Economic Modelling 106 (2022) 105709

Contents lists available at ScienceDirect

Economic Modelling

journal homepage: www.journals.elsevier.com/economic-modelling

# Financial distress prediction by combining sentiment tone features

Shuping Zhao [a,b], Kai Xu [a,b,*], Zhao Wang [a,b], Changyong Liang [a,b], Wenxing Lu [a,b], Bo Chen [a]

[a] School of Management, Hefei University of Technology, Hefei, Anhui, 230009, China
[b] Key Laboratory of Process Optimization and Intelligent Decision-Making, Ministry of Education, Hefei, 230009, China

ARTICLE INFO

*JEL classification:*
G32
G33
G34
G41

*Keywords:*
Financial distress prediction
Comments on online stock forums
Management discussion and analysis
Financial statement notes
CatBoost

ABSTRACT

In addition to financial features, we propose a novel framework that combines sentiment tone features extracted from comments on online stock forums, management discussion and analysis, and financial statement notes, to predict financial distress. We evaluate the proposed framework using data from the Chinese stock market between 2016 and 2020. We find that financially distressed companies are more likely to have weak sentiment tones as investors have a negative attitude toward the operation and financial status of the companies, while normal companies are to the contrary. Additionally, the sentiment tones of comments within one month most effectively reflect such correlations. We recommend incorporating sentiment tone features as they contribute to predictive performance improvements of all models using financial features only, and using the CatBoost model as it outperforms all benchmarked models with its ability to capture complex feature relationships. Economic benefits analysis shows that the proposed framework can correctly identify more financially distressed companies.

## 1. Introduction

With the continuous expansion of the scale of the securities market and the gradual development of the capital market, companies are facing highly complex and changeable economic environments and huge competitive pressure (Liu et al., 2020). Without a good mechanism to adapt to this environment, companies are more likely to encounter problems, such as insufficient liquidity and excessive liabilities, which eventually lead to financial distress (Wang et al., 2018). As one of the most concerning issues in the economic field, financial distress can adversely affect the sustainable development of companies, threaten their survival, and result in huge losses to investors, creditors, customers, and other stakeholders (Kim and Upneja, 2014). When the number of companies in financial distress accumulates to a certain level, social financial distress may occur and the stability and sustainability of macroeconomic development may be endangered (Mai et al., 2019).

Life cycle theory holds that the crisis of any company or industry is a process of gradual deterioration (Bhandari and McGrattan, 2020). If stakeholders can reduce information asymmetry and receive warning signals in the early stages of a crisis, then they can make reasonable decisions and reduce or avoid huge losses from financial distress. Therefore, building an accurate financial distress early warning model and reducing information asymmetry for economic entities are of great importance (Wang et al., 2018). Specifically, investors can make investment decisions in advance based on predicted results to avoid or reduce investment losses; creditors can rely on the probability of default to make debt decisions and rating pricing; suppliers can adjust their sales strategy; and regulators can monitor the financial situation of companies and curb systemic risks. To this end, financial distress prediction has become a task related to academic researchers and practitioners of common concern.

Previous studies of financial distress prediction are mainly based on market and accounting information from financial statements. As early as the 1960s, some scholars built a prediction model of financial distress by using the current ratio, asset turnover, and other financial indicators (Kim and Upneja, 2014). In recent years, some scholars have recognized the significance of textual information and used financial statement notes (FSN) and management discussion and analysis (MD&A) to predict corporate bankruptcy (Mai et al., 2019; Wang et al., 2018). The results show that the emotion of the text relates to financial distress and improves the performance of the prediction model. However, MD&A and FSN are derived from annual financial reports of listed companies. Annual financial reports are frequently falsified because of financing, tax, and other interests, and this situation can easily pass off incorrect information (Beatty et al., 2013). The release of text information such as financial reports also often lag behind, which hinders immediate judgments from stakeholders (Breuer, 2021). Therefore, finding more timely and accurate valuable information to add to the financial distress

* Corresponding author. School of Management, Hefei University of Technology, Hefei, Anhui, 230009, China.
*E-mail address:* xukaik5@163.com (K. Xu).

https://doi.org/10.1016/j.econmod.2021.105709
Received 29 April 2021; Received in revised form 14 November 2021; Accepted 15 November 2021
Available online 17 November 2021

prediction is important.

The online stock forum is a forum dedicated to investors to enable them to exchange their experience in the stock market in real-time, release the latest financial news, invite popular experts to write special comments, and participate in online Q&A. This online forum is an important way for investors to publish and share information, which promotes information dissemination and interaction while reducing information asymmetry (Jiang et al., 2019). Investors' comments on online stock forums often contain their own sentiments and opinions on company operations and stock price changes (Li et al., 2018). In recent years, researchers have discovered that comments on online stock forums (COSF) can influence stock trends and facilitate the risk perception of listed companies (Antweiler and Frank, 2004; Ruan et al., 2020). In this research, we obtain effective information from COSF and use them as the supplement of accounting information to overcome the shortcomings of easy falsification and lag of financial reports. Further, we try to combine COSF with MD&A and FSN to predict financial distress.

However, effectively integrating accounting and text information to establish a financial distress prediction model is challenging (Lang and Stice-Lawrence, 2015), especially after multiple types of text information are added. To address this, we propose a framework to combine heterogeneous text information for predicting financial distress. In our proposed framework, we consider three types of text information as follows: COSF, MD&A, and FSN. First, we obtain effective information from COSF, classify each comment according to emotional bias, and convert it into features. Second, the COSF are divided into four categories according to the periods, and the influence of the COSF during different periods on the financial distress prediction model is examined. Third, we add COSF, MD&A, and FSN to the prediction model and compare the importance of different types of text features. Fourth, we introduce an advanced ensemble learning method, namely CatBoost, to build the prediction model to accommodate complex relationships between features, especially when multiple text features are added. CatBoost is a new machine learning algorithm that can automatically process category features, utilize the relationships between features, and greatly enrich the feature dimension (Xia et al., 2020). Finally, we compare the discrimination performance of CatBoost with benchmark models, namely, logistic regression (LR), support vector machine (SVM), decision tree (DT), XGBoost, and artificial neural network (ANN) on multiple feature sets. Empirical evaluation using data of Chinese listed companies show that the introduced text features and CatBoost significantly improve the discrimination performance of the financial distress prediction model.

The main contributions of this study are as follows. First, we propose a framework of financial distress prediction that combines three sentiment tone features. To the best of our knowledge, this study is the first to apply COSF in financial distress prediction, and we compare its contribution during different periods to financial distress prediction models. We also compare the importance of different semantic text features. Second, we introduce CatBoost to accommodate complex relationships between features, which is expected to enrich the modeling tools for predicting financial distress. Third, empirical evaluation using the data of Chinese listed companies shows that the discrimination performance of CatBoost is better than the benchmark methods on the data sets.

The remainder of this paper is as follows. Section 2 reviews the relevant literature and outlines our goals. Section 3 presents a financial distress prediction framework that combines heterogeneous text information. Section 4 conducts empirical research based on sample data. Section 5 analyzes and discusses the experimental results in detail and conducts a robustness test. Section 6 summarizes the prospects of this paper and future work.

## 2. Literature review

Given the importance of predicting financial distress, it has received extensive attention from researchers since the 1960s. Throughout the existing research, obtaining valid features and building high-performance models are two important directions (Wang et al., 2018).

