Minging Semantic Features in Current Reports for Financial Distress Prediction: Empirical Evidence from Unlisted Public Firms

Ximei LYU, Zhao Wang, Cuiqing Jiang

Abstract: It is difficult to predict financial distress for unlisted public firms due to their longer disclosure cycle of accounting information, as well as poorer continuity of market trading information in comparison with listed firms. In order to alleviate the problem of information asymmetry effectively, we propose a framework of financial distress prediction by minging semantic features in current reports. We first use a textual mining technology to extract semantic features in current reports, then combine with accounting features to predict financial distress for unlisted public firms. At the same time, we also compare the prediction performances of semantic features in current reports with that in periodic reports, and with topic features in current reports. The results show that semantic features in current reports could significantly improve prediction performance. And their prediction abilities are superior to semantic features in periodic reports and to topic features in current reports.

Keywords: Financial distress; Current report; Semantic features; Unlisted public firms

1 Introduction

The failure of a firm is an event of major concern in economic life. Consequently, the prediction of corporate financial distress has been receiving considerable attention in the field of finance and management. Financial distress here refers to a state in which a firm has not enough assets to repay matured debt. In particular, the unlisted public firm whose shares are traded in an over-the-counter (OTC) markets rather than a stock exchange, since they have relatively small size and weak risk-resistance capabilities, the financial risk events occur frequently, predicting financial distress of these firms is especially significant[1-2]. For example, similar to the OTC bulletin board (OTCBB) market of USA, an OTC market in China named as National Equities Exchange and Quotations (NEEQ), also known as “New Third Board” established in 2013 plays a more and more important role in providing financial serves for Small and medium-sized enterprises (SMEs)[3]. However, with the rapid expansion of NEEQ market, the number of unlisted public firms in financial distress is increasing year by year. According to statistics by NEEQ website, in the year of 2015, 2016, 2017, 2018 and 2019, there were 9, 20, 66, 112 and 145 financial distress new cases respectively, with the new market share accounts for 0.18%, 0.20%, 0.57%, 1.05% and 1.62% respectively. These bring increased risk in investment. Emphatically, investors and creditors may be off-guard about financial distress of firms, thus may suffer huge losses. Therefore, the effective prediction of financial distress for unlisted

public firms can provide strong support for investors and creditors to make investment decision and avoid investment risk.

Although the financial distress prediction for unlisted public firms is of great significance, there still be lack of researches on them because of the insufficient information disclosure or other reasons, and the object of financial distress prediction still be focused on listed firms. For a long time, scholars mainly use Z-Score model, market structure model and KMV model etc. based on accounting information and market information to predict the financial distress of listed firms[4-6]. However, compared with listed firms, most unlisted public firms do not disclose quarterly reports, which makes the update cycle of accounting information longer. At the same time, due to poor market liquidity and low turnover rate, it's difficult to obtain continuous and effective market trading information for unlisted public firms. It is reported that the average annual turnover rate of Chinese listed firms is over 500%, while that of unlisted public firms is only 4% in the year of 2019. As a result, the existing financial distress prediction methods for listed firms can not achieve expected performance for unlisted public firms. Therefore, it is urgent to find new and effective supplementary information.

In recent years, some researches have begun to focus on valid information in annual reports texts disclosed by firms. And these researches show that qualitative and unstructured text information can also convey the business status and development trend of firms[7-9]. Few of scholars use Management Discussion and Analysis (MD&A) in annual reports to predict the bankruptcy or financial distress of listed firms, and find that textual soft information plays an important role in supplementing traditional accounting and market features[10-11]. However, annual reports belong to periodic reports that are static, hysteresis and incomplete to some extent. And they don't take into account timely the impact of major events occurring irregularly in business operation on financial distress prediction.

