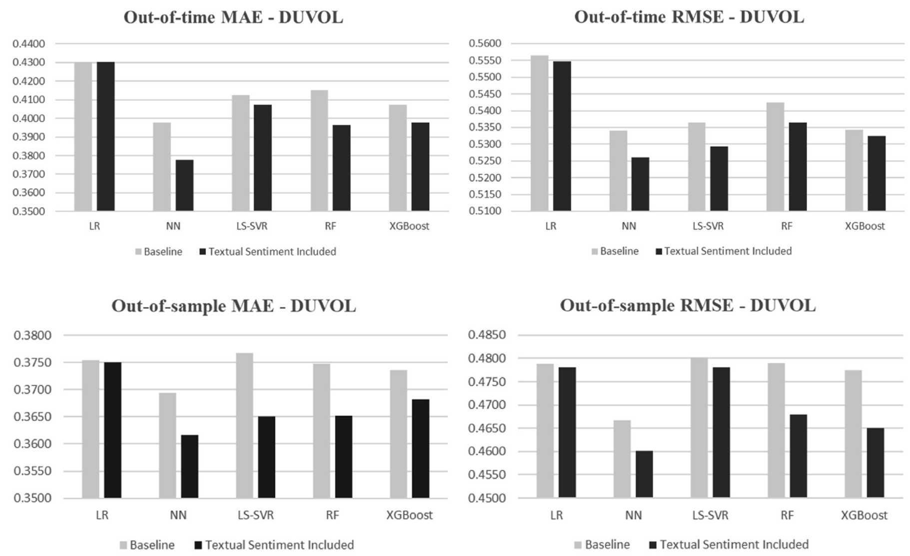


Figure 3.
Comparisons of predictive performance – DUVOL

Source(s): Authors' own work

	(1) *NCSKEW*	(2) *DUVOL*	(3) *CRASH*
MDA_SCORE	−0.078**	−0.098**	−0.005*
	(−1.998)	(−2.071)	(−1.792)
Lagged	0.110***	0.181***	–
	(11.410)	(13.470)	–
SIZE	−0.031	−0.038**	−0.830***
	(−1.590)	(−2.240)	(−3.660)
BTM	−0.051***	−0.046***	−0.008
	(−4.390)	(−3.960)	(−1.030)
LEV	−0.002	−0.001	0.001
	(−0.200)	(−0.150)	(0.190)
ROA	−0.310*	−0.356**	−0.066
	(−1.950)	(−2.250)	(−0.670)
ACCRM	0.009***	0.009***	−0.001
	(3.120)	(3.090)	(−0.390)
RET	7.267	9.018	6.211***
	(0.330)	(0.410)	(3.300)
SIGMA	2.953**	2.914**	2.774***
	(2.210)	(2.190)	(3.310)
DTURN	−0.110***	−0.114***	−0.054***
	(−6.710)	(−6.930)	(−5.200)
Constant	0.936	0.091	0.234
	(1.610)	(0.480)	(1.930)
Year FE	*Yes*	*Yes*	*Yes*
Industry FE	*Yes*	*Yes*	*Yes*
N	7,267	7,267	7,267
R-squared	0.3048	0.3109	0.2207

Note(s): The *t*-values are reported in parentheses. ***, ** and * represent statistical significance at the 1%, 5% and 10% levels, respectively
Source(s): Authors' own work

Table 10.
Fixed effect regression

	Out-of-time		Out-of-sample	
Model	Baseline	Treatment	Baseline	Treatment
LR	0.6427	0.6688	0.6574	0.6653
NN	0.7403	0.7628	0.7425	0.7649
LS-SVR	0.7198	0.7246	0.7237	0.7388
RF	0.7326	0.7414	0.7355	0.7567
GBDT	0.7354	0.7503	0.7459	0.7535

Source(s): Authors' own work

Table 11.
Out-of-time and out-of-sample prediction - *CRASH*

to other techniques. Such evidence also validates the effectiveness of using soft information to predict crash risk.