### 2.1. Features in financial distress prediction

In the early 1930s, some scholars pioneered an attempt to compare the financial ratios of failed and successful companies (Almamy et al., 2016). The results proved that financial indicators are closely related to companies' financial performance. Since then, most research has used financial ratios to predict financial distress. Kim and Upneja (2014) used the financial ratios of profitability, solvency, liquidity, activity and growth to distinguish between restaurants with financial difficulties and restaurants with non-financial difficulties. However, the financial indicators are calculated under the specific financial supervision framework, which can only reflect past operations and financial situations of companies and not other important information (Wang et al., 2018). Researchers gradually realized the limitations of financial indicators and began to introduce more dimensions of information (Chen, 2014). The research of Liang et al. (2020) proved that the shareholding ratio of major shareholders is an important feature for predicting financial distress.

Recently, more research has shown that text information can be an effective supplement to quantitative financial information (Mai et al., 2019). The text information in annual financial reports and news reports contain a lot of supplementary information on companies' current financial situations and future prospects for development. Thus, some researchers have tried to apply these multi-source text information to predict credit risk and financial distress (Mai et al., 2019). Wang et al. (2018) extracted key phrases/descriptions from annual reports to predict financial distress. The results showed that the annual report can effectively distinguish between normal companies and financially distressed ones. Mai et al. (2019) used MD&A as text features into the model and showed that the discrimination performance of financial distress prediction can be further improved.

The application of text features has become an important research direction to improve the discrimination performance of predicting financial distress. However, most studies only add a single text feature, derive features from the annual financial report of companies, and lack the exploration of other text information and understanding of the role of different text features in the financial distress prediction model.

### 2.2. Models of predicting financial distress

In the past decades, scholars have developed several models to predict financial distress effectively. These models can be roughly categorized into two types: models based on statistical methods and models based on artificial intelligence technology. Regarding statistical methods, the main methods used are discriminant, factor, and LR analysis because of their low complexity and easy operation (Mai et al., 2019). As early as the 1960s, some scholars used multiple linear discriminant analyses and LR to predict financial distress and prove the effectiveness of the models (Kim and Upneja, 2014). However, these methods of traditional statistical have many restrictive hypotheses, such as linear, normality, and independence hypotheses (Wang et al., 2018). In practice, these hypotheses are difficult to satisfy simultaneously. Hence, the effectiveness and applicability of these models are often limited.

Recently, methods of artificial intelligence technology have received widespread attention and provided many fruitful research routes (Mai et al., 2019). Compared with traditional statistical methods, methods of artificial intelligence technology do not have strict restrictive assumptions on the distribution of data; they can also handle large scale data sets and express nonparametric and nonlinear relationships (Wang et al., 2018). For example, Geng et al. (2015) established a model of predicting financial crisis based on three different time windows using data mining technology and neural network. They found that the discrimination performance of the model is accurate than those of other classifiers.

Simultaneously, SVM is widely used in financial distress prediction because of its strong nonlinear mapping and generalization ability. Further, Mai et al. (2019) used deep learning technology to extract information from text data for building predictive models and showed that deep learning has high discrimination performance in predicting corporate bankruptcy.

Although many methods for predicting financial distress are available, the complexity of data and difficulty of practical application often make the single classification method ineffective, especially after many categories and text features are added. Therefore, researchers have investigated the integration of multiple classification methods, that is, ensemble learning (Wang et al., 2018). Carmona et al. (2019) used the XGBoost algorithm to predict bank failure and found that it has better discrimination performance in predicting financial distress than other methods. Some scholars established models based on bagging and AdaBoost and compared them with a single neural network classifier. The results showed that the ensemble algorithm can significantly improve prediction performance (Jayasekera, 2018). Tsai et al. (2021) proved that classifier ensembles are likely to outperform single classifiers in the imbalanced sample of financial distress prediction.

A large number of research show that the prediction performance of ensemble learning is significantly better than that of a single classifier. In this paper, we try to introduce an advanced integrated classification algorithm, namely CatBoost, to financial distress prediction for accommodating complex relationships between features. We also compare it with benchmark methods, namely, LR, DT, SVM, XGBoost, and ANN. The effectiveness and superiority of CatBoost in financial distress prediction are verified on the data of Chinese listed companies.

## 3. Research design

This paper proposes a framework of financial distress prediction that combines sentiment tone features extracted from COSF, MD&A, and FSN. Specifically, this research attempts to analyze and quantify COSF as semantic tone features and input into a financial distress prediction model to improve the discrimination performance. Concurrently, we also use the semantic tone features extracted from MD&A and FSN to study whether multiple text features can provide further incremental information for financial distress prediction. Additionally, we introduce CatBoost to accommodate complex relationships between features, especially when multiple sentiment tone features are added.

The overall prediction framework (Fig. 1) mainly includes three parts: (1) raw data acquisition; (2) data processing and feature extraction; and (3) model construction. In the data collection process, financial and text data of companies are collected from research databases and websites; in the data processing and feature extraction step, accounting information is preprocessed, feature selection is performed, and text information is semantically analyzed and quantified; and in model construction, we built CatBoost, LR, DT, SVM, XGBoost, and ANN models.

### 3.1. Theoretical foundation

In market economic activities, information asymmetry theory holds that people's mastery of information is different. The party having more information is at an advantage and can benefit from the market by transmitting reliable information to others, while the party with less information is at a disadvantage and will take the initiative to obtain information through various sources (Hu and Prigent, 2018). The dissemination of information has greatly eased the differences in the degree of information mastery and plays an important role in regulating information asymmetry (Li et al., 2018). Signaling theory holds that under information asymmetry, common signals for companies to transmit internal company information to the outside world generally include announcements of important information such as profits, dividends, and financing (Al-Malkawi et al., 2014). In the Internet era, the interactivity and contagion of information are continuously strengthened, and the platform has become an important means of signal transmission (Engelberg and Parsons, 2011; Li et al., 2018).

The stock market is highly information-sensitive. Investors with more information will proactively release value information to obtain the attention or economic interests of other investors, while investors with less information will more actively obtain information to make the most satisfactory decisions (Li et al., 2018; Aouadi et al., 2018). The online stock forum is important in allowing companies and investors to publish and share information (Jiang et al., 2019). COSF are a collection of company announcements, expert opinions, and investor discussions, which directly or indirectly reflect companies' operation and financial statuses (Antweiler and Frank, 2004; Li et al., 2018). Further, as the link of information reception and dissemination, COSF contains frequent and rich information and feedback, accelerates the dissemination of positive, negative, disclosed, and undisclosed information, and greatly reduces information asymmetry (Li et al., 2018; Jiang et al., 2019). Our research aims to extract valuable information from the information disseminated in the COSF to help predict financial distress. This information can be positive or negative, but it does not include neutral or invalid information.