The current report is another important document required by information disclosure system of unlisted public firms. Compared with the periodic report, it pays more attention to the timeliness, objectivity and importance of the information, and it covers a wide range and has strong time efficient. So the current report could make up for the lag of information in the periodic report. Surprisingly, the study on financial distress prediction using current reports has not been seen. Despite its availability, effective extraction of soft information in current reports is a big challenge due to its qualitative and unstructured characteristics. Different from the periodic report, current reports contain a variety of event types, and there are different descriptions for different event types with complex syntactical and semantic relationships. It is important to find a method to examine the lexical features of the current report text. Therefore, we propose a financial distress prediction framework for unlisted public firms by

mining semantic features in current reports. We first adopt a text representation method to extract and quantify semantic features in current reports, and then combine with accounting features to predict financial distress for unlisted public firms, and finally evaluate the prediction performance of extracted semantic features through some comparative experiment results analysis.

The main contributions of this paper are as follows. First of all, we focus on the unlisted public firms, a new object of interest in financial distress prediction. Unlisted public firms play an important role in many countries, but their financial distress prediction has not been taken seriously. We provide an favorable analysis clue of financial distress prediction for them. Secondly, we propose a new financial distress prediction framework and the semantic features in current reports is applied to financial distress prediction for the first time. So far, we can only find few researches using MD&A in annual reports to predict the bankruptcy or financial distress of listed firms, and they still face the problem of information lag. Our study provides new insights about whether the current reports could help predict financial distress of unlisted public firms. Thirdly, we extract the semantic features in current reports based on the word embedding technology, and they show good predictive ability. We provide strong evidence that the extraction method of semantic features is effective for current reports, and semantic features in current reports could promote the discriminant performance of financial distress firms and normal firms. Fourthly, our processing method for current reports can also be applied to other fields. The agency features we have seen in other field are limited to the number and frequency of disclosure, and our research can provide reference for them.

The rest of this paper are arranged as follows. Section 2 reviews the relevant literature. Section 3 focuses on financial distress prediction method integrating the semantic features in current reports. Section 4 carries out experimental research based on sample data. Section 5 discusses and analyzes the experimental results. Section 6 summarizes the conclusion and discussion.

2 Literature review

There are different definitions of financial distress from different perspectives. As widely known, Ross et al.[17] summarized previous studies and concluded that financial distresses consist of the following four conditions: business failure, that is, a company cannot pay the outstanding debt after liquidation; legal bankruptcy, namely, a company or its creditors applies to the court for a declaration of bankruptcy; technical bankruptcy, namely, a company cannot fulfill the contract on schedule to repay principal and interest; and accounting bankruptcy, namely: a company's book net assets are negative. In recent years, some scholars have defined financial distress based on market performance of a firm. For example, Tinoco et al.[19] defined a firm getting into financial distress when it meets that EBITDA are lower than financial expenses for two consecutive years, and that market value exists a

negative growth for two consecutive periods. Liang D et al.[20] defined financial distress based on the business regulations of Taiwan Stock Exchange. Sun J et al.[21], and Geng R et al.[22] defined that a firm gets into financial distress if it is identified as special treatment (ST) by Chinese Stock Exchange, which means that the profits of the firm are negative for two consecutive years and the per-share net assets are lower than per-share stock face value. In this paper, we refer to the above definition of Chinese listed firms[21-22], the financial distress of unlisted public firms is defined as ST firms in NEEQ market, which means that net assets of a firm are negative at the end of last year, or their financial reports are issued with negative or unable to express opinions.

Researches on financial distress prediction focus on two aspects: the selection of prediction features and the construction of prediction models. As for the selection of prediction features, scholars routinely use accounting information and market features. Altman et al.[4] tested the performance of Z-score model for bankruptcy prediction based on accounting information. Doumpos et al.[5] proposed a prediction framework by constructing a structural model based on accounting data. Chen et al.[6] used the KMV model to measure credit risk of listed SMEs in China. Although these methods have been widely used in the financial distress or bankruptcy prediction for listed firms, they often fail to achieve the expected effect among unlisted public firms because of the irregular accounting information and missing market features.

