Contents lists available at ScienceDirect

International Journal of Forecasting

journal homepage: www.elsevier.com/locate/ijforecast

Mining semantic features in current reports for financial distress prediction: Empirical evidence from unlisted public firms in China

Cuiqing Jiang [a], Ximei Lyu [a,*], Yufei Yuan [b], Zhao Wang [a], Yong Ding [a]

[a] School of Management, Hefei University of Technology, Hefei, Anhui, 230009, China
[b] DeGroote School of Business, McMaster University, Hamilton, Ontario, L8S 4M4, Canada

ARTICLE INFO

Keywords:
Financial distress
Current report
Semantic feature
Word embedding
Unlisted public firm

ABSTRACT

It is difficult to predict the financial distress of unlisted public firms due to their longer disclosure cycle of accounting information and more inadequate continuity of market trading information compared to listed firms. In this paper, we propose a framework to predict the financial distress of unlisted public firms using current reports. Specifically, to better represent the meaning of current report texts, we propose a semantic feature extraction method based on a word embedding technology. Empirical results show that current reports contain more effective information for predicting the financial distress of unlisted public firms compared with periodic reports. In addition, semantic features extracted using our proposed method significantly improve the predictive performance, and their enhancing effect is superior to that of topic features and sentiment features. Our study also provides implications for stakeholders such as investors and creditors.

1. Introduction

An unlisted public firm, also known as unquoted public, is a firm that has issued equity shares that are traded on an over-the-counter (OTC) market (a market of private brokers and dealers) rather than a stock exchange. Compared to listed firms, whose stocks trade on a stock exchange, unlisted public firms are unlisted and trade over-the-counter.[1] Similar to the OTC bulletin board (OTCBB) market in the United States (Bushee & Leuz, 2005), an OTC market in China named the New Third Board plays an important role in providing financial services for small and medium-sized enterprises (Liang et al., 2015). However, with the rapid expansion of the New Third Board market, the number of firms in financial distress is increasing year by year. According to the New Third Board website statistics, from 2015 to 2019, there were 9, 20, 66, 112, and 145 new financial distress cases, respectively, with the market share of these cases being 0.18%, 0.20%, 0.57%, 1.05%, and 1.62%, respectively. These cases bring severe risk for investments. With the emergence of China as one of the leading markets for international investors, the New Third Board market has attracted increasing attention. Hence, predicting the financial distress of firms in this market could provide strong support for investors and creditors, especially those in China, to make investment decisions and avoid investment risk.

While the financial distress prediction of unlisted public firms is of great significance, there is still a lack of research on the financial distress prediction of these firms. Research on financial distress prediction has primarily focused on listed firms, and has generally used the accounting and market features (Chen et al., 2010; Doumpos et al., 2015). However, compared with listed firms, most unlisted public firms in China do not disclose quarterly reports, making the update cycle of accounting information

* Corresponding author.
E-mail address: lvximei92@163.com (X. Lyu).

[1] https://www.investopedia.com/terms/u/unquoted_public_company.asp.

https://doi.org/10.1016/j.ijforecast.2021.06.011
0169-2070/© 2021 International Institute of Forecasters. Published by Elsevier B.V. All rights reserved.

longer. Meanwhile, due to poor market liquidity and low turnover rate, it is difficult to obtain their market trading information continuously and effectively. As a result, using traditional accounting and market features, it may be hard to get the expected performance for predicting the financial distress of unlisted public firms.

Recently, a few studies have tried to use management discussion and analysis (MD&A) in annual reports to predict the financial distress of listed firms (Tsai & Wang, 2017). However, annual reports belong to the category of periodic reports, which include annual reports, semi-annual reports, and quarterly reports. Periodic reports are generally static, experience hysteresis, and do not take into account in a timely manner the impact on financial distress prediction of major events that occur irregularly in business operation. In addition, most unlisted public firms do not disclose forward-looking statements in the MD&A of annual reports, which may lose important information for distress prediction. Therefore, new and effective supplementary information is highly desired to aid in the financial distress prediction of unlisted public firms.

Current reports (similar to the US 8-K filings) are another kind of document that the information disclosure system requires a company to disclose. A current report is disclosed when a major event occurs that may significantly impact securities prices. Such reports may also convey the signals of financial distress. Compared with periodic reports, current reports have three advantages in terms of financial distress prediction. First, the information disclosed is more timely. Current reports are disclosed during the "blank period" between periodic reports, indicating a stronger time efficiency. Second, the information reflects important events. Current reports disclose emergencies and other significant events occurring in the business's operation. Third, the information disclosed has objective and detailed descriptions. As shown by a real example of a description in a current report—"the company was suspected of violating rules, and relevant responsible personnel were subject to disciplinary measures and self-regulatory measures"—there are specific descriptions about the risk event. Despite these advantages, studies on financial distress prediction using current reports have not been seen yet.

Although current reports are publicly available, using them for financial distress prediction is not easy. The effective information in current reports is hidden in the qualitative and unstructured texts. Moreover, a company may produce a series of current reports over a period of time. These current reports contain multiple types of events with different descriptions. The syntactical and semantic relationships of descriptions in the reports are difficult to grasp. How to extract effective features, specifically semantic features, from current reports is a big challenge.

Therefore, in this paper, we propose a framework to incorporate current reports into the financial distress prediction of unlisted public firms. As an essential part of the framework, we propose a semantic feature extraction method to mine effective predictive information from current report texts. First, we adopt the word embedding algorithm BERT (Bidirectional Encoder Representations from Transformers) to turn words into real-valued vectors. Second, we calculate the term frequency inverse document frequency (TFIDF) of each word as its weight, and obtain the document vectors. Third, we use principal component analysis to reduce the dimension of the document vectors. Finally, we take the arithmetical average of all of a company's current report documents at each dimension as the company's semantic features.

We evaluate the predictive power of extracted semantic features using a real data set of unlisted public firms from the New Third Board market in China. We perform five groups of comparative experiments: using versus not using semantic features, current reports versus periodic reports, the semantic model versus the topic model, semantic features versus sentiment features, and using versus not using Synthetic Minority Over-sampling Technique (SMOTE). The results show that the predictive performance of current reports is superior to that of periodic reports. In addition, semantic features, topic features, and sentiment features extracted from current reports improve financial distress prediction performance for unlisted public firms. Still, the extracted semantic features lead to a bigger improvement.

The contributions of this study are threefold. First, to the best of our knowledge, this study is the first attempt to predict financial distress using current reports. Our study broadens the literature on using textual information for financial distress prediction. Second, we propose a method to extract semantic features in current reports using a contextual word embedding model (i.e., BERT). While previous studies have used Word2vec and Glove to predict financial risk, it would be hard for these methods to be adapted for contexts with a small number of documents. Therefore, we bridge this gap by introducing the advanced embedding method, i.e., BERT, to predict financial distress. Third, our study also provides important implications for stakeholders, such as investors and creditors.

The rest of this paper is arranged as follows. Section 2 reviews the relevant literature. Section 3 focuses on the methodology of mining semantic features in current reports for financial distress prediction. Section 4 presents the experimental research based on real data. Section 5 discusses and analyzes the experimental results. Section 6 is the conclusion.

2. Literature review

2.1. The definition of financial distress

Regarding the definition of financial distress, different studies have different perspectives. Hillier et al. (2013) defined financial distress as a situation where a company's operating cash flows are not sufficient to satisfy current obligations. Tinoco and Wilson (2013) defined a company getting into financial distress as the point when it meets the criteria that its earnings before interest, taxes, depreciation, and amortization (EBITDA) are lower than its financial expenses for two consecutive years, and that its market value has shown negative growth for two consecutive periods. Sun et al. (2017) and Geng et al. (2015) defined a company getting into financial distress as the point when it is identified as a special treatment (ST) by the Chinese Stock Exchange, which indicates negative cumulative earnings over two consecutive years or net asset value (NAV) per share below par book value.

In this paper, unlisted public firms under financial distress are defined as firms with ST. Based on the business rule of our focal market, i.e., the New Third Board market, an unlisted public firm is labeled ST when its net assets are negative at the end of one fiscal year.

2.2. Financial distress prediction

Research on financial distress prediction focuses on the construction of prediction models and the selection of prediction features. The prediction model of financial distress can be divided into two categories. The first category is mathematical statistical models, such as the famous Z-score model (Altman, 1968), and market structure model (Merton, 1974). There are also other statistical techniques for predicting financial distress, such as probit regression (Antunes et al., 2018). Statistical models have a simple operation and strong interpretability, but they have strict requirements on data assumptions. The second category is machine learning models, including single classifier models, such as decision tree, neural network, support vector machine, and K-nearest neighbor (Geng et al., 2015; Olson et al., 2012), and ensemble models, such as boosting, bagging, and random forest (Barboza et al., 2017; Petropoulos et al., 2020). This model is widely used in classification because it runs quickly and does not require too many assumptions.