### 3.2. Data acquisition

The data needed can be categorized into two types as follows: financial data based on accounting information and nonfinancial data based on text information. Listed companies will regularly disclose

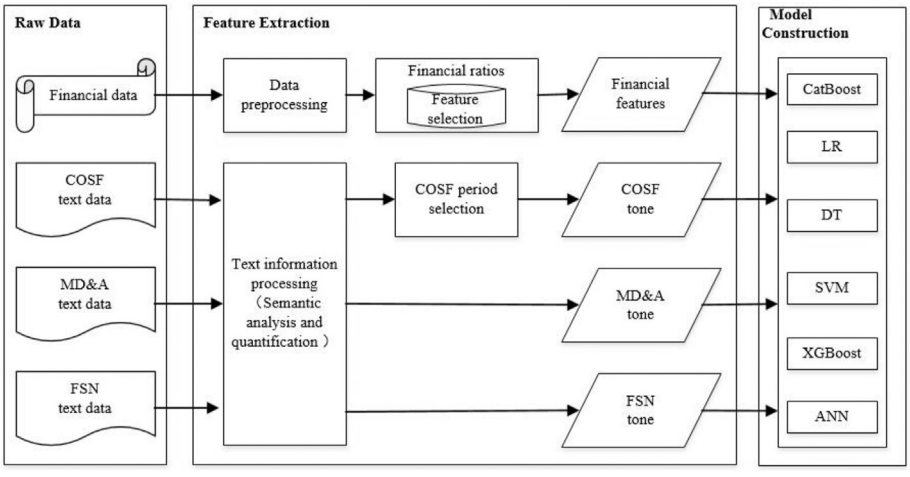


**Fig. 1.** Framework of predicting financial distress that combines sentiment tone features.

corporate accounting information online; thus, it can usually be obtained from accounting research databases (Wang et al., 2018). For the nonfinancial data, the MD&A and FSN all come from companies' annual financial reports, which are published regularly and available. COSF are information published, shared, and stored by investors in online stock forums. We can directly use information crawling tools to extract this information. Additionally, some research databases have conducted special collection and basic sentiment analysis of this text information, such as the Chinese Research Data Services Platform (CNRDS). Therefore, we can obtain text information from these research databases.

### 3.3. Data processing and feature extraction

#### 3.3.1. Financial features

According to previous research, the financial features used in financial distress prediction can be categorized into two types. The first is the financial ratios that reflect the solvency, profitability, growth ability, operating capacity, and cash flow of a company (Mai et al., 2019; Wang et al., 2018). These financial ratios are obtained from balance sheets, income statements, and other basic financial statements, which can directly reflect companies' financial and operating conditions. The other is financial indicators that reflect corporate governance and supervision capabilities (Mai et al., 2019), such as the proportion of major shareholders and shares held by the board of directors. This information will be regularly disclosed online. In predicting financial distress, selected financial ratios are generally based on three standards as follows (Wang et al., 2018): First, the ratios that have been used in existing studies. Second, calculation information for the selected financial ratio should be obtained. Third, the ratio of choice should satisfy the requirements for research. For example, researchers can choose according to their own preliminary experiments. This research also follows similar standards. We collect financial ratios and introduce the CatBoost algorithm to select important financial ratios according to the importance of characteristics. Additionally, According to the latest research results of Liang et al. (2020), we add some financial features that reflect the ability of corporate governance and supervision, that is, the shareholding ratio of major shareholders and related transactions.

#### 3.3.2. Text features

One of the key innovations of our research is that we consider using the undeveloped text data source—COSF—to predict financial distress. COSF often contain sentiments and opinions of investors on company operations and stock price changes (Tetlock et al., 2008; Li et al., 2018). The text information we use also includes MD&A and FSN. The challenge of using text information is in analyzing and quantifying the text semantically and then inputting it into the model as features. We obtain the text data from the CNRDS. Based on the dictionary built by Loughran and Mcdonald (2011) for emotional or intonation analysis of financial texts, the English vocabulary in LM dictionary is translated and combined with the Chinese context to expand and improve. Finally, artificial intelligence algorithms are used to analyze the positive and negative aspects of the text. The vocabulary is judged and recognized, and the number of positive and negative posts/words is counted. The data provided by this database have been widely used in research (Jiang et al., 2019; Brown et al., 2020; Hemmings et al., 2020; Jensen and Plumlee, 2020). Drawing on the practices of Price et al. (2012), this study measures the textual information as

$$TONE = \frac{POS - NEG}{POS + NEG} \tag{1}$$

We construct three sentiment tone features using three types of text information: COSF, MD&A, and FSN. For COSF, *POS* and *NEG* are the number of positive and negative posts, respectively, in COSF of listed companies in a certain period of T−2; for MD&A, *POS* and *NEG* are the number of positive and negative words, respectively, in MD&A of listed companies in a period T−2; and for FSN, *POS* and *NEG* are the number of positive and negative words, respectively, in FSN of listed companies in a period T−2.

#### 3.3.3. Selecting the period of comments on online stock forums

The number of comments and the information provided in the online stock forums are different in different periods. In general, a longer time means that more comments and more useful or unhelpful information are provided. However, whether greater number of comments and more information provided correspond to greater contribution to the model is unclear. This study needs to select the best period of COSF for the model to improve the discrimination performance to the greatest extent. First, we define the COSF during different periods. "COSF 12 months" is the COSF during the whole year in T−2; "COSF 6 months" is the COSF from July to December in T−2; "COSF 3 months" is the COSF from October to December in T−2; "COSF 1 month" is the COSF during December in T−2. Next, we collect the COSF in these four periods, process and quantify them, and then add them to the model. Finally, we select the period of COSF by observing the relationship between these characteristics and financial distress.

### 3.4. Model construction

The prediction of financial distress can be seen as a classification problem, divided into financial "distressed companies" and "normal companies." We introduce an advanced ensemble classification method (CatBoost) such that the prediction model can accommodate complex relationships between features and fully utilize the value of sentiment tone features. The method is compared with benchmark methods, namely, LR, DT, SVM, XGBoost, and ANN, to verify its actual prediction performance. It is worth noting that we used the default parameter settings for all the models. Among them, ANN has a two-layer hidden layer, and the number of neurons in each layer is set as 8.

#### 3.4.1. CatBoost

CatBoost is a new DT model based on gradient boosting (GB), proposed by Yandex, a Russian search giant, in 2017 (Xia et al., 2020). GB is an effective machine learning technology that can solve problems such as noise data, heterogeneous data, and complex dependencies. Compared with other algorithms of gradient boosting decision tree (GBDT), CatBoost has the following characteristics (Xia et al., 2020).

First, the algorithm adopts an effective strategy, which can convert categories into numbers without any explicit preprocessing, that is, it can directly use category features for modeling. Before modeling, traditional algorithms need to use label coding, hot coding, and other methods to deal with category features. The specific strategies adopted by the CatBoost algorithm are as follows:

(1) Sort the input sample set randomly and generate multiple groups of random sequences;
(2) Given a sequence, calculate the mean sample value of the same class for each instance;
(3) Convert all classification characteristic values for the numerical results.

With $\sigma = (\sigma_1, \cdots, \sigma_n)$ recorded as a permutation, $x_{\sigma_{p,k}}$ can be replaced by (Xia et al., 2020)

$$\frac{\sum_{j=1}^{p-1}\left[x_{\sigma_{j,k}} = x_{\sigma_{p,k}}\right] \cdot Y_{\sigma_j} + \beta \cdot P}{\sum_{j=1}^{p-1}\left[x_{\sigma_{j,k}} = x_{\sigma_{p,k}}\right] + \beta} \tag{2}$$

where $P$ is the previous value, and $\beta$ is the weight of the previous value. This algorithm helps reduce the noise obtained from the low-frequency class.

Second, CatBoost combines different types of features into new

features to obtain high-order dependencies. CatBoost will adopt strategies of greed to consider the combination when building a new division point of the current tree. For the first split of the tree, the combination is not considered. For the next split, CatBoost combines all combination and classification features of the current tree with all classification features in the dataset and dynamically converts the new combined categorical features into numerical features. CatBoost uses the following method to form a combination of numbers and classification features: all selected split points are regarded as classification features with two values and are combined and considered similar to classification features.

Third, the gradient deviation can be overcome by CatBoost. In GBDT, each iteration generates a weak learner, training for each learner of the learner based on the gradient former, and the classification results of all learners are accumulated to provide an output. However, the distribution of the estimated gradient in any domain of the feature space deviates from the real distribution of the gradient in that domain, which leads to overfitting. CatBoost improves the classical gradient lifting algorithm and uses the ordered boosting method to solve the aforementioned problem. This method further improves the model's generalization ability because it overcomes the overfitting problem caused by gradient deviation. The pseudo-codes of ordered boosting (Xia et al., 2020) are in Table 1.