6. Conclusions

Stock price crash risk is important for investors and regulators, and forecasting crash risk is a challenging topic in financial market. The determinants of crash risk have been largely investigated based on the econometric models. However, there is lack of studies that explore the prediction of crash risk. Since the management statements reflect the views of corporate senior managers, it is promising to extract the textual sentiment embedded in such context as complementary information in addition to the hard information presented in the annual reports to predict stock price crash risk. Our study aims to predict stock price crash risk using machine learning techniques with the incorporation of textual sentiment. To be specific, we first extract the textual sentiment from the MD&A sections in annual reports using the Naïve Bayes method. The extracted sentiment score is bounded between -1 and 1 indicating how optimistic or pessimistic of the attitude reflected from the management statement. We then include the textual sentiment in the predictive models to forecast crash risk which is proxied by negative skewness of stock price return distribution and DUVOL. The modeling results demonstrate that the model predictive accuracy in terms of MAE and RMSE is incrementally improved with the inclusion of textual sentiment for both out-of-time and out-of-sample predictions. In addition, we find machine learning techniques present better predictive performance compared to the traditional LR.

The main contribution of this study is proposing to apply the soft information embedded in the annual reports for crash risk prediction, which has not been presented in literature. In fact, prior studies related to stock price crash risk have prominently focused on exploring the explanation of crash risk rather than investigate how to predict crash risk more accurately. Our study shows that the textual sentiment reflected from the management statement is not only related to crash risk, and the predictive performance is significantly improved with the incorporation of textual sentiment. Our findings highlight the information value embedded in the management statements in the China market. Our study also shows implications to the importance of utilizing soft information in financial risk prediction. Moreover, the regulators need to ensure that managers should not exaggerate their statement to mislead the market investors. Highly volatile markets will bring losses to investors and damage the financial stability of the financial system where the evaluation of many assets rely on the securities market. If crash risk can be foreseen and addressed, the management of market risk can be effective and assets can be secured.

It should be noted that our study focuses on examining the predictive power of textual sentiment in crash risk prediction, while other dimensions of textual features such as readability and thematic contents are not considered. More analysis is needed to explore the predictive power of textual features from various dimensions with the most recent sample data included in future study. Beside when computing capabilities improve in recent years, large language models can empower more deep mining on textual information, which is also an interesting topic to explore.

Notes

1. The circuit breaker of USA stock market is referenced to the S&P 500 index. According to the current guidelines, there are three levels of breaks when the index drops by 7%, 13 and 20%, respectively, based on the previous close price. In 2020, the circuit breaker of S&P 500 index was triggered in March 9th, March 12th, March 16th and March 18th, all of which were Level 1 break.
2. Luckin Coffee has been suspended trading on 29th June and filed for delisting from Nasdaq.
3. Guidelines for the Content and Format of Corporate Information Disclosure in Public Securities Issues No. 2 - Content and Format of Annual Report (Revised in 2012), China Securities Regulatory Commission, available at http://www.gov.cn/gongbao/content/2012/content_2292066.htm

References

Al Nasseri, A., Tucker, A. and de Cesare, S. (2015), “Quantifying StockTwits semantic terms’ trading behavior in financial markets: an effective application of decision tree algorithms”, *Expert Systems with Applications*, Vol. 42 No. 23, pp. 9192-9210.

Atsalakis, G.S. and Valavanis, K.P. (2009), “Surveying stock market forecasting techniques - Part II: soft computing methods”, *Expert Systems with Applications*, Vol. 36 No. 3, pp. 5932-5941.

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

Baek, Y. and Kim, H.Y. (2018), “Modaugnet: a new forecasting framework for stock market index value with an overfitting prevention LSTM module and a prediction LSTM module”, *Expert Systems with Applications*, Vol. 113, pp. 457-480.

Ballings, M., den Poel, D.V., Hespeels, N. and Gryp, R. (2015), “Evaluating multiple classifiers for stock price direction prediction”, *Expert Systems with Applications*, Vol. 42 No. 20, pp. 7046-7056.

Bezerra, P.C.S. and Albuquerque, P.H.M. (2017), “Volatility forecasting via SVR-GARCH with mixture of Gaussian kernels”, *Computational Management Science*, Vol. 14 No. 2, pp. 179-196.

Breiman, L. (2001), “Random forests”, *Machine Learning*, Vol. 45, pp. 5-32.

Brown, S.V. and Tucker, J.W. (2011), “Large-sample evidence on firms’ year-over-year MD&A modifications”, *Journal of Accounting Research*, Vol. 49 No. 2, pp. 309-346.

Bryan, S.H. (1997), “Incremental information content of required disclosures contained in management discussion and analysis”, *The Accounting Review*, Vol. 4, pp. 285-301.