Regarding the selection of prediction features, scholars routinely use accounting and market features. Li et al. (2015) selected financial ratios as input features for a distress prediction model. Chen et al. (2010) used market features to measure the credit risk of listed small and medium enterprises (SMEs) in China. Doumpos et al. (2015) proposed a prediction framework combining accounting features and market features. In addition, Tinoco and Wilson (2013) tested the distress predictive utility of accounting, market, and macroeconomic features for listed firms. However, these studies have mainly focused on listed firms, there being a lack of research on unlisted public firms. Compared with listed firms, unlisted public firms in China have a longer disclosure cycle of accounting features and poor continuity of market trading features. These characteristics lead to the result that traditional accounting and market features cannot achieve the expected performance for financial distress prediction of unlisted public firms.

In recent years, a few scholars have extracted effective features from MD&A texts in annual reports to supplement accounting and market features for the financial distress prediction of listed firms. Chen (2019) analyzed management tone in MD&A and found that a negative tone is more effective than a net tone for the financial distress prediction of listed Chinese firms. Tsai and Wang (2017) divided risk into five levels according to the volatility of equity value, used the MD&A corpus and a sentiment dictionary to analyze the management tones and then studied the impact of management tones on risk prediction. However, the annual report must be disclosed within four months after the end of one fiscal year, which raises the problem of information lag. Furthermore, most unlisted public firms in China do not disclose forward-looking statements in the MD&A texts of annual reports, which could significantly impact the predictive performance.

The current report is another important document required by information disclosure regulation. These reports pay more attention to the timeliness, emergency nature, and importance of the information, which may dynamically convey financial distress signals. Interestingly, a study of financial distress prediction using current reports has not been seen.

2.3. Text mining technology

Research related to textual analysis is mainly carried out from sentiment, topic, and semantics. A common method to extract meaning from texts is sentiment analysis based on "bag-of-words" techniques. Loughran and McDonald (2011) created a glossary based on the annual reports of firms in the United States from 1994 to 2008 and found that the negative glossary could better reflect the tone of the financial texts. Sentiment analysis calculates the number of positive and negative words in the document, ignoring the contextual meaning.

A popular method of topic modeling, the latent Dirichlet allocation (LDA), identifies potential topic structures from the document set and gets the distribution of words on each topic. Jiang et al. (2018) used the LDA topic model to extract topic-related features from description texts for P2P loan default prediction. Huang et al. (2018) employed the LDA model to compare the thematic contents of analyst reports and earnings conference calls and thus examined analysts' different roles in information discovery and interpretation. LDA is still within the "bag-of-words" realm, ignoring the word sequence and the word context.

A recent text representation method, word embedding, converts textual terms into digital vectors according to the semantic information of the context. In this method, the meaning of a word is derived from its context and grammatical structure, and thus the semantic information can be better captured. Mai et al. (2019) used the Word2Vec algorithm to convert MD&A text to word vectors and then input word vectors into deep learning models to improve the predictive effect of bankruptcy. Wang et al. (2020) adopted the Glove algorithm to train the word vectors of P2P loan description texts and then clustered word vectors into different cliques as semantic soft factors to evaluate lending credit risk.

These text mining techniques can quantify text data well, as long as one selects proper methods according to the characteristics of different texts or improves the original techniques. In this paper, we use the current report as supplementary information to forecast the financial distress of unlisted public firms. As for the current report texts, which are characterized by containing a variety of event types, and there being different descriptions for different event types with complex syntactical and semantic relationships, we mine effective features in current reports from the semantic aspect.

3. Methodology

3.1. Proposed framework

We propose a framework of financial distress prediction for unlisted public firms by integrating the semantic features in current reports. As for the framework, we try to extract and quantify the semantic features in current reports as a complement of the accounting features to improve the performance of the financial distress prediction of unlisted public firms. At the same time, to test whether the periodic reports of unlisted public firms can provide valuable information for financial distress prediction, like that of listed firms, we also extract the semantic features in periodic reports.

Fig. 1 shows our proposed framework, which consists of data preprocessing, prediction feature extraction, and prediction model construction. (1) In the data preprocessing stage, the accounting data's missing values and outliers are processed. Preprocessing textual data includes removing punctuation and numbers, word segmentation, and stop words and sparse words. (2) In the stage of prediction feature extraction, a set of selected accounting features and processed semantic features in disclosure reports are extracted as the input of the prediction models. (3) In the stage of prediction model construction, some machine learning models are selected to judge the validity of prediction features.

More specifically, Section 3.2 introduces the accounting feature screening method in this paper. Section 3.3 proposes the semantic feature extraction method in this study. Section 3.4 introduces four financial distress prediction models and three evaluation criteria.

3.2. Accounting feature selection

Previous studies have proposed different accounting features as the basis of financial distress prediction, in which two types of methods are generally used to acquire these features. One method is to select certain accounting features persistently based on some models, such as the Z-score model (Altman et al., 2017) or Shumway's model (Shumway, 2001). The other method is to conduct feature selection from many collected accounting features (Li et al., 2015). Relative to the first method, the second method can take advantage of more comprehensive accounting features. Thus, we refer to the second method, and compile a list of 25 accounting features reflecting unlisted public firms' profitability, solvency, operating capacity, and growth capacity.

However, the collected features may have collinearity and redundancy problems, so selecting simplified and practical features becomes a key step. There are usually filter-, wrapper- and embedded-based feature selection methods. Filter-based methods select features via univariate statistics, such as Information Gain Hancer et al. (2018). Wrapper-based methods search the space of all possible subsets of features, assessing their quality by learning and evaluating a classifier with that feature subset (Mafarja & Mirjalili, 2018). Embedded-based methods select features as part of the model construction process (Guo et al., 2019). We consider two feature selection methods: the least absolute shrinkage and selection operator (LASSO) method, and stepwise regression. The LASSO method introduces a penalty for least squares regression, uses the penalty to compress the estimated coefficient of features with less influence to 0, and then obtains the sparse coefficient model. The basic idea of stepwise regression is to add features into the model one by one, and the features that are considered insignificant by the test are deleted. In the end, all features in the regression model are significant, and there is no serious multicollinearity among them.

3.3. Proposed semantic feature extraction method

3.3.1. Word embedding

The identification of complex syntactical and semantic relationships is the key to extracting and quantifying features from texts. A common discrete representation model, the one-hot model, represents the text in binary terms with the dimension of each word being the number of words in the vocabulary; however, it is easy to cause the problem of the "dimensionality curse", where words are independent of each other so the model cannot reflect sequence information. A distributed representation model, the LDA topic model, is used to find representative topics from the text library and get the distribution of words on each topic but is insufficient at forming a vector space structure for capturing semantic information.

Word embedding is another method of text representation that can convert textual terms into digital vectors according to the semantic information of the context. A common word embedding model based on local context windows, Word2Vec, can learn the co-occurrence relation between words, and the word vectors are built on the premise of distributed assumption. Still, the model is expected to create a specific vocabulary before training word vectors and lacks dynamism.

In this study, we want to train word embedding that can express the true meaning of words more accurately, and the size of the corpus does not limit that. We start with a recently proposed contextual embedding method called BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2018), which is a method of pretrained text representation that can be used to extract high-quality language features from textual data to generate state-of-the-art predictions. The architecture of this method is based on multi-layer bidirectional transformation decoding, and its main innovation is using a masked language model pretraining method. Some words are first masked randomly; then, these words are predicted through the training model, and as a result, the word embedding representations are captured.