CatBoost trains a separate model $M_k$ (which is composed of multiple trees) for each sample $X_k$ to obtain unbiased gradient estimation, and the model is never updated with gradient estimation based on the sample. We use $M_k$ to get the gradient estimation of sample $X_k$ (that is, the value of leaf node). Additionally, the final model of the basic learner will be trained using the gradient.

#### 3.4.2. Other models

This study introduces popular financial distress prediction methods as benchmark methods to verify the prediction performance of CatBoost.

(1) LR. This method is mature and widely used and has the advantages of simplicity, high efficiency, good interpretability, and dynamic expansion. It is broadly suitable for various classification tasks and is usually used as the benchmark model for risk prediction and analysis (Mai et al., 2019).
(2) DT. This is a typical nonparametric method and is a common machine learning algorithm. It does not need to assume a prior probability distribution and has good flexibility and robustness. Simultaneously, it can effectively suppress the problem of sample noise and missing attributes. It is an effective algorithm for predicting financial distress (Korol, 2013; Kim and Upneja, 2014).
(3) SVM. This is a generalized linear classifier that uses a supervised learning method to perform binary classification of data. The principle of classification is to maximize the interval, which overcomes the problems of overfitting, nonlinearity, dimension disaster, and small local pole in traditional methods of machine learning. It is widely used to predict credit risk and financial distress (Mai et al., 2019; Zoričák et al., 2020).
(4) XGBoost. This adopts a series GB strategy, which is representative of nonlinear and integrated learning methods. Its effectiveness has been verified in many challenges of data science and machine learning (Carmona et al., 2019).
(5) ANN. This has the characteristics of self-organization, self-adaptive, and real-time learning and overcomes the shortcomings of traditional methods in dealing with unstructured information. Its effectiveness in predicting financial distress has been verified (Mai et al., 2019).

**Table 1**
Pseudo-code of the ordered boosting.

| Algorithm: Ordered boosting |
|---|
| Input: $\{(X_k, Y_k)\}_{k=1}^{n}$ ordered according to$\sigma$, the number of trees $I$; |
| $\sigma \leftarrow$ random permutation of $[1, n]$ |
| $M_i \leftarrow 0$ for $i = 1, \cdots, n$ for t $\leftarrow$ 1 to $I$ do |
| for $i \leftarrow 1$ to $n$ do |
| $r_i \leftarrow y_i \leftarrow M_{\sigma(i)-1}(X_i)$; |
| for $i \leftarrow 1$ to $n$ do |
| $\Delta M \leftarrow LearnModel[(X_i, r_j) : \sigma(j) \leq i]$ |
| $M_i \leftarrow M_i + \cdot M$ |
| Return $M_n$ |

## 4. Empirical evaluation

This section conducts experiments based on real datasets to verify the feasibility and superiority of extracting sentiment tone features from COSF, MD&A, and FSN and using the CatBoost algorithm to predict financial distress.

### 4.1. Experimental dataset

For the experimental dataset, we firstly focus on the symbol of financial distress; however, domestic and overseas researchers are not uniform in the symbol. Some overseas researchers believed that if a company faces bankruptcy or is in arrears with preferred stock dividends and cannot repay debts, then the company is facing a financial crisis (Kim and Upneja, 2014). Mai et al. (2019) considered the entry of a company into liquidation or bankruptcy reorganization proceedings as financial distress. In comparison, Chinese scholars generally define financial distress as a company with a negative net profit for two consecutive fiscal years or a company with net assets per share that is lower than the par value of each share because of a substantial loss for one year, which is also the main standard for the special treatment (ST) of listed companies in China by the stock exchange because of abnormal financial conditions (Wang et al., 2018). This paper aligns with Chinese scholars in defining financial distress.

Given this mechanism, this paper selects the data of two years before the financial distress of companies to establish a more scientific and effective model; specifically, the financial distress is assumed to occur in year T; then, the data in year T−2 are used to establish the model (Mai et al., 2019). We selected 1427 listed manufacturing companies from the Shanghai Stock Exchange and the Shenzhen Stock Exchange as samples to avoid the heterogeneous influence of different industries and consider the economic and societal importance of the manufacturing industry. The companies were selected from 2018 to 2019, and the corresponding data interval was 2016–2017. Particularly, 67 companies were in financial distress, including 27 in 2018 and 40 in 2019. The remaining 1360 companies were normal companies, 560 in 2018 and 800 in 2019; thus, the financial distress ratio is 4.695%. We collected and extracted the financial and nonfinancial data of the 1427 listed companies from the CNRDS. Particularly, the financial data included 20 financial ratios, major shareholders' shareholding ratios, and related transactions. These financial data reflect the solvency, profitability, growth capabilities, operating capabilities, and governance and supervision capabilities of the companies. Table 2 lists the 22 financial variables. Nonfinancial data include COSF, MD&A, and FSN.

### 4.2. Selecting financial features

We have collected the financial data related to the experiment and preprocessed the data, but not all financial characteristics contribute similarly to the prediction model. We introduce the CatBoost algorithm to calculate the importance of each financial feature for reducing the model's feature dimension and improving its generalization ability. This algorithm is a feature ranking method based on the learning model, which can effectively reduce the number of features and feature dimensions as well as improve model efficiency. Fig. 2 shows the importance of 22 financial features. We have selected 16 features with a feature contribution degree greater than 2, which are referred to as basic features hereinafter.

**Table 2**
Financial variable list.

| Variable | Definition | Min | Max | Mean | SD |
|---|---|---|---|---|---|
| $X_1$ | Current assets/ current liabilities | 0.18 | 49.63 | 2.59 | 3.00 |
| $X_2$ | Total liabilities/ total assets | 0.01 | 2.58 | 0.39 | 0.20 |
| $X_3$ | (Current assets-inventory)/current liabilities | 0.03 | 44.59 | 2.08 | 2.68 |
| $X_4$ | Total liabilities/ total shareholders' equity | −18.05 | 1556.43 | 2.17 | 41.36 |
| $X_5$ | (Monetary capital + trading financial assets)/current liabilities | 0.01 | 20.51 | 1.01 | 1.50 |
| $X_6$ | Main business cost/average inventory | 0.07 | 302.07 | 4.94 | 10.61 |
| $X_7$ | Main business income/average total assets | 0.02 | 2.63 | 0.60 | 0.33 |
| $X_8$ | Main business income/average balance of accounts receivable | 0.04 | 20873.51 | 46.29 | 638.97 |
| $X_9$ | Main business income/average current assets | 0.02 | 7.14 | 1.12 | 0.70 |
| $X_{10}$ | Net profit/sales revenue | −815.64 | 100.65 | 7.65 | 27.37 |
| $X_{11}$ | Main business income of this year/main business income of last year | −95.33 | 2399.84 | 29.54 | 94.00 |
| $X_{12}$ | Net profit/average shareholders' equity | −1036.14 | 871.50 | 7.50 | 41.53 |
| $X_{13}$ | Net operating cash flow/current liabilities | −153.62 | 555.94 | −0.04 | 17.41 |
| $X_{14}$ | Net profit/average total assets | −273.55 | 112.50 | 4.05 | 13.54 |
| $X_{15}$ | Net profit of this year/net profit of last year | −11014.02 | 13132.37 | 3.87 | 679.73 |
| $X_{16}$ | Net profit/number of ordinary shares | −5.81 | 8.09 | 0.37 | 0.59 |
| $X_{17}$ | Net operating cash flow/financial expenses | −8516.92 | 111746.03 | 92.53 | 2999.80 |
| $X_{18}$ | Net profit/sales revenue | −7.39 | 1.05 | 0.06 | 0.28 |
| $X_{19}$ | Operating profit of this year/operating profit of last year | −6026.72 | 28098.52 | 105.79 | 1056.33 |
| $X_{20}$ | Total assets of this year/total assets of last year | −78.67 | 1135.46 | 21.17 | 61.88 |
| $X_{21}$ | Shareholding ratio of the largest shareholder | 0.04 | 0.89 | 0.32 | 0.14 |
| $X_{22}$ | Whether there is any related transaction | Yes: 64%, No: 36% | | | |