Callen, J.L. and Fang, X.H. (2013), “Institutional investor stability and crash risk: monitoring versus short-termism?”, *Journal of Banking and Finance*, Vol. 37, pp. 3047-3063.

Callen, J.L. and Fang, X.H. (2017), “Crash risk and auditor-client relationship”, *Contemporary Accounting Research*, Vol. 34 No. 3, pp. 1715-1750.

Cao, Q., Leggio, K.B. and Schniederjans, M.J. (2005), “A comparison between Fama and French’s model and artificial neural networks in predicting the Chinese stock market”, *Computers and Operations Research*, Vol. 32 No. 10, pp. 2499-2512.

Cao, J., Li, Z. and Li, J. (2019), “Financial time series forecasting model based on ceemdan and LSTM”, *Physica A: Statistical Mechanics and Its Applications*, Vol. 519, pp. 127-139.

Chatzis, S.P., Siakoulis, V., Petropoulos, A., Stavroulakis, E. and Vlachogiannakis, N. (2018), “Forecasting stock market crisis events using deep and statistical machine learning techniques”, *Expert Systems with Applications*, Vol. 112, pp. 353-371.

Chen, J., Hong, H. and Stein, J.C. (2001), “Forecasting crashes: trading volume, past returns, and conditional skewness in stock prices”, *Journal of Financial Economics*, Vol. 61, pp. 345-381.

Chen, Y.-S., Cheng, C.-H. and Tsai, W.-L. (2014), “Modeling fitting-function-based fuzzy time series patterns for evolving stock index forecasting”, *Applied Intelligence*, Vol. 41 No. 2, pp. 327-347.

Chen, H., Xiao, K., Sun, J. and Wu, S. (2017), “A double-layer neural network framework for high-frequency forecasting”, *ACM Transactions on Management Information Systems (TMIS)*, Vol. 7 No. 4, pp. 1-17, 11:17.

Clarkson, P.M., Kao, J.L. and Richardson, G.D. (1999), “Evidence that management discussion and analysis (MD&A) is a part of a firm’s overall disclosure package”, *Contemporary Accounting Research*, Vol. 16 No. 1, pp. 111-134.

Davis, A.K. and Tama, S.I. (2012), “Managers’ use of language across alternative disclosure outlets: earnings press releases versus MD&A”, *Contemporary Accounting Research*, Vol. 29 No. 3, pp. 804-837.

Dietterich, T.G. (1998), “Approximate statistical tests for comparing supervised classification learning”, *Neural Computation*, Vol. 10, pp. 1895-1923.

Du, Z.J., Huang, A.G., Wermers, R.R. and Wu, W.F. (2022), “Language and domain specificity: a Chinese financial sentiment dictionary”, *Review of Finance*, Vol. 26 No. 3, pp. 673-719.

Friedman, J. (2001), “Greedy function approximation: a gradient boosting machine”, *Annals of Statistics*, Vol. 29 No. 5, pp. 1189-1232.

Fu, X., Wu, X. and Zhang, Z. (2021), "The information role of earnings conference call tone: evidence from stock price crash risk", *Journal of Business Ethics*, Vol. 173, pp. 643-660.

Hagenau, M., Liebmann, M. and Neumann, D. (2013), "Automated news reading: stock price prediction based on financial news using context-captureing features", *Decision Support Systems*, Vol. 55, pp. 685-697.

Henrique, B.M., Sobreiro, V.A. and Kimura, H. (2019), "Literature review: machine learning techniques applied to financial market prediction", *Expert Systems with Applications*, Vol. 124, pp. 226-251.

Hochreiter, S. and Schmidhuber, J. (1997), "Long short-term memory", *Neural Computation*, Vol. 9 No. 8, pp. 1735-1780.

Hong, H. and Stein, J.C. (2003), "Differences of opinion, short-sales constraints, and market crashes", *Review of Financial Studies*, Vol. 16 No. 2, pp. 487-525.

Hu, G. and Wang, Y. (2018), "Political connections and stock price crash risk: the role of intermediary information disclosure", *China Finance Review International*, Vol. 8 No. 2, pp. 140-157.

Huang, C.-L. and Tsai, C.-Y. (2009), "A hybrid SOFM-SVR with a filter-based feature selection for stock market forecasting", *Expert Systems with Applications*, Vol. 36 No. 2, pp. 1529-1539.