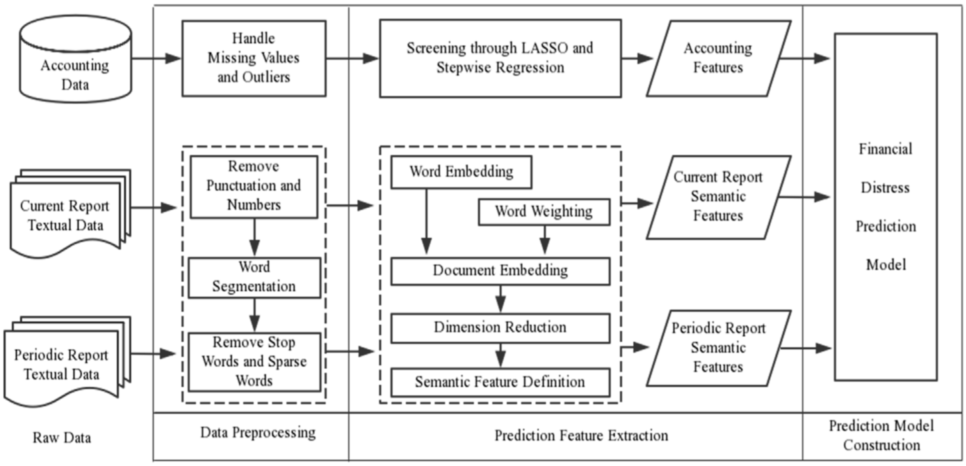
Accounting Data
Handle Missing Values and Outliers
Screening through LASSO and Stepwise Regression
Accounting Features
Current Report Textual Data
Remove Punctuation and Numbers
Word Segmentation
Remove Stop Words and Sparse Words
Periodic Report Textual Data
Word Embedding
Word Weighting
Document Embedding
Dimension Reduction
Semantic Feature Definition
Current Report Semantic Features
Periodic Report Semantic Features
Financial Distress Prediction Model
Raw Data
Data Preprocessing
Prediction Feature Extraction
Prediction Model Construction

Fig. 1. The framework of unlisted public firms' financial distress prediction.

3.3.2. Document embedding

After obtaining the word vector of each word in a document, we proceed to the embedding representation of each document since the disclosure reports (current and periodic reports) are in the form of documents. A document consists of a series of words, and document embedding can usually be approximated by simply adding up all the word vectors. Still, the importance of each word in the document is different. To better obtain the representation of document embedding, we weight the words in the document by considering their importance.

First, we construct a document-term matrix (DTM), and each entry in the DTM is the term frequency inverse document frequency (TFIDF) for each term in each current report document. TFIDF is a statistical method used to assess the importance of a word to one document in a corpus. The TF part represents the frequency at which a word appears in a document. TF has an obvious disadvantage: it gives higher weight to words that appear often but lack the power to distinguish. As a measure of the general importance of a word, the IDF part can effectively make up for the disadvantage. Given a word t and a document d, the TFIDF of t in d is calculated as Eq. (1).

$$TFIDF_{i,j} = TF_{i,j} \times IDF_i = \frac{n_{i,j}}{\sum_k n_{k,j}} \times \lg \frac{|D|}{1 + |\{d_j \in D : t_i \in d_j\}|} \tag{1}$$

where $n_{i,j}$ is the number of times that word t_i appears in document d_j. $\sum_k n_{k,j}$ is the total number of words in document d_j. $|D|$ is the total number of documents, and $|d_j \in D : ti \in d_j|$ is the number of documents where ti appears.

Second, we obtain the document embedding by multiplying the TFIDF of each word in each document by its word embedding. Assume a document d has n words, forming a sequence: $W = \{w_1, w_2, \ldots\ldots, w_n\}$. The embedding of words in the document forms the following sequence: $V = \{v_1, v_2, \ldots\ldots, v_n\}$. And the TFIDF of words in d forms a sequence: $TFIDF = \{tfidf_1, tfidf_2, \ldots\ldots, tfidf_n\}$. In the end, the embedding of document d is expressed as Eq. (2):

$$V_d = tfidf_1 \times v_1 + tfidf_2 \times v_2 + \cdots\cdots + tfidf_n \times v_n \tag{2}$$

3.3.3. Dimension reduction

In principle, high-dimensional document embedding can be directly substituted into the prediction model as semantic features. However, in practical applications, there should be an appropriate proportion between the number of experimental samples and the number of feature dimensions. A high number of feature dimensions may make the performance of a model worse, so reducing the feature dimensions is necessary.

There are two dimension reduction methods. One method is feature selection, that is, selecting a subset of the existing feature set. The other method is feature combination; several features are combined to form new features. Through the similar relationship between attributes, feature combination compresses the high-dimensional features, which is more suitable for the case of large data volume and multiple features. In our research, the high-dimensional document vectors with similar attributes can be compressed and merged according to the second dimension-reduction method.

Therefore, we adopt a common dimension reduction technology, principal component analysis, to reduce the dimension of the document embedding. As an advantage, principal component analysis maps high-dimensional coordinates to low-dimensional coordinates, compressing data while minimizing information loss. The criterion for selecting a low-dimensional coordinate is to find the eigenvalues and eigenvectors of the covariance matrix, with the eigenvectors representing the coordinate system and the eigenvalues representing the lengths mapped to the new coordinates.

3.3.4. Semantic feature definition

After dimension reduction of the document embedding, each document is represented with the same vector dimensions. We define the number of semantic features as equal to the remaining dimensions. Typically, a company has more than one current report document. We define the company's semantic features as the arithmetical average of these documents at each dimension.

As a result, we realize the extraction and quantification of semantic features in disclosure reports. These features can then be used as a complement to the accounting features to improve the performance of financial distress prediction of unlisted public firms.

3.4. Prediction modeling construction and evaluation

Financial distress prediction can be regarded as a dichotomy problem, with firms divided into financial distress and normal firms. In this paper, we use four common dichotomy-problem methods, namely logistic regression (LR), CART decision tree (CART), K-nearest neighbor (KNN), and random forest (RF), to construct the models of financial distress prediction for unlisted public firms. The LR model assumes a non-linear relationship between binary variables and explanatory variables, and the maximum likelihood estimation method is used to obtain the parameter estimation of regression coefficients. The probability threshold is set for classification. The records with the same target attribute value are divided recursively into binary nodes in the CART model to obtain binary classification results. The KNN model determines the test set's classification by finding the set of records most similar to the unclassified records from the training set. The RF model is the integration of CART and a bagging method. It first samples the training set according to a bootstrap method, then builds a CART decision tree for each sample subset, and finally synthesizes multiple decision trees to obtain the final classification result.

The area under the receiver operating characteristic curve (AUC), H-measure, and Kolmogorov–Smirnov (KS) test are selected to judge the prediction performance of each model. The AUC index reflects the comprehensive discriminability of the model to financially distressed firms and normal firms. The H-measure index sets the model loss of classification error based on data distribution, which can overcome the deficiency of the loss function change of the AUC index. The KS index is the maximum difference between the cumulative distribution of financially distressed firms and normal firms predicted by the model, reflecting the model's ability to distinguish between the two types of samples. The model's effect is better with larger values of these three indexes.

4. Empirical evaluation

4.1. Data

We have evaluated our proposed framework using data collected from the New Third Board market. To exclude the influence of the heterogeneity of different industries, we choose the information technology service industry as a representative. Our collected dataset covers all information technology service firms in 2018 and 2019, i.e., normal or under financial distress in 2018 or 2019. We follow the general practice on defining a firm under financial distress, i.e., firms with special treatment (ST). An unlisted public firm is labeled as ST when its net assets are negative at the end of one fiscal year based on the focal market's business rule.[2] Accordingly, the dataset collected for evaluation consists of 1,128 normal firms, 36 firms firstly being ST in 2018, and 33 firms firstly being ST in 2019.

Our dataset covers three types of information, i.e., accounting information, MD&A information in periodic reports, and current reports. Periodic and current reports are both written in Chinese. We set up a timeframe for obtaining features (independent variables). For normal firms and ST firms in 2019, we measure all the accounting features using data in mid-2018 (using semi-annual reports in 2018) and extract semantic features using current reports from mid-2018 to the end of 2018 (before the release of annual reports in 2018). For ST firms in 2018, we measure all the accounting features using data in mid-2017 (using semi-annual reports in 2017) and extract semantic features using current reports from mid-2017 to the end of 2017 (before the release of annual reports in 2017).

4.2. Processing of accounting data

The accounting data is collected from the China Stock Market and Accounting Research (CSMAR) Database and Choice Database. The original accounting data includes 25 features reflecting each firm's profitability, operation ability, growth ability, and debt-paying ability. We detect outliers using the boxplot method and imputed the missing values using the KNN method. We also conduct a robustness check (available in Appendix A), comparing the predictive performance of each model with and without handling outliers. We select the feature selection method with a pilot experiment using RF. The results show that stepwise regression gives a better predictive performance with a smaller number of selected features. Specifically, stepwise regression yields 0.871 in AUC with 12 selected accounting features, while LASSO yields 0.861 with 18 selected accounting features. Hence, we choose stepwise regression for our empirical evaluation. Table 1 summarizes the selected accounting features for empirical evaluation.

4.3. Processing of textual data

We collect a total of 13,798 current reports. The mean and standard deviation of the number of current reports (of each firm) are 11.527 and 9.324, respectively. An illustration of the distribution of current reports among firms is available in Appendix B.

We conduct the stage of preprocessing the text information (MD&A and current reports). First, we remove punctuation and numbers to get the text-only expressions. Second, we create a custom dictionary with 1,398 words that have specific meanings and are indivisible.