### 4.3. Evaluation metrics

As the two types of samples (i.e., ST and non-ST) are largely unbalanced and the costs of the two types of errors (i.e., false positive and false negative) are largely asymmetric, we did not use the standard error rate. We addressed this by using performance metrics that are not sensitive to the imbalanced data, i.e., the area under ROC curve (AUC) and Kolmogorov–Smirnov (KS) test (Kleinberg et al., 2018; Jones et al., 2019). The AUC indicator is the probability that a good sample is ranked in front of a bad sample when the classifier randomly selects a sample (Kleinberg et al., 2018). The KS indicator is a measure of the difference between the cumulative distribution of positive and negative samples, which reflects the ability of the model to distinguish between two types of samples (Jones et al., 2019). The higher the values of these two metrics, the better the model's performance.

Concurrently, we conducted 10 independent 10-fold cross-validations, obtained 100 performance evaluation values, and calculated the average value to reduce the negative effects of the variability of training set. We also used complete pairwise comparisons to verify that the introduction of COSF into the model significantly improves the discrimination performance of models with only basic features.

### 4.4. Selecting the period of comments on online stock forums

We introduce the CatBoost algorithm to rank the feature importance of COSF during different periods for selecting the best period of COSF applied to the prediction model.

Fig. 3 shows the importance of COSF during different periods based on the CatBoost algorithm. “COSF 1 month” is the most important variable, while “COSF 12 months” is the least important one. The results show that the COSF during December in T−2 contribute the most to the prediction model, and the COSF during the whole year in T−2 contribute the least. These results have two possible reasons. First, investors will have a better understanding of operating and financial status of companies closer to December because of the disclosure of more reports, and the content discussed in COSF is more related to the operating and financial situation of the company. Second, the comment timeline of the whole year is longer, the influence factors are more, and the comment base is larger. The number of positive and negative COSF of most companies is closer, and the difference in the tone of COSF is small.

We then use CatBoost, LR, DT, SVM, XGBoost, and ANN to create prediction models and add COSF during different periods to the basic features. Table 3 shows the values of AUC and KS of the five models when the COSF during different periods are added. The maximum AUC and KS values are obtained on five different models when joining “COSF 1 month,” which indicates that the variable “COSF 1 month” has the strongest predictive ability. Thus, we select December in T−2 as the best period of COSF.

## 5. Experimental results and discussion

### 5.1. Experimental results and analysis

We analyzed the discrimination performance of multiple features including accounting information and discussed the prediction and discrimination ability of different sentiment tone features. Table 4 shows the classification effect of each model with different features, and the best performance value under each feature is in bold.

First, the AUC and KS of each model under the feature set “B + COSF” are significantly higher than those of similar models that only use basic features, indicating that adding sentiment tone feature extracted from COSF can significantly enhance the prediction model's discrimination ability. Second, the AUC and KS values of each model under “B + MD&A” and “B + FSN” have been improved and are better than similar models that only use financial features. Therefore, adding MD&A or FSN improves the prediction and discrimination ability, which complements previous conclusions reached by researchers (Mai et al., 2019; Wang et al., 2018). AUC and KS indicators show that the incremental prediction performance of COSF for the model is greater than that of FSN and less than that of MD&A. Finally, the combined use of COSF, MD&A, and FSN can reach the maximum AUC and KS values in multiple models. Therefore, we can reach the following conclusions: first, adding the information extracted from COSF improves the discrimination performance. Second, the incremental prediction performance brought by COSF is

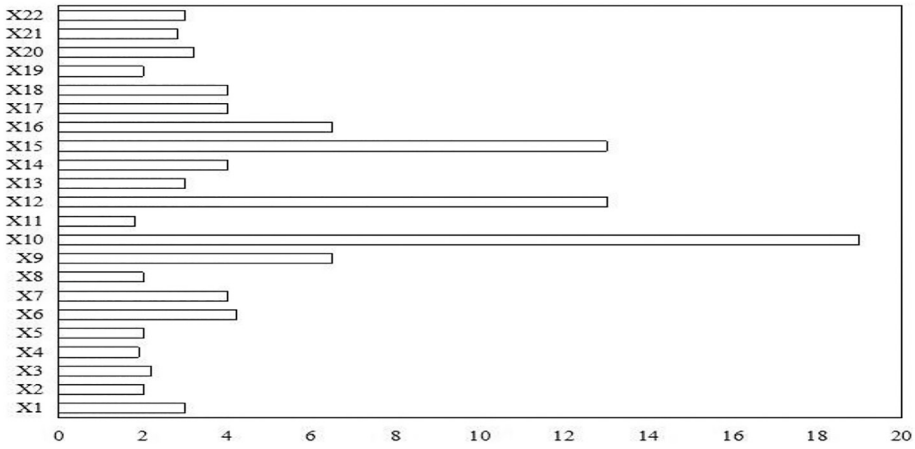


**Fig. 2.** Ranking of the importance of financial features based on CatBoost.

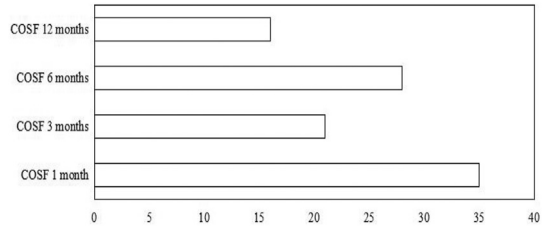


**Fig. 3.** Ranking of the importance of COSF during different time periods based on CatBoost. Notes: the best performance value under each model is shown in bold.

greater than that of the FSN and less than that of the MD&A. Third, the combined use of COSF, MD&A, and FSN can maximize the model's discrimination performance.

These results confirm that extracting valuable information from COSF can help solve the problem of financial distress prediction. On the one hand, online stock forums have become an important way for companies and investors to publish and share information, promoting the rapid and widespread dissemination of information (Li et al., 2018). This greatly reduces the cost of information acquisition for investors, improves investors' understanding of companies' operation and financial statuses, and reduces information asymmetry. On the other hand, the interactivity and contagion of information are constantly increasing. The valuable information on online stock forums is quickly disseminated and discussed after being released and shared. After frequent and repeated interactions,