Huang, A., Wu, W. and Yu, T. (2020), "Textual analysis for China's financial markets: a review and discussion", *China Finance Review International*, Vol. 10 No. 1, pp. 1-15.

Hutton, A.P., Marcus, A.J. and Tehranian, H. (2009), "Opaque financial reports, R^2, and crash risk", *Journal of Financial Economics*, Vol. 94 No. 1, pp. 67-86.

Jiang, C., Wang, Z., Wang, R.,. and Ding, Y. (2017), "Loan default prediction by combining soft information extracted from descriptive text in online peer-to-peer lending", *Annuals of Operations Research*, Vol. 266, pp. 511-529.

Jin, L. and Myers, S. (2006), "R^2 around the world: new theory and new tests", *Journal of Financial Economics*, Vol. 79 No. 2, pp. 257-292.

Jin, Y., Yan, M., Xi, Y. and Liu, C. (2016), "Stock price synchronicity and stock price crash risk: based on the mediating effect of herding behavior of QFII", *China Finance Review International*, Vol. 6 No. 3, pp. 230-244.

Jin, X., Liang, S. and Yu, J. (2021), "Management geographical proximity and stock price crash risk", *China Finance Review International*, Vol. 12 No. 4, pp. 601-622.

Jin, Z., Yang, Y. and Liu, Y. (2020), "Stock closing price prediction based on sentiment analysis and LSTM", *Neural Computing and Applications*, Vol. 32, pp. 9713-9729.

Kim, K. (2003), "Financial time series forecasting using support vector machines", *Neurocomputing*, Vol. 55 Nos 1-2, pp. 307-319.

Kim, K.-j. and Han, I. (2000), "Genetic algorithms approach to feature discretization in artificial neural networks for the prediction of stock price index", *Expert Systems with Applications*, Vol. 19 No. 2, pp. 125-132.

Kim, H. and Won, C. (2018), "Forecasting the volatility of stock price index: a hybrid model integrating LSTM with multiple GARCH-type models", *Expert Systems with Applications*, Vol. 103, pp. 25-37.

Kim, J.-B. and Zhang, L. (2016), "Accounting conservatism and stock price crash risk: firm-level evidence", *Contemporary Accounting Research*, Vol. 33 No. 1, pp. 412-441.

Kim, Y., Li, H. and Li, S. (2014), "Corporate social responsibility and stock price crash risk", *Journal of Banking and Finance*, Vol. 43, pp. 1-13.

Klußmann, A.G., König, S. and Ebner, M. (2019), "Buzzwords build momentum: global financial Twitter sentiment and the aggregate stock market", *Expert Systems with Applications*, Vol. 136, pp. 171-186.

Kumar, M. and Thenmozhi, M. (2014), "Forecasting stock index returns using ARIMA-SVM, ARIMA-ANN, and ARIMA-random forest hybrid models", *International Journal of Banking, Accounting and Finance*, Vol. 5 No. 3, pp. 284-308.

Li, F. (2010), "The information content of forward-looking statements in corporate filings - a naive Bayesian machine learning approach", *Journal of Accounting Research*, Vol. 48, pp. 1049-1102.

Liang, Q., Ling, L., Tang, J., Zeng, H. and Zhuang, M. (2020), "Managerial overconfidence, firm transparency, and stock price crash risk: evidence from an emerging market", *China Finance Review International*, Vol. 10 No. 3, pp. 271-296.

Lleo, S. and Ziemba, W.T. (2015), "Some historical perspectives on the Bond-Stock Earnings Yield Model for crash prediction around the world", *International Journal of Forecasting*, Vol. 31, pp. 399-425.

Loughran, T. and Mcdonald, B. (2011), "When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks", *Journal of Finance*, Vol. 66 No. 1, pp. 35-65.

Maghyereh, A. and Abdoh, H. (2022), "Can news-based economic sentiment predict bubbles in precious metal markets?", *Financial Innovation*, Vol. 8 No. 1, pp. 1-29.

Mai, F., Tian, S., Lee, C. and Ma, L. (2019), "Deep learning models for bankruptcy prediction using textual disclosures", *European Journal of Operational Research*, Vol. 274 No. 2, pp. 743-758.