[2] The business regulation in our focal context is available at: http://www.neeq.com.cn/rule/Business_rules.html.

Table 1
Description of accounting features.

Number	Description	Mean	SD	Min	Max
1	Net profit/total equity	−0.007	0.258	−1.95	2.240
2	Capital reserve/total equity	0.572	0.840	0.000	7.900
3	Undistributed profits/total equity	0.158	0.895	−4.140	4.520
4	Operating income	16.689	1.507	8.609	24.517
5	Current net profit/prior period net profit-1	−52.087	600.461	−5296.920	4913.130
6	Net profit per share/net assets per share	−8.619	38.141	−396.010	217.200
7	Pre-tax profits/average total assets	−2.276	14.302	−123.490	71.920
8	Net interest/sales revenue	−50.209	293.790	−5053.830	78.440
9	Main business income/total assets	0.407	0.461	0.000	4.780
10	Net operating income/receivables	24.563	105.651	0.013	1636.364
11	Total liabilities/total assets	32.871	22.878	0.320	115.390
12	Current assets/current liabilities	5.285	6.760	0.200	59.000

Note: Operating income feature was $\ln(x)$ transformed.

Based on this dictionary, we use the Jieba natural language processing tool to cut the sentences into a series of separate words. Third, we remove the stop words and sparse words, which are insignificant for extracting semantic features. We create a new stop word list including 282 words by adding some meaningless words in the corpus to the Chinese general stop word list. We also delete those words appearing very rarely in all documents by setting the sparsity threshold to 0.99; that is, for each word, we divide the number of documents where it appears by the total number of documents. After the above processing, we finally produce a word list including 6,070 words by removing duplicates of all words in all documents.

In light of the practice in natural language processing, BERT can do the preprocessing by itself. However, the dimension of word or document vectors generated by BERT (e.g., 768-dimension) may be too high to train a robust model for financial distress prediction because the sample size of firms is generally small. In this regard, dimension reduction may be necessary for our focal context. Removing redundant information (e.g., stopwords and sparse words) in the text may be highly desired to yield a dense semantic feature set. We also conduct an experiment using BERT without text preprocessing and report the results in Appendix C.

After preprocessing, each word in each document is included in the final word list produced. We then conduct the stage of extracting the semantic features in four steps. First, we adopt the word embedding algorithm BERT to capture the current reports' semantic meaning and turn words into real-valued vectors. We use the BERT-Base-Chinese pre-trained model[3] with default parameters, i.e., the hidden-size, layer, and attention head are 768, 12, and 12, respectively. In the end, each word in each document is turned into a 768-dimensional real-valued vector. Second, we calculate the TFIDF of each word as its weight and obtain the weighted sum of all the word vectors in the document to get the 768-dimensional document vector. Third, we use principal component analysis to reduce the dimension of the document vectors. Based on common practice in dimension reduction, we use the covariance ratio to retain 85% of the information; then, the document vectors of the current reports are reduced from 768 dimensions to 29 dimensions, and the document vectors of the periodic reports are reduced from 768 dimensions to 38 dimensions. Finally, taking the arithmetical average of all of a company's current report documents at each dimension as the company's semantic features, we get a 1197 × 29 semantic-feature matrix.

4.4. Experimental design

We construct the financial distress indicator as a binary response variable, setting the ST company to 1 and setting the normal company to 0. Explanatory variables are the accounting features, the current-report-related features, the periodic-report-related features, or their combination. We input the response and explanatory variables into the models, namely, the LR, CART, KNN, and RF models. The output results, namely, AUC, KS, and H-measure, are used to judge the predictive power of the explanatory variables.

Before the experiment, the data set should be divided into the training set and the test set; then, we can observe the prediction effects of the models on the test set. Different scholars have different partitioning methods for different data sets, such as the leave-one method (Cecchini et al., 2010), random proportion (Geng et al., 2015), time window (Mai et al., 2019), and K-fold cross-validation (Olson et al., 2012). In this work, we make ten independent 10-fold cross-validations based on the same data set for out-of-sample prediction. This method can repeatedly use randomly generated sub-samples for training and verification at the same time, which largely avoids the limitation of insufficient or excessive training. We also conduct an out-of-time validation in Appendix D, training prediction models using the sample in the historical period and evaluating the performance of the trained models on the sample in the future period.

5. Results and analysis

5.1. Out-of-sample Prediction performance of semantic features in current reports

To test whether semantic features in current reports can improve the performance of financial distress prediction, we compare three groups of results using either

[3] https://storage.googleapis.com/bert_models/2018_11_03/chinese_L-12_H-768_A-12.zip.

Table 2
Predictive performance of semantic features in current reports.

Model	Measure	Feature		
		A	CS	A+CS
LR	AUC	0.853(0.835–0.871)	0.773(0.754–0.791)	**0.894(0.879–0.909)**
	KS	0.703(0.674–0.733)	0.561(0.531–0.591)	**0.745(0.720–0.771)**
	H-measure	0.611(0.578–0.645)	0.404(0.373–0.436)	**0.657(0.628–0.687)**
CART	AUC	0.721(0.698–0.744)	0.693(0.675–0.712)	**0.788(0.764–0.811)**
	KS	0.494(0.458–0.531)	0.451(0.420–0.482)	**0.599(0.563–0.635)**
	H-measure	0.415(0.378–0.451)	0.264(0.235–0.293)	**0.472(0.432–0.512)**
KNN	AUC	0.760(0.738–0.781)	0.711(0.694–0.728)	**0.830(0.811–0.848)**
	KS	0.540(0.502–0.577)	0.467(0.440–0.495)	**0.659(0.626–0.693)**
	H-measure	0.433(0.397–0.470)	0.258(0.232–0.283)	**0.563(0.528–0598)**
RF	AUC	0.871(0.855–0.887)	0.783(0.765–0.801)	**0.936(0.926–0.945)**
	KS	0.711(0.679–0.742)	0.571(0.542–0.600)	**0.819(0.797–0.841)**
	H-measure	0.621(0.586–0.656)	0.416(0.384–0.448)	**0.724(0.697–0.750)**

Note: "A" refers to accounting features; "CS" refers to semantic features in current reports; the best performance is in boldface; 95% confidence interval is in the parentheses.

accounting features, semantic features in current reports, or their combination. Accounting features are described in Section 4.2, and semantic features in current reports are described in Section 4.3. Table 2 summarizes the mean values and 95% confidence intervals of 100 results coming from 10 independent 10-fold cross-validations, in terms of the three evaluation standards (i.e., AUC, KS, and H-measure) of the four prediction models (i.e., LR, CART, KNN, and RF).

As shown in Table 2, AUC, KS, and H-measure show consistent patterns; that is, they have a bigger or smaller value simultaneously, suggesting that the results of the models are quite robust. Across every type of feature set (i.e., accounting features, semantic features in current reports, and their combination), the four prediction models listed in descending order of performance always are RF, LR, KNN, and CART. This reflects that the integrated model (i.e., RF) benefits more than the single-classifier models (i.e., LR, KNN, and CART). When using accounting features or semantic features alone, the accounting features are significantly superior to the semantic features in every prediction model in terms of every evaluation metric. However, when combining semantic features in current reports with accounting features, the predictive performance is better than using accounting features only. Especially, the RF model provides the best performance, and its AUC, KS, and H-measure increase to 0.936, 0.819, and 0.724, respectively. The KNN model delivers the biggest improvement, and its AUC, KS, and H-measure increase by 0.070 (from 0.760 to 0.830), 0.120 (from 0.540 to 0.659), and 0.130 (from 0.433 to 0.563), respectively.

5.2. Comparison with semantic features in periodic reports

After confirming that the semantic features extracted from current reports using our proposed framework can significantly improve the predictive performance, we now compare the predictive power of current reports with periodic reports. In addition, we can also test whether the periodic reports of unlisted public firms can provide information value for financial distress prediction, as they can with listed firms. We extract semantic features from periodic reports using the same method as current reports, as described in Section 4.3. We compare five groups of results using (1) semantic features in periodic reports, (2) semantic features in current reports, (3) the combination of accounting features and semantic features in periodic reports, (4) the combination of accounting features and semantic features in current reports, and (5) the combination of accounting features, semantic features in periodic reports, and semantic features in current reports. The comparison results are shown in Table 3.

Table 3 shows that using the semantic features in current reports demonstrates a noticeably higher AUC, KS, and H-measure compared to using semantic features in periodic reports, in terms of every prediction model. Similarly, using the combination of accounting features and semantic features in current reports always shows better performance than using the combination of accounting features and semantic features in periodic reports. These confirm that the predictive performance of current reports is superior to periodic reports.