**Table 3**
Discrimination performance (mean and 95% confidence interval) of COSF during different periods.

| Model | Metrics | COSF | COSF | COSF | COSF |
|---|---|---|---|---|---|
| | | 1 month | 3 months | 6 months | 12 months |
| LR | AUC | **0.819(0.799–0.839)** | 0.809 (0.789–0.829) | 0.812 (0.793–0.831) | 0.804 (0.783–0.825) |
| | KS | **0.635(0.603–0.667)** | 0.608 (0.575–0.641) | 0.624 (0.593–0.655) | 0.613 (0.579–0.647) |
| DT | AUC | **0.798(0.776–0.819)** | 0.781 (0.758–0.804) | 0.780 (0.757–0.803) | 0.766 (0.744–0.788) |
| | KS | **0.577(0.533–0.621)** | 0.553 (0.508–0.598) | 0.557 (0.511–0.603) | 0.527 (0.481–0.573) |
| SVM | AUC | **0.765(0.738–0.792)** | 0.760 (0.732–0.788) | 0.763 (0.734–0.792) | 0.753 (0.722–0.784) |
| | KS | **0.578(0.543–0.613)** | 0.576 (0.542–0.610) | 0.573 (0.538–0.608) | 0.577 (0.541–0.613) |
| XGBoost | AUC | **0.944(0.933–0.955)** | 0.939 (0.928–0.950) | 0.940 (0.927–0.953) | 0.937 (0.926–0.948) |
| | KS | **0.774(0.757–0.791)** | 0.769 (0.753–0.785) | 0.770 (0.755–0.785) | 0.769 (0.753–0.785) |
| ANN | AUC | **0.840(0.819–0.861)** | 0.832 (0.807–0.857) | 0.835 (0.809–0.861) | 0.830 (0.806–0.854) |
| | KS | **0.592(0.577–0.607)** | 0.572 (0.544–0.600) | 0.586 (0.562–0.610) | 0.561 (0.543–0.579) |
| CatBoost | AUC | **0.959(0.950–0.968)** | 0.953 (0.946–0.960) | 0.957 (0.951–0.963) | 0.953 (0.946–0.960) |
| | KS | **0.778(0.764–0.792)** | 0.771 (0.752–0.790) | 0.777 (0.762–0.792) | 0.766 (0.743–0.789) |

**Table 4**
AUC and KS under different features and methods.

| Model | Metrics | B | B + COSF | B + MD&A | B + FSN | B + ALL |
|---|---|---|---|---|---|---|
| LR | AUC | 0.791 (0.767–0.815) | 0.819 (0.799–0.839) | 0.825 (0.806–0.844) | 0.814 (0.794–0.627) | **0.835(0.817–0.853)** |
| | KS | 0.596 (0.559–0.633) | 0.635 (0.603–0.667) | 0.635 (0.602–0.668) | 0.627 (0.596–0.658) | **0.662(0.631–0.693)** |
| DT | AUC | 0.785 (0.763–0.807) | 0.798 (0.776–0.819) | 0.806 (0.789–0.823) | 0.792 (0.772–0.812) | **0.816(0.799–0.833)** |
| | KS | 0.564 (0.522–0.606) | 0.577 (0.533–0.621) | 0.609 (0.576–0.642) | 0.574 (0.531–0.617) | **0.620(0.590–0.650)** |
| SVM | AUC | 0.746 (0.718–0.774) | 0.765 (0.738–0.792) | 0.770 (0.741–0.799) | 0.765 (0.737–0.793) | **0.777(0.750–0.804)** |
| | KS | 0.557 (0.522–0.592) | 0.578 (0.543–0.613) | 0.592 (0.554–0.630) | 0.570 (0.535–0.605) | **0.598(0.562–0.634)** |
| XGBoost | AUC | 0.926 (0.914–0.938) | 0.944 (0.933–0.955) | 0.952 (0.941–0.963) | 0.944 (0.933–0.955) | **0.968(0.961–0.975)** |
| | KS | 0.764 (0.741–0.787) | 0.774 (0.757–0.791) | 0.776 (0.756–0.796) | 0.771 (0.753–0.789) | **0.781(0.760–0.802)** |
| ANN | AUC | 0.829 (0.808–0.850) | 0.840 (0.819–0.861) | 0.848 (0.822–0.874) | 0.837 (0.812–0.863) | **0.867(0.847–0.887)** |
| | KS | 0.571 (0.546–0.596) | 0.592 (0.577–0.607) | 0.619 (0.598–0.640) | 0.583 (0.564–0.602) | **0.631(0.594–0.668)** |
| CatBoost | AUC | 0.943 (0.932–0.954) | 0.959 (0.950–0.968) | 0.963 (0.955–0.971) | 0.954 (0.942–0.966) | **0.976(0.969–0.983)** |
| | KS | 0.770 (0.755–0.785) | 0.778 (0.764–0.792) | 0.780 (0.764–0.796) | 0.773 (0.756–0.790) | **0.788(0.769–0.807)** |

Notes: 95% confidence interval in the parentheses, ALL refers to COSF, MD&A, and FSN.

**Table 5**
Confusion matrices.

| | Predictive class | | | |
|---|---|---|---|---|
| | Benchmark model | | CatBoost model with COSF | |
| Actual Class | ST | Non-ST | ST | Non-ST |
| ST | 25 | 13 | 30 | 4 |
| Non-ST | 97 | 573 | 32 | 648 |

COSF often contains investors' opinions and feedback on the information obtained, and these positive or negative feedbacks largely convey the signals that investors are optimistic or bearish about companies' operation and financial statuses (Antweiler and Frank, 2004; Li et al., 2018). These signals provide incremental information for predicting financial distress, that is, financially distressed companies are more likely to have weak sentiment tones as investors have a negative attitude toward the operation and financial status of the companies, while normal companies are to the contrary.

Additionally, we compare CatBoost with benchmark methods on different feature sets to explore its prediction performance. Table 4 shows the prediction performance of each model under different methods.

First, CatBoost is significantly better than other benchmark methods in the basic feature set. Compared with the AUC and KS values of LR, DT, SVM, XGBoost, and ANN, the AUC of CatBoost is increased by 19.22%, 20.13%, 26.41%, 1.84%, and 13.75%, respectively, and its KS is increased by 29.19%, 36.52%, 38.24%, 0.79%, and 34.85%, respectively, compared with LR, DT, SVM, XGBoost, and ANN. Second, on "B + COSF," CatBoost increases the AUC by 17.09%, 20.18%, 25.36%, 1.59%, and 14.17%, respectively, compared with LR, DT, SVM, XGBoost, and ANN; it also increases KS by 22.52%, 34.84%, 34.60%, 0.52%, and 31.42%, respectively, compared with LR, DT, SVM, XGBoost, and ANN. Third, on "B + MD&A," CatBoost increases the AUC by 16.73%, 19.48%, 25.06%, 1.16%, and 13.56%, respectively, compared with LR, DT, SVM, XGBoost, and ANN, and it increases KS by 22.83%, 28.08%, 31.76%, 0.52%, and 26.01%, respectively, compared with LR, DT, SVM, XGBoost, and ANN. Fourth, on "B + FSN," CatBoost increases the AUC by 17.20%, 20.45%, 24.71%, 1.06%, and 13.99%, respectively, compared with LR, DT, SVM, XGBoost, and ANN, and it increases KS by 23.29%, 34.67%, 35.61%, 0.26%, and 32.59%, respectively, compared with LR, DT, SVM, XGBoost, and ANN. Fifth, in "B + ALL," CatBoost increases the AUC by 16.89%, 19.61%, 25.61%, 0.83%, and 12.57%, respectively, compared with LR, DT, SVM, XGBoost, and ANN, and it increases KS by 19.03%, 27.10%, 31.77%, 0.9%, and 24.88%, respectively, compared with LR, DT, SVM, XGBoost, and ANN. In summary, the prediction performance of CatBoost on all feature sets is better than those of other benchmark methods, showing that the CatBoost method is effective and superior in solving important economic problems.