Matin, R., Hansen, C., Hansen, C. and Mølgaard, P. (2019), "Predicting distresses using deep learning of text segments in annual reports", *Expert Systems with Applications*, Vol. 132, pp. 199-208.

Mayew, W.J., Sethuraman, M. and Venkatachalam, M. (2015), "MD&A disclosure and the firm's ability to continue as a going concern", *The Accounting Review*, Vol. 90 No. 4, pp. 1621-1651.

Meng, Q.B., Yang, J.H. and Lu, B. (2017), "The informative content of management discussion and analysis and stock price crash risk—based on text vectorization method", *China Industrial Economics*, Vol. 12, pp. 132-150.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S. and Dean, J. (2013), "Distributed representations of words and phrases and their compositionality", *Proceedings of the 26th International Conference on Neural Information Processing Systems*, 26, Advances in Neural Information Processing Systems, Vol. 2, pp. 3111-3119.

Nassirtoussi, A.K., Aghabozorgi, S., Wah, T.Y. and Ngo, D.C.L. (2014), "Text mining for market prediction: a systematic review", *Expert Systems with Applications*, Vol. 41, pp. 7653-7670.

Nguyen, T.H., Shirai, K. and Velcin, J. (2015), "Sentiment analysis on social media for stock movement prediction", *Expert Systems with Applications*, Vol. 42 No. 24, pp. 9603-9611.

Oliveira, N., Cortez, P. and Areal, N. (2017), "The impact of microblogging data for stock market prediction: using Twitter to predict returns, volatility, trading volume and survey sentiment indices", *Expert Systems with Applications*, Vol. 73 No. 1, pp. 125-144.

Patel, J., Shah, S., Thakkar, P. and Kotecha, K. (2015), "Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques", *Expert Systems with Applications*, Vol. 42 No. 1, pp. 259-268.

Schumaker, R.P., Zhang, Y., Huang, C.-N. and Chen, H. (2012), "Evaluating sentiment in financial news articles", *Decision Support Systems*, Vol. 53 No. 3, pp. 458-464.

Sun, Y. (2010), "Do MD&A disclosures help users interpret disproportionate inventory increases", *The Accounting Review*, Vol. 85 No. 4, pp. 1411-1440.

Suykens, J. and Vandewalle, J. (1999), "Least squares support vector machine classifiers", *Neural Processing Letters*, Vol. 9, pp. 293-300.

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

Wang, Y.-F. (2002), "Predicting stock price using fuzzy grey prediction system", *Expert Systems with Applications*, Vol. 22 No. 1, pp. 33-38.

Wang, Y.-F. (2003), "Mining stock price using fuzzy rough set system", *Expert Systems with Applications*, Vol. 24 No. 1, pp. 13-23.

Wang, K. (2021), "Is the tone of risk disclosures in MD&As relevant to debt markets? Evidence from the pricing of credit default swaps", *Contemporary Accounting Research*, Vol. 38 No. 2, pp. 1465-1501.

Weng, B., Ahmed, M.A. and Megahed, F.M. (2017), "Stock market one-day ahead movement prediction using disparate data sources", *Expert Systems with Applications*, Vol. 79 No. 1, pp. 153-163.

Xu, J. and Zou, L. (2019), "The impact of CEO pay and its disclosure on stock price crash risk: evidence from China", *China Finance Review International*, Vol. 9 No. 4, pp. 479-497.

Xue, S., Xiao, Z.Z. and Pan, M.l. (2010), "Does management discussion and analysis provide useful information - empirical exploration based on loss-making listed companies", *Management World*, Vol. 5, pp. 130-140.

Yu, L., Chen, H., Wang, S. and Lai, K.K. (2009), "Evolving least squares support vector machines for stock market trend mining", *IEEE Transactions on Evolutionary Computation*, Vol. 13 No. 1, pp. 87-102.

Yu, Y., Duan, W. and Cao, Q. (2013), "The impact of social and conventional media on firm equity value: a sentiment analysis approach", *Decision Support Systems*, Vol. 55 No. 4, pp. 919-926.

Zhang, N., Lin, A. and Shang, P. (2017), "Multidimensional k-nearest neighbor model based on EEMD for financial time series forecasting", *Physica A: Statistical Mechanics and Its Applications*, Vol. 477 No. 1, pp. 161-173.

Corresponding author
Zhiyong Li can be contacted at: liz@swufe.edu.cn