When using semantic features in periodic reports alone, the AUC values are below 0.650 for three prediction models (i.e., CART, KNN, RF). Still, the LR model shows a higher AUC of 0.722, indicating that this type of feature has a weak predictive power. After incorporating the semantic features in periodic reports based on accounting information, the performance increases only in the CART model. In contrast, the performance decreases in the LR, KNN, and RF models. The predictive performance also decreases when adding the semantic features of periodic reports to the combination of accounting features and semantic features of current reports. These results show that the semantic features in periodic reports do not complement the accounting features and semantic features in current reports but interfere with them.

5.3. Comparison with topic features

In addition to comparing the predictive power of current reports with periodic reports, we also further compare our proposed method of mining semantic features with another well-known textual analysis method, the LDA topic model. The LDA model can find representative

Table 3
Comparison between semantic features in current reports and those in periodic reports.

Model	Measure	Feature				
		PS	CS	A+PS	A+CS	A+PS+CS
LR	AUC	0.722(0.703–0.742)	0.773(0.754–0.791)	0.833(0.812–0.853)	**0.894(0.879–0.909)**	0.874(0.858–0.891)
	KS	0.486(0.457–0.515)	0.561(0.531–0.591)	0.677(0.646–0.707)	**0.745(0.720–0.771)**	0.731(0.708–0.754)
	H-measure	0.301(0.273–0.329)	0.404(0.373–0.436)	0.579(0.545–0.613)	**0.657(0.628–0.687)**	0.619(0.591–0.646)
CART	AUC	0.586(0.572–0.599)	0.693(0.675–0.712)	0.751(0.726–0.776)	**0.788(0.764–0.811)**	0.768(0.748–0.787)
	KS	0.239(0.212–0.265)	0.451(0.420–0.482)	0.549(0.512–0.585)	**0.599(0.563–0.635)**	0.565(0.528–0.603)
	H-measure	0.099(0.077–0.120)	0.264(0.235–0.293)	0.439(0.401–0.478)	**0.472(0.432–0.512)**	0.470(0.432–0.508)
KNN	AUC	0.587(0.575–0.599)	0.711(0.694–0.728)	0.757(0.737–0.776)	**0.830(0.811–0.848)**	0.813(0.794–0.832)
	KS	0.328(0.307–0.349)	0.467(0.440–0.495)	0.538(0.506–0.570)	**0.659(0.626–0.693)**	0.632(0.598–0.667)
	H-measure	0.137(0.122–0.151)	0.258(0.232–0.283)	0.372(0.339–0.405)	**0.563(0.528–0598)**	0.540(0.504–0.577)
RF	AUC	0.641(0.620–0.662)	0.783(0.765–0.801)	0.853(0.836–0.869)	**0.936(0.926–0.945)**	0.928(0.918–0.938)
	KS	0.315(0.282–0.349)	0.571(0.542–0.600)	0.688(0.661–0.715)	**0.819(0.797–0.841)**	0.800(0.777–0.822)
	H-measure	0.189(0.159–0.218)	0.416(0.384–0.448)	0.570(0.539–0.602)	**0.724(0.697–0.750)**	0.707(0.679–0.735)

Note: "A" refers to accounting features; "CS" refers to semantic features in current reports; "PS" refers to semantic features in periodic reports; the best performance is in boldface; 95% confidence interval is in the parentheses.

Table 4
Comparison between semantic features and topic features.

Model	Measure	Feature			
		A+CT	A+CS	A+PT	A+PS
LR	AUC	0.871(0.855–0.887)	**0.894(0.879–0.909)**	0.831(0.810–0.852)	0.833(0.812–0.853)
	KS	0.726(0.700–0.752)	**0.745(0.720–0.771)**	0.677(0.646–0.707)	0.677(0.646–0.707)
	H-measure	0.634(0.604–0.664)	**0.657(0.628–0.687)**	0.565(0.531–0.600)	0.579(0.545–0.613)
CART	AUC	0.726(0.702–0.750)	**0.788(0.764–0.811)**	0.747(0.742–0.752)	0.751(0.726–0.776)
	KS	0.524(0.489–0.558)	**0.599(0.563–0.635)**	0.552(0.527–0.598)	0.549(0.512–0.585)
	H-measure	0.425(0.389–0.460)	**0.472(0.432–0.512)**	0.463(0.405–0.520)	0.439(0.401–0.478)
KNN	AUC	**0.849(0.833–0.865)**	0.830(0.811–0.848)	0.721(0.701–0.741)	0.757(0.737–0.776)
	KS	**0.698(0.669–0.728)**	0.659(0.626–0.693)	0.493(0.462–0.524)	0.538(0.506–0.570)
	H-measure	**0.579(0.547–0.611)**	0.563(0.528–0598)	0.307(0.275–0.338)	0.372(0.339–0.405)
RF	AUC	0.919(0.909–0.929)	**0.936(0.926–0.945)**	0.852(0.845–0.859)	0.853(0.836–0.869)
	KS	0.775(0.752–0.798)	**0.819(0.797–0.841)**	0.683(0.676–0.690)	0.688(0.661–0.715)
	H-measure	0.672(0.643–0.701)	**0.724(0.697–0.750)**	0.567(0.544–0.590)	0.570(0.539–0.602)

Note: "A" refers to accounting features; "CT" refers to topic features in current reports; "CS" refers to semantic features in current reports; "PT" refers to topic features in periodic reports; "PS" refers to semantic features in periodic reports; the best performance is in boldface; 95% confidence interval is in the parentheses.

topics from disclosure reports and obtain the distribution of words on each topic. The difference is that the LDA model can only obtain sparse features according to the global structure from topics to documents without considering the semantics of the word context to form a word vector space, as our method can. A parameter, the number of topics, needs to be preset in the LDA model. Considering the effect of several features, we select the same number of topic features and semantic features, as mentioned in Section 4.3: 29 topic features in current reports and 38 topic features in periodic reports. Table 4 shows the comparison results.

As shown in Table 4, both semantic features and topic features extracted from current reports improve the performance over accounting features alone. In addition, the performance of semantic features is better than that of topic features for the LR, CART, and RF models. Similarly, compared with accounting features alone, adding semantic features or topic features extracted from periodic reports improves the predictive performance only in the CART model. However, the semantic features perform better than the topic features in terms of every performance measure of the LR, KNN, and RF models.

5.4. Comparison with sentiment features

In addition to the LDA topic model, we also compare our proposed framework with sentiment analysis based on a "bag-of-words" method. We use the Chinese translated Loughran and McDonald (LM) word list to gauge the positive and negative tones. In addition, we use five Degree Dictionaries to judge the sentiment strength. They are divided into five levels, namely, most, very, more, weak, and less, weighted by 5, 4, 3, 2, and 1, respectively. Then, we use the simple proportional weighting method to calculate each document's positive sentiment score and negative sentiment score as the sentiment features. Table 5 shows the comparison results.

As shown in Table 5, both semantic features and sentiment features extracted from current reports improve the performance over accounting features alone. In addition, the performance of semantic features is better than that of sentiment features in terms of every performance measure of the four models. Similarly, compared with accounting features alone, adding sentiment features extracted from periodic reports only improves the KNN model's predictive performance. Furthermore, in contrast with current reports, the sentiment features in

Table 5
Comparison between semantic features and emotional features in current reports.

Model	Measure	Feature			
		A+CSE	A+CS	A+PSE	A+PS
LR	AUC	0.888(0.874–0.902)	**0.894(0.879–0.909)**	0.844(0.826–0.861)	0.833(0.812–0.853)
	KS	0.740(0.716–0.763)	**0.745(0.720–0.771)**	0.688(0.658–0.717)	0.677(0.646–0.707)
	H-measure	0.644(0.617–0.672)	**0.657(0.628–0.687)**	0.592(0.559–0.624)	0.579(0.545–0.613)
CART	AUC	0.741(0.719–0.764)	**0.788(0.764–0.811)**	0.713(0.689–0.737)	0.751(0.726–0.776)
	KS	0.531(0.495–0.567)	**0.599(0.563–0.635)**	0.487(0.451–0.523)	0.549(0.512–0.585)
	H-measure	0.442(0.407–0.477)	**0.472(0.432–0.512)**	0.406(0.370–0.442)	0.439(0.401–0.478)
KNN	AUC	0.803(0.784–0.822)	**0.830(0.811–0.848)**	0.789(0.770–0.808)	0.757(0.737–0.776)
	KS	0.611(0.577–0.646)	**0.659(0.626–0.693)**	0.591(0.558–0.624)	0.538(0.506–0.570)
	H-measure	0.509(0.473–0.545)	**0.563(0.528–0598)**	0.473(0.441–0.506)	0.372(0.339–0.405)
RF	AUC	0.924(0.915–0.934)	**0.936(0.926–0.945)**	0.867(0.850–0.884)	0.853(0.836–0.869)
	KS	0.791(0.768–0.813)	**0.819(0.797–0.841)**	0.709(0.678–0.740)	0.688(0.661–0.715)
	H-measure	0.696(0.670–0.722)	**0.724(0.697–0.750)**	0.619(0.584–0.653)	0.570(0.539–0.602)

Note: "A" refers to accounting features; "CSE" refers to sentiment features in current reports; "CS" refers to semantic features in current reports; "PSE" refers to sentiment features in periodic reports; "PS" refers to semantic features in periodic reports; the best performance is in boldface; 95% confidence interval is in the parentheses.