### 5.2. The economic benefits of the proposed prediction framework

We analyze the benefits of our proposed framework from an economic perspective. Specifically, we simulate real credit scenarios and analyze the added value of the proposed sentiment tone features and CatBoost model compared with the benchmark model, i.e., LR. First, we extract the annual new long-term borrowings of the 1427 sample companies and calculate the average value, which is approximately RMB90 million. Given that financial distress prediction involves a long time period, we only consider the long-term borrowing. Second, we assume that the percentage of approved loans is 50%, that is, creditors decide to lend to 50% of the companies (714 companies). Next, we calculate the confusion matrix between the benchmark model and the model under our proposed framework (Table 5).

From the benchmark model, the number of ST companies is expected to be 25, whereas our proposed model indicates 30. Therefore, in comparison with the benchmark model, using our proposed model increases the efficiency of financial distress prediction by 20% (i.e., (30−25)/25). Consequently, the creditors may avoid the average loss of RMB90 million caused by each company because they predict 5 ST companies in financial distress in advance. In addition, the number of non-ST companies predicted as ST companies from the benchmark model is 97, whereas our proposed model is 32. Therefore, in comparison, our model reduces the proportion of false prediction by 67% (i.e., (32–97)/97), which reduces the opportunity cost of creditors. On this basis, creditors can increase the interest income generated by an average of RMB90 million loans per company because of correctly classifying 65 non-ST companies.

Notably, our framework also benefits to investors, operators, regulators, and other stakeholders. For example, investors can make investment decisions in advance based on predicted results to avoid or reduce investment losses; operators can make strategic adjustments and business transformations based on predicted results to reduce the possibility of financial distress; and regulators can supervise companies and industries based on predicted results to curb systemic risks.

### 5.3. Robustness test

To enhance the robustness of the above results, we selected a total of 1405 manufacturing companies from 2019 to 2020 for a robustness test, and the corresponding data interval was 2017–2018. Particularly, 55 companies were in financial distress, 40 in 2019 and 15 in 2020. The remaining 1350 companies were normal companies, 1000 in 2019 and 350 in 2020. We also conducted 10 independent 10-fold cross-validations (Table 6). The best performance value under each feature is in bold. From the AUC and KS of each model in different feature sets, the experimental results are still robust in the subperiod. Text information, such as COSF, brings incremental benefits to financial distress prediction, and the CatBoost model achieves the most satisfactory results in each

**Table 6**
Discrimination performance (mean and 95% confidence interval) of different features and methods.

| Model | Metrics | B | B + COSF | B + MD&A | B + FSN | B + ALL |
|---|---|---|---|---|---|---|
| LR | AUC | 0.780 (0.757–0.803) | 0.799 (0.785–0.813) | 0.806 (0.791–0.821) | 0.794 (0.781–0.807) | **0.819(0.806–0.832)** |
| | KS | 0.553 (0.522–0.584) | 0.566 (0.538–0.594) | 0.578 (0.547–0.609) | 0.560 (0.530–0.590) | **0.609(0.578–0.640)** |
| DT | AUC | 0.784 (0.746–0.832) | 0.804 (0.772–0.836) | 0.812 (0.784–0.840) | 0.795 (0.766–0.824) | **0.828(0.801–0.855)** |
| | KS | 0.579 (0.523–0.634) | 0.593 (0.540–0.646) | 0.596 (0.542–0.650) | 0.588 (0.533–0.643) | **0.622(0.568–0.676)** |
| SVM | AUC | 0.749 (0.734–0.764) | 0.762 (0.744–0.780) | 0.766 (0.749–0.783) | 0.759 (0.743–0.775) | **0.781(0.764–0.798)** |
| | KS | 0.556 (0.529–0.584) | 0.577 (0.545–0.609) | 0.583 (0.552–0.614) | 0.562 (0.530–0.594) | **0.604(0.573–0.635)** |
| XGBoost | AUC | 0.902 (0.890–0.914) | 0.915 (0.897–0.923) | 0.922 (0.906–0.938) | 0.909 (0.895–0.923) | **0.939(0.924–0.954)** |
| | KS | 0.749 (0.714–0.784) | 0.755 (0.723–0.777) | 0.760 (0.732–0.788) | 0.751 (0.722–0.780) | **0.769(0.742–0.796)** |
| ANN | AUC | 0.811 (0.794–0.827) | 0.835 (0.812–0.858) | 0.843 (0.819–0.867) | 0.823 (0.801–0.845) | **0.852(0.834–0.871)** |
| | KS | 0.567 (0.526–0.608) | 0.592 (0.571–0.613) | 0.598 (0.561–0.635) | 0.581 (0.552–0.610) | **0.626(0.593–0.659)** |
| CatBoost | AUC | 0.918 (0.901–0.935) | 0.933 (0.915–0.951) | 0.942 (0.931–0.953) | 0.929 (0.912–0.946) | **0.951(0.938–0.964)** |
| | KS | 0.752 (0.733–0.771) | 0.762 (0.735–0.789) | 0.768 (0.753–0.783) | 0.759 (0.732–0.786) | **0.779(0.759–0.799)** |

Notes: ALL refers to COSF, MD&A, and FSN.

feature set.

## 6. Conclusion

The ability to predict financial distress timely and effectively is important for the decision-making of companies and stakeholders. Therefore, researchers have paid close attention toward developing a financial distress prediction model with higher discrimination performance for a long time. Obtaining valid features and building high-performance models have proven to be important directions for improving the ability to accurately predict financial distress. In this paper, we study the use of COSF to supplement basic financial information and other text information and CatBoost to improve the discrimination performance of predicting the financial distress of listed companies. We propose a framework that combines sentiment tone features extracted from COSF, MD&A, and FSN. Specifically, we extract relevant features from COSF and prove that these features can significantly improve the accuracy of predicting financial distress. We examine the impact of COSF during different periods and different sentiment tone features on the financial distress prediction model. We also build six models using CatBoost, LR, DT, SVM, XGBoost, and ANN, and the performance of CatBoost is compared with the other five benchmark models on different feature sets.

The results show that, compared with the COSF in other periods, the incremental discrimination performance of the model is the highest when "COSF 1 month" is added and lowest when "COSF 12 months" is added. When using the three sentiment tone features, separately, the incremental prediction performance of COSF for the model is greater than that of FSN and less than that of MD&A. When the three sentiment tone features are used in combination, the model discrimination performance can be improved to the greatest extent. Comparing the discrimination performance of different models on multiple feature sets shows that the discrimination performance of CatBoost is significantly better than those of other benchmark methods, verifying CatBoost's effectiveness and superiority in solving important economic problems.

From a research perspective, first, we apply the COSF to financial distress prediction, verify its effectiveness, and improve the discrimination performance of the prediction model. Second, we identify the importance of the features and compare the incremental contributions of COSF during different periods and different sentiment tone features to the prediction model. Third, the CatBoost algorithm is introduced and its effectiveness and superiority are proven. From a practical perspective, the features and methods used can be utilized by investors, creditors, management, and regulatory agencies in real-world practices. The information extracted from online stock forums and other sources can help alleviate information asymmetry. It can also be used to evaluate the financial and operating status of listed companies more comprehensively and immediately for making better decisions and avoiding or reducing huge losses from financial distress.

This study has several future directions. First, more valuable text features or other features from social media can be applied in the model to further improve the discrimination performance and given the continuous changes in the financial environment. Second, this research treats financial distress prediction as a dichotomy problem. However, financial distress has different degrees, and future research needs to explore prediction models with different degrees.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgment

This work was funded by the National Natural Science Foundation of China (Grant Nos. 71601061, 71771075, 71771077, and 71731005) and the Fundamental Research Funds for the Central Universities (Grant Nos. JZ2021HGTB0066).

## References

Al-Malkawi, H., Bhatti, M.I., Magableh, S.I., 2014. On the dividend smoothing, signaling and the global financial crisis. Econ. Modell. 42, 159–165. https://doi:10.1016/j.econmod.2014.06.007.