Table 6
Predictive performance of semantic features in current reports using SMOTE.

Model	Measure	Feature		
		A	CS	A+CS
LR	AUC	0.834(0.814–0.854)	0.770(0.751–0.788)	**0.884(0.867–0.902)**
	KS	0.688(0.658–0.718)	0.562(0.531–0.594)	**0.749(0.724–0.774)**
	H-measure	0.592(0.559–0.625)	0.390(0.356–0.424)	**0.656(0.628–0.684)**
CART	AUC	0.792(0.768–0.815)	0.717(0.695–0.739)	**0.814(0.790–0.838)**
	KS	0.609(0.576–0.642)	0.478(0.445–0.511)	**0.683(0.653–0.714)**
	H-measure	0.481(0.445–0.517)	0.330(0.296–0.364)	**0.541(0.507–0.575)**
KNN	AUC	0.770(0.751–0.788)	0.733(0.714–0.752)	**0.896(0.883–0.909)**
	KS	0.562(0.531–0.594)	0.511(0.483–0.540)	**0.776(0.753–0.799)**
	H-measure	0.390(0.356–0.424)	0.323(0.294–0.352)	**0.650(0.623–0.677)**
RF	AUC	0.865(0.851–0.879)	0.798(0.777–0.805)	**0.939(0.932–0.946)**
	KS	0.711(0.683–0.739)	0.596(0.571–0.620)	**0.833(0.816–0.851)**
	H-measure	0.598(0.565–0.630)	0.392(0.365–0.420)	**0.730(0.707–0.753)**

Note: "A" refers to accounting features; "CS" refers to semantic features in current reports; the best performance is in boldface; 95% confidence interval is in the parentheses.

periodic reports perform better than the semantic features in terms of the LR, KNN, and RF models.

5.5. Sensitivity analysis using SMOTE

Considering that our experiment is based on unbalanced samples, that is, the number of normal firms is far more than that of financially distressed firms, the validity of the experimental results may be affected. To avoid the possible effects, we further test the robustness of the above experimental results by using the Synthetic Minority Over-sampling Technique (SMOTE) to process the unbalanced data and observing whether the experimental results are consistent with those when not using SMOTE. Table 6 through Table 9 show the performances.

As shown in Table 6, when using SMOTE, when we incorporate the semantic features in the current reports in addition to the accounting features, the prediction effect is improved. Table 7 shows that the predictive performance of current reports is superior to that of periodic reports. Table 8 and Table 9 show that the topic features and sentiment features in current reports also improve the performance over accounting features alone. Still, this improvement is inferior to that of semantic features in current reports. However, neither the semantic features nor the topic features in periodic reports could improve the prediction performance, except the KNN model. These results are consistent with the experimental results when not using SMOTE.

5.6. Discussion and implications

First, as a contextual embedding method, BERT has its unique advantages in mining semantic features. Compared to alternative embedding models, such as Word2vec (Mai et al., 2019) and Glove (Wang et al., 2020), which generate context-independent embeddings, BERT can yield multiple vector representations based on the context. In addition, BERT is pretrained, and thus it may be more useful when the number of documents is too small, such as in our focal context, to train a high-quality language model from scratch.

Second, current reports can predict the financial distress of unlisted public firms. Previous studies have mostly focused on using the MD&A texts in periodic reports for predicting financial distress, but the utilities of text information from other channels, especially current reports, have remained ambiguous. Our empirical results

Table 7
Comparison between semantic features in current reports and those in periodic reports using SMOTE.

Model	Measure	Feature				
		PS	CS	A+PS	A+CS	A+PS+CS
LR	AUC	0.721(0.704–0.738)	0.770(0.751–0.788)	0.824(0.804–0.845)	**0.884(0.867–0.902)**	0.868(0.849–0.886)
	KS	0.502(0.473–0.530)	0.562(0.531–0.594)	0.681(0.651–0.711)	**0.749(0.724–0.774)**	0.737(0.712–0.762)
	H-measure	0.297(0.269–0.324)	0.390(0.356–0.424)	0.584(0.550–0.617)	**0.656(0.628–0.684)**	0.636(0.607–0.665)
CART	AUC	0.614(0.597–0.631)	0.717(0.695–0.739)	0.790(0.768–0.813)	**0.814(0.790–0.838)**	0.784(0.761–0.807)
	KS	0.312(0.285–0.338)	0.478(0.445–0.511)	0.600(0.566–0.635)	**0.683(0.653–0.714)**	0.637(0.607–0.668)
	H-measure	0.137(0.116–0.158)	0.330(0.296–0.364)	0.466(0.428–0.503)	**0.541(0.507–0.575)**	0.490(0.457–0.523)
KNN	AUC	0.618(0.602–0.635)	0.733(0.714–0.752)	0.797(0.777–0.816)	**0.896(0.883–0.909)**	0.876(0.861–0.892)
	KS	0.349(0.324–0.375)	0.511(0.483–0.540)	0.598(0.565–0.631)	**0.776(0.753–0.799)**	0.752(0.725–0.780)
	H-measure	0.138(0.119–0.156)	0.323(0.294–0.352)	0.429(0.395–0.463)	**0.650(0.623–0.677)**	0.628(0.597–0.659)
RF	AUC	0.671(0.651–0.692)	0.798(0.777–0.805)	0.868(0.853–0.884)	**0.939(0.932–0.946)**	0.933(0.925–0.942)
	KS	0.422(0.388–0.455)	0.596(0.571–0.620)	0.722(0.695–0.750)	**0.833(0.816–0.851)**	0.824(0.805–0.842)
	H-measure	0.251(0.217–0.284)	0.392(0.365–0.420)	0.610(0.577–0.643)	**0.730(0.707–0.753)**	0.723(0.698–0.747)

Note: “A” refers to accounting features; “CS” refers to semantic features in current reports; “PS” refers to semantic features in periodic reports; the best performance is in boldface; 95% confidence interval is in the parentheses.

Table 8
Comparison between semantic features and topic features using SMOTE.

Model	Measure	Feature			
		A+CT	A+CS	A+PT	A+PS
LR	AUC	0.849(0.829–0.868)	**0.884(0.867–0.902)**	0.825(0.805–0.844)	0.824(0.804–0.845)
	KS	0.719(0.693–0.745)	**0.749(0.724–0.774)**	0.658(0.627–0.689)	0.681(0.651–0.711)
	H-measure	0.617(0.588–0.647)	**0.656(0.628–0.684)**	0.561(0.526–0.596)	0.584(0.550–0.617)
CART	AUC	0.814(0.793–0.836)	**0.814(0.790–0.838)**	0.785(0.761–0.808)	0.790(0.768–0.813)
	KS	0.676(0.646–0.705)	**0.683(0.653–0.714)**	0.599(0.564–0.634)	0.600(0.566–0.635)
	H-measure	0.521(0.487–0.554)	**0.541(0.507–0.575)**	0.458(0.421–0.496)	0.466(0.428–0.503)
KNN	AUC	0.885(0.875–0.896)	**0.896(0.883–0.909)**	0.752(0.734–0.771)	0.797(0.777–0.816)
	KS	0.770 (0.752–0.788)	**0.776(0.753–0.799)**	0.538(0.507–0.569)	0.598(0.565–0.631)
	H-measure	0.602(0.581–0.624)	**0.650(0.623–0.677)**	0.352(0.321–0.383)	0.429(0.395–0.463)
RF	AUC	0.937(0.929–0.944)	**0.939(0.932–0.946)**	0.855(0.837–0.872)	0.868(0.853–0.884)
	KS	0.828(0.810–0.847)	**0.833(0.816–0.851)**	0.700(0.672–0.729)	0.722(0.695–0.750)
	H-measure	0.722(0.698–0.746)	**0.730(0.707–0.753)**	0.587(0.555–0.620)	0.610(0.577–0.643)

Note: “A” refers to accounting features; “CT” refers to topic features in current reports; “CS” refers to semantic features in current reports; “PT” refers to topic features in periodic reports; “PS” refers to semantic features in periodic reports; the best performance is in boldface; 95% confidence interval is in the parentheses.

Table 9
Comparison between semantic features and sentiment features using SMOTE.

Model	Measure	Feature			
		A+CSE	A+CS	A+PSE	A+PS
LR	AUC	0.863(0.846–.880)	**0.884(0.867–0.902)**	0.826(0.805–0.847)	0.824(0.804–0.845)
	KS	0.709(0.685–0.733)	**0.749(0.724–0.774)**	0.681(0.650–0.712)	0.681(0.651–0.711)
	H-measure	0.617(0.590–0.643)	**0.656(0.628–0.684)**	0.583(0.550–0.616)	0.584(0.550–0.617)
CART	AUC	0.822(0.801–0.844)	0.814(0.790–0.838)	0.796(0.774–0.819)	0.790(0.768–0.813)
	KS	0.655(0.626–0.683)	**0.683(0.653–0.714)**	0.611(0.578–0.645)	0.600(0.566–0.635)
	H-measure	0.509(0.476–0.542)	**0.541(0.507–0.575)**	0.469(0.432–0.505)	0.466(0.428–0.503)
KNN	AUC	0.844(0.826–0.861)	**0.896(0.883–0.909)**	0.790(0.771–0.810)	0.797(0.777–0.816)
	KS	0.689(0.659–0.719)	**0.776(0.753–0.799)**	0.594(0.559–0.628)	0.598(0.565–0.631)
	H-measure	0.534(0.502–0.567)	**0.650(0.623–0.677)**	0.478(0.443–0.513)	0.429(0.395–0.463)
RF	AUC	0.920(0.912–0.929)	**0.939(0.932–0.946)**	0.868(0.852–0.884)	0.868(0.853–0.884)
	KS	0.779(0.758–0.800)	**0.833(0.816–0.851)**	0.718(0.690–0.746)	0.722(0.695–0.750)
	H-measure	0.669(0.643–0.696)	**0.730(0.707–0.753)**	0.611(0.578–0.645)	0.610(0.577–0.643)

Note: “A” refers to accounting features; “CSE” refers to sentiment features in current reports; “CS” refers to semantic features in current reports; “PSE” refers to sentiment features in periodic reports; “PS” refers to semantic features in periodic reports; the best performance is in boldface; 95% confidence interval is in the parentheses.

provide strong evidence that combining semantic features extracted from current reports with accounting features could significantly improve the predictive performance. We also find that topic and sentiment features extracted from current reports have such an enhancing effect, but their effect sizes are smaller than those of semantic features. Our study opens up a new avenue of research in mining the effective information hidden in current reports and other text information to improve the performance of financial distress prediction. Investors and creditors may consider mining the semantic features from current reports and combining them with accounting features to make better investment decisions.

Third, the predictive performance of current reports is superior to that of periodic reports in the context of the financial distress prediction of unlisted public firms. Our empirical results show that although semantic features in current reports could significantly improve predictive performance, such features in periodic reports are useless and even weaken the performance of the financial distress prediction models. Interestingly, previous studies found that periodic reports, especially the MD&A texts, effectively predict the financial distress of listed firms, but most unlisted public firms do not disclose forward-looking statements in their MD&A, which may significantly limit its utility in financial distress prediction. Therefore, the financial distress risk of listed public firms and unlisted public firms may be heterogeneous. We suggest that future research distinguish listed firms from unlisted public firms when predicting financial distress. We also recommend that investors and creditors consider using current reports rather than periodic reports when predicting the financial distress of unlisted public firms.

Fourth, our proposed method of mining semantic features is superior to the LDA method and the sentiment analysis method. Previous studies mainly used only one aspect of the features, such as sentiment, without comparing the predictive performance of diverse features from different aspects. In this regard, our study provides new insights about the predictive performance of features extracted from three aspects (i.e., semantic, topic, and sentiment). It extends the application of word embedding technology, the LDA topic model, and the sentiment analysis method. For non-professional information users, it is difficult to analyze information from qualitative and unstructured texts. Our research provides a valuable tool for them to mine the effective features from texts such as current reports to predict financial distress better.

6. Conclusion

With the financial distress risk of unlisted public firms increasing year by year, investors and creditors demand more effective information to help them make investment and loan decisions. Considering the value of textual information in conveying firms' business status and development trends, we introduce the current report, a new textual channel, to predict the financial distress of unlisted public firms. Given the difficulty of quantifying the text of current reports, with their multiple types, various descriptions, and complex semantic relationships, we propose a method to mine the semantic features from current reports. Then we collect a data set of unlisted public firms in China's New Third Board market and construct prediction models to test the predictive power of these features. We also compare the predictive performances of current reports and periodic reports and the predictive performances of semantic features and topic features. In addition, we test the robustness of our experimental results by using SMOTE. The results show that current reports indeed contain effective information for financial distress prediction of unlisted public firms and that the semantic features extracted from current reports can significantly improve the predictive performance. The utility of periodic reports is ambiguous in the context of the financial distress prediction of unlisted public firms.

This work has several limitations, which may be addressed in future research. First, while the word embedding method (i.e., BERT) could effectively measure the semantics of current reports, it would be hard to explain the meanings of each dimension of the document vector. In this regard, the increased predictive performance comes at a price in model interpretability. Future research may devise ways to enhance the interpretability of our proposed framework (e.g., using LIME and Shapley values). Second, we validate the ability of current reports to predict the financial distress of unlisted public firms. However, we do not verify whether current reports are equally applicable to forecasting the financial distress of listed firms, although we think they would be. Third, this paper treats financial distress prediction as a dichotomy problem. However, there are different degrees of financial distress. Future research could explore prediction models of financial distress with different degrees, such as mild, moderate, and bankruptcy. Finally, in China's stock markets, whether a company is state-owned or has a certain percentage of stocks owned by the state may be a potential feature affecting financial distress. However, currently, this feature is unavailable in our dataset. Future research may devise ways to collect such information and incorporate it into prediction models.

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.

Acknowledgments

This work was funded by the National Natural Science Foundation of China (Grant Nos. 71731005). We thank the Special Issue Editors Professors Anthony Bellotti, Galina Andreeva, and Zhiyong Li, and the anonymous referees for their valuable comments. This paper was presented and discussed at the First Credit Scoring and Credit Rating Conference (CSCR I 2020). We thank the conference participants for their helpful comments on this paper.

Appendix A. Results of handling and not handling outliers

Table A.1 summarizes the predictive performance using accounting data with and without handling outliers. The results show that the influences of handling outliers for KNN and RF are negligible. Besides, CART benefits from handling outliers, while LR does not (i.e., a slight predictive performance decrease).

Table A.1
Results of handling and not handling outliers.

	Measure	LR	CART	KNN	RF
Handling outliers	AUC	0.853(0.835–0.871)	0.721(0.698–0.744)	0.760(0.738–0.781)	0.871(0.855–0.887)
	KS	0.703(0.674–0.733)	0.494(0.458–0.531)	0.540(0.502–0.577)	0.711(0.679–0.742)
	H-measure	0.611(0.578–0.645)	0.415(0.378–0.451)	0.433(0.397–0.470)	0.621(0.586–0.656)
Not handling outliers	AUC	0.873(0.858–0.887)	0.680(0.654–0.705)	0.758(0.739–0.777)	0.873(0.861–0.885)
	KS	0.717(0.688–0.746)	0.465(0.428–0.502)	0.542(0.509–0.576)	0.710(0.687–0.732)
	H-measure	0.621(0.590–0.652)	0.396(0.360–0.432)	0.444(0.411–0.476)	0.625(0.601–0.649)

Note: 95% confidence interval is in the parentheses.

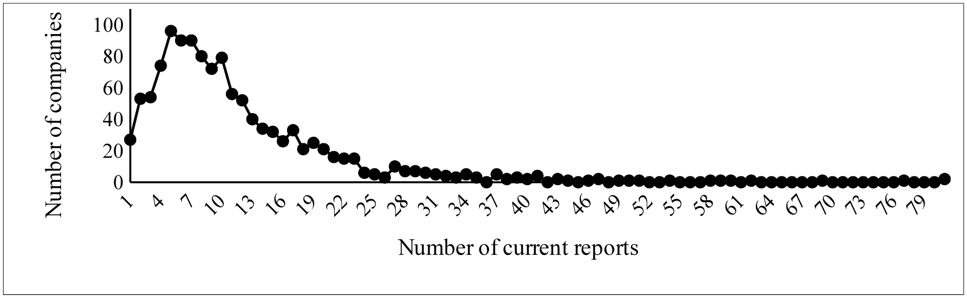

Fig. B1. Distribution of current reports among firms.