Almamy, J., Aston, J., Ngwa, N.L., 2016. An evaluation of Altman's z-score using cash flow ratio to predict corporate failure amid the recent financial crisis: evidence from the UK. J. Corp. Finance 36, 278–285. https://doi:10.1016/j.jcorpfin.2015.12.009.

Antweiler, W., Frank, M.Z., 2004. Is all that talk just noise? The information content of internet stock message boards. J. Finance 59 (3), 1259–1294. https://doi:10.1111/j.1540-6261.2004.00662.x.

Aouadi, A., Arouri, M., Roubaud, D., 2018. Information demand and stock market liquidity: international evidence. Econ. Modell. 70, 194–202. https://doi:10.1016/j.econmod.2017.11.005.

Beatty, A., Liao, S., Yu, J.J., 2013. The spillover effect of fraudulent financial reporting on peer firms' investments. J. Account. Econ. 55 (2–3), 183–205. https://doi:10.1016/j.jacceco.2013.01.003.

Bhandari, A., McGrattan, E.R., 2020. Sweat equity in U.S. private business. Q. J. Econ. 136 (2), 727–781. https://doi:10.1093/qje/qjaa041.

Breuer, M., 2021. How does financial-reporting regulation affect industry-wide resource allocation? J. Account. Res. 59 (1), 59–110. https://doi:10.1111/1475-679X.12345.

Brown, N.C., Crowley, R.M., Elliott, W.B., 2020. What are you saying? Using topic to detect financial misreporting. J. Account. Res. 58 (1), 237–291. https://doi:10.1111/1475-679X.12294.

Carmona, P., Climent, F., Momparler, A., 2019. Predicting failure in the u.s. banking sector: an extreme gradient boosting approach. Int. Rev. Econ. Finance 61, 304–323. https://doi:10.1016/j.iref.2018.03.008.

Chen, M.Y., 2014. Using a hybrid evolution approach to forecast financial failures for taiwan-listed companies. Quant. Finance 14 (6), 1047–1058. https://doi:10.1080/14697688.2011.618458.

Engelberg, J.E., Parsons, C.A., 2011. The causal impact of media in financial markets. J. Finance 66 (1), 67–97. https://doi:10.1111/j.1540-6261.2010.01626.x.

Geng, R., Bose, I., Chen, X., 2015. Prediction of financial distress: an empirical study of listed Chinese companies using data mining. Eur. J. Oper. Res. 241 (1), 236–247. https://doi:10.1016/j.ejor.2014.08.016.

Hemmings, D., Hodgkinson, L., Williams, G., 2020. It's OK to pay well, if you write well: the effects of remuneration disclosure readability. J. Bus. Finance Account. 47 (5–6), 547–586. https://doi:10.1111/jbfa.12431.

Hu, Y., Prigent, J.L., 2018. Information asymmetry, cluster trading, and market efficiency: evidence from the Chinese stock market. Econ. Modell. 80, 11–22. https://doi:10.1016/j.econmod.2018.04.001.

Jayasekera, R., 2018. Prediction of company failure: past, present and promising directions for the future. Int. Rev. Financ. Anal. 55, 196–208. https://doi:10.1016/j.irfa.2017.08.009.

Jensen, T.K., Plumlee, M.A., 2020. Measuring news in management range forecasts. Contemp. Account. Res. 37 (3), 1687–1719. https://doi:10.1111/1911-3846.12570.

Jiang, L., Liu, J., Yang, B., 2019. Communication and comovement: evidence from online stock forums. Financ. Manag. 48 (3), 805–847. https://doi:10.1111/fima.12245.

Jones, D., Molitor, D., Reif, J., 2019. What do workplace wellness programs do? evidence from the Illinois workplace wellness study. Q. J. Econ. 134, 1747–1791. https://doi:10.1093/qje/qjz023.

Kim, S.Y., Upneja, A., 2014. Predicting restaurant financial distress using decision tree and AdaBoosted decision tree models. Econ. Modell. 36, 354–362. https://doi:10.1016/j.econmod.2013.10.005.

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., Mullainathan, S., 2018. Human decisions and machine predictions. Q. J. Econ. 133 (1), 237–293. https://doi:10.1093/qje/qjx032.

Korol, T., 2013. Early warning models against bankruptcy risk for Central European and Latin American enterprises. Econ. Modell. 31, 22–30. https://doi:10.1016/j.econmod.2012.11.017.

Lang, M., Stice-Lawrence, L., 2015. Textual analysis and international financial reporting: large sample evidence. J. Account. Econ. 60 (2–3), 110–135. https://doi:10.2139/ssrn.2407572.

Li, X., Shen, D., Zhang, W., 2018. Do Chinese internet stock message boards convey firm-specific information? Pac. Basin Finance J. 49, 1–14. https://doi:10.1016/j.pacfin.2018.03.003.

Liang, D., Tsai, C.-F., Lu, H.-Y., Chang, L.-S., 2020. Combining corporate governance indicators with stacking ensembles for financial distress prediction. J. Bus. Res. 120, 137–146. https://doi:10.1016/j.jbusres.2020.07.052.

Liu, B., Xia, X.Y., Xiao, W., 2020. Public information content and market information efficiency: a comparison between China and the U.S. China Econ. Rev. 60, 101405. https://doi:10.1016/j.chieco.2020.101405.

Loughran, T., Mcdonald, B., 2011. When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. J. Finance 66 (1), 35–65. https://doi:10.1111/j.1540-6261.2010.01625.x.

Mai, F., Tian, S., Lee, C., Ma, L., 2019. Deep learning models for bankruptcy prediction using textual disclosures. Eur. J. Oper. Res. 274 (2), 743–758. https://doi:10.1016/j.ejor.2018.10.024.

Price, S., Doran, J., Peterson, D., Bliss, B., 2012. Earnings conference calls and stock returns: the incremental informativeness of textual tone. J. Bank. Finance 36 (4), 992–1011. https://doi:10.1016/j.jbankfin.2011.10.013.

Ruan, Q., Wang, Z., Zhou, Y., Lv, D., 2020. A new investor sentiment indicator (ISI) based on artificial intelligence: a powerful return predictor in China. Econ. Modell. 88, 47–58. https://doi:10.1016/j.econmod.2019.09.009.

Tetlock, P.C., Saar-Tsechansky, M., Macskassy, S., 2008. More than words: quantifying language to measure firms' fundamentals. J. Finance 63 (3), 1437–1467. https://doi:10.1111/j.1540-6261.2008.01362.x.

Tsai, C.F., Sue, K.L., Hu, Y.H., Chiu, A., 2021. Combining feature selection, instance selection, and ensemble classification techniques for improved financial distress prediction. J. Bus. Res. 130, 200–209. https://doi:10.1016/j.jbusres.2021.03.018.

Wang, G., Chen, G., Chu, Y., 2018. A new random subspace method incorporating sentiment and textual information for financial distress prediction. Electron. Commer. Res. Appl. 29, 30–49. https://doi:10.1016/j.elerap.2018.03.004.

Xia, Y., He, L., Li, Y., Liu, N., Ding, Y., 2020. Predicting loan default in peer-to-peer lending using narrative data. J. Forecast. 39, 260–280. https://doi.org/10.1002/for.2625.

Zoričák, M., Gnip, P., Drotár, P., Gazda, V., 2020. Bankruptcy prediction for small-and medium-sized companies using severely imbalanced datasets. Econ. Modell. 84, 165–176. https://doi.org/10.1016/j.econmod.2019.04.003.