Table C.1
Predictive performance of using the BERT model without preprocessing text.

Model	Measure	Feature		
		A	CS	A+CS
LR	AUC	**0.853(0.835–0.871)**	0.717(0.697–0.736)	0.707(0.687–0.727)
	KS	**0.703(0.674–0.733)**	0.487(0.458–0.516)	0.424(0.385–0.463)
	H-measure	**0.611(0.578–0.645)**	0.331(0.302–0.361)	0.321(0.283–0.359)
CART	AUC	**0.721(0.698–0.744)**	0.616(0.596–0.636)	0.697(0.674–0.721)
	KS	**0.494(0.458–0.531)**	0.270(0.237–0.304)	0.458(0.422–0.494)
	H-measure	**0.415(0.378–0.451)**	0.144(0.115–0.173)	0.373(0.337–0.409)
KNN	AUC	**0.760(0.738–0.781)**	0.584(0.571–0.597)	0.617(0.603–0.631)
	KS	**0.540(0.502–0.577)**	0.225(0.201–0.248)	0.265(0.239–0.292)
	H-measure	**0.433(0.397–0.470)**	0.115(0.095–0.135)	0.177(0.154–0.200)
RF	AUC	0.871(0.855–0.887)	0.661(0.642–0.680)	0.874(0.859–0.888)
	KS	**0.711(0.679–0.742)**	0.417(0.393–0.441)	0.707(0.680–0.734)
	H-measure	**0.621(0.586–0.656)**	0.225(0.202–0.247)	0.599(0.570–0.628)

Note: “A” refers to accounting features; “CS” refers to semantic features in current reports; the best performance is in boldface; 95% confidence interval is in the parentheses.

Appendix B. Distribution of current reports among firms

See Fig. B1.

Appendix C. Results of using BERT without text preprocessing

Table C.1 shows the predictive performance using semantic features extracted from current reports using BERT without text preprocessing. Directly using BERT without text preprocessing results in 103 features after dimension reduction, higher than using our proposed framework (i.e., 29 features). The results show that the predictive performance of methods using our proposed framework is superior to that of models using BERT without text preprocessing.

Appendix D. Out-of-time validation

We also conduct an out-of-time validation to evaluate the utility of extracted semantic features in a real prediction scenario. We train prediction models using the sample consisting of firms that are normal or firstly ST in 2018 (historical period) and evaluate the performance of the trained models on the sample consisting of firms that are normal or firstly ST in 2019 (i.e., firms that are ST before 2019 were excluded) (future period). Table D.1 summarizes the predictive performance in terms of AUC, KS, and H-measure, of the four prediction models (i.e., LR, CART, KNN, and RF). The results show that adding the extracted semantic features in current reports improves out-of-time prediction performance over the accounting features, except KNN in KS and RF in H-measure.

Table D.1
Results of out-of-time prediction performance.

Model	Measure	Feature		
		A	CS	A+CS
LR	AUC	0.680	0.644	**0.779**
	KS	0.507	0.281	**0.515**
	H-measure	0.264	0.142	**0.297**
CART	AUC	0.841	0.524	**0.841**
	KS	0.690	0.105	**0.690**
	H-measure	0.583	0.035	**0.583**
KNN	AUC	0.819	0.595	**0.830**
	KS	0.671	0.197	0.645
	H-measure	0.486	0.111	**0.553**
RF	AUC	0.836	0.689	**0.867**
	KS	0.664	0.391	**0.684**
	H-measure	0.599	0.146	0.554

Note: “A” refers to accounting features; “CS” refers to semantic features in current reports; the best performance is in boldface.

References

Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *Journal of Finance*, *23*, 589–609.

Altman, E. I., Iwanicz-Drozdowska, M., Laitinen, E. K., & Suvas, A. (2017). Financial distress prediction in an international context: A review and empirical analysis of Altman’s Z-score model. *Journal of International Financial Management & Accounting*, *28*, 131–171.

Antunes, A., Bonfim, D., Monteiro, N., & Rodrigues, P. M. (2018). Forecasting banking crises with dynamic panel probit models. *International Journal of Forecasting*, *34*, 249–275.

Barboza, F., Kimura, H., & Altman, E. (2017). Machine learning models and bankruptcy prediction. *Expert Systems with Applications*, *83*, 405–417.

Bushee, B. J., & Leuz, C. (2005). Economic consequences of SEC disclosure regulation: Evidence from the OTC bulletin board. *Journal of Accounting and Economics*, *39*, 233–264.

Cecchini, M., Aytug, H., Koehler, G. J., & Pathak, P. (2010). Making words work: Using financial text as a predictor of financial events. *Decision Support Systems*, *50*, 164–175.

Chen, Y. (2019). Forecasting financial distress of listed companies with textual content of the information disclosure: A study based on MD & A in Chinese annual reports. *Journal of Management Science in China*, *27*, 23–34 (in Chinese).

Chen, X., Wang, X., & Wu, D. D. (2010). Credit risk measurement and early warning of SMEs: An empirical study of listed SMEs in China. *Decision Support Systems*, *49*, 301–310.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. ArXiv preprint arXiv:1810.04805.

Doumpos, M., Niklis, D., Zopounidis, C., & Andriosopoulos, K. (2015). Combining accounting data and a structural model for predicting credit ratings: Empirical evidence from European listed firms. *Journal of Banking & Finance*, *50*, 599–607.

Geng, R., Bose, I., & Chen, X. (2015). Prediction of financial distress: An empirical study of listed Chinese companies using data mining. *European Journal of Operational Research*, *241*, 236–247.

Guo, Y., Chung, F. L., Li, G., & Zhang, L. (2019). Multi-label bioinformatics data classification with ensemble embedded feature selection. *IEEE Access*, *7*, Article 103863-103875.

Hancer, E., Xue, B., & Zhang, M. (2018). Differential evolution for filter feature selection based on information theory and feature ranking. *Knowledge-Based Systems*, *140*, 103–119.

Hillier, D., Clacher, I., Ross, S., Westerfield, R., & Jordan, B. (2013). *Corporate finance (2nd European Edition)*.

Huang, A. H., Lehavy, R., Zang, A. Y., & Zheng, R. (2018). Analyst information discovery and interpretation roles: A topic modeling approach. *Management Science*, *64*, 2833–2855.

Jiang, C., Wang, Z., Wang, R., & Ding, Y. (2018). Loan default prediction by combining soft information extracted from descriptive text in online peer-to-peer lending. *Annals of Operations Research*, *266*, 511–529.

Li, Y., Meng, X., & Wei, X. (2015). China’s new third board market: Opportunities and challenges. *Procedia Computer Science*, *55*, 1050–1059.

Liang, D., Tsai, C. F., & Wu, H. T. (2015). The effect of feature selection on financial distress prediction. *Knowledge-Based Systems*, *73*, 289–297.

Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of Finance*, *66*, 35–65.

Mafarja, M., & Mirjalili, S. (2018). Whale optimization approaches for wrapper feature selection. *Applied Soft Computing*, *62*, 441–453.

Mai, F., Tian, S., Lee, C., & Ma, L. (2019). Deep learning models for bankruptcy prediction using textual disclosures. *European Journal of Operational Research*, *274*, 743–758.

Merton, R. C. (1974). On the pricing of corporate debt: The risk structure of interest rates. *The Journal of Finance*, *29*, 449–470.

Olson, D. L., Delen, D., & Meng, Y. (2012). Comparative analysis of data mining methods for bankruptcy prediction. *Decision Support Systems*, *52*, 464–473.

Petropoulos, A., Siakoulis, V., Stavroulakis, E., & Vlachogiannakis, N. E. (2020). Predicting bank insolvencies using machine learning techniques. *International Journal of Forecasting*, *36*, 1092–1113.

Shumway, T. (2001). Forecasting bankruptcy more accurately: A simple hazard model. *Journal of Business*, *74*, 101–124.

Sun, J., Fujita, H., Chen, P., & Li, H. (2017). Dynamic financial distress prediction with concept drift based on time weighting combined with adaboost support vector machine ensemble. *Knowledge-Based Systems*, *120*, 4–14.

Tinoco, M. H., & Wilson, N. (2013). Financial distress and bankruptcy prediction among listed companies using accounting, market and macroeconomic variables. *International Review of Financial Analysis*, *30*, 394–419.

Tsai, M. F., & Wang, C. J. (2017). On the risk prediction and analysis of soft information in finance reports. *European Journal of Operational Research*, *257*, 243–250.

Wang, Z., Jiang, C., Zhao, H., & Ding, Y. (2020). Mining semantic soft factors for credit risk evaluation in peer-to-peer lending. *Journal of Management Information Systems*, *37*, 282–308.