In recent years, with the development of natural language processing technology, scholars have begun to extract effective features from information disclosure texts as a beneficial supplement to accounting and market features.The MD&A in annual reports has become one of the most concerned contents. Scholars mainly extract the features from two aspects and conduct exploratory research on its usefulness. One is word vector. Tsai et al.[7] divided five risk levels according to the volatility of equity value, and used MD&A corpus and emotion dictionary to extract word vectors respectively, then studied their impact on risk prediction. Mai et al.[9] used deep learning model, such as average embedding model and convolutional neural network, to extract word vectors and study their predictive effect on financial distress. The other one is management tone. Loughran & McDonald[23] created the glossary based on the annual reports of the United States from 1994 to 2008, and found that the negative glossary could better reflect the tone of financial texts. Davis et al.[24] compared the management tone in MD&A and in earnings reports, and found that the tone of earnings report was more positive than in MD&A. However, the annual report can only be disclosed regularly, which cannot fully meet investors' requirements for information disclosure timely.

The current report is another important document required by the information disclosure regulation. The regulation claims that public firms should disclose information immediately in case of major events that may have

a great impact on securities prices. So the current report could make up for the lag of information in the annual report. Mcmullin et al.[12] examined the relationship between disclosure frequency of 8-K filings (current reports) and price formation, and Mcmullin et al.[13] examined the relationship between disclosure frequency and information asymmetry. As for the effect of current reports to periodic reports, DeFond et al.[14] compared the annual earnings reports of 26 different countries in the world and found that the current reports are disclosed more frequently, the less information is contained in the annual earnings reports. However, Lerman & Livnat[15] drew the opposite conclusion that current reports are a powerful supplement to annual reports, and they together convey more information about a firm. It is clear that scholars pay more attention to some features easily observed, such as the time of publication, the number or frequency of the current report, but the automatic text analysis method for the current report is still rare, and the study on financial distress prediction using current reports has not been seen.

The prediction model of financial distress can be divided into two categories: one is mathematical statistical model. For example, Altman[25] designed the Z-score model based on accounting information. Merton[26] proposed the market structure model based on market features, and Crosbie & Bohn[27] developed the KMV model based on accounting and market features. This kind of model has simple operation and strong interpretability, but it relies on structured hard information too much and has strict requirements on data assumptions.The other one is machine learning model. Olson et al.[28], Ruibin et al.[29] and Barboza[30] compared the prediction performances of decision tree (DT), neural network (NN), support vector machine (SVM) etc., and verified the applicability of these models in financial distress prediction. This kind of model is widely used in classification because it runs fast and does not require too much assumptions.

In general, researches on financial distress prediction mainly focuses on listed firms, and the data sources are mainly accounting and market features. However, long disclosure cycle of accounting information and poor continuity of market trading information lead to that the empirical model of listed firms is not applicable in unlisted public firms to a large extent. Although some studies about financial distress prediction for listed firms begin to focus on the supplementary information in the annual report texts, they still face the problem of information lag. Therefore, we use the current report as supplementary information to forecast the financial distress for unlisted public firms. We mining effective soft information in current reports from the semantic aspect and conduct experimental research on its predictive ability. At the same time, the prediction performances of semantic features and topic features in current reports are compared.

3 Methodology

3.1 Framework

This section focuses on constructing the framework of financial distress prediction for unlisted public firms by mining semantic features in current reports. As for the framework, we try to extract and quantify the semantic features in current reports as complement of accounting information, so as to improve the performance of financial distress prediction of unlisted public firms. At the same time, in order to test whether the periodic reports of unlisted public firms can provide information value for financial distress prediction like that of listed firms, we also extract the semantic features in periodic reports accordingly.

Figure 1 shows our analysis processes, mainly including data preprocessing, prediction information extraction and prediction model construction. (1) In the data preprocessing stage, the missing value and outliers of accounting information are processed. And preprocessing of text information includes removing punctuation and numbers, word segmentation, removing stop words and sparse words. (2) In the stage of prediction information extraction, a set of selected accounting information and processed semantic features in disclosure reports are extracted as the input of the prediction models. (3) In the construction stage of prediction model, LR, CART, KNN and RF are selected to judge the validity of prediction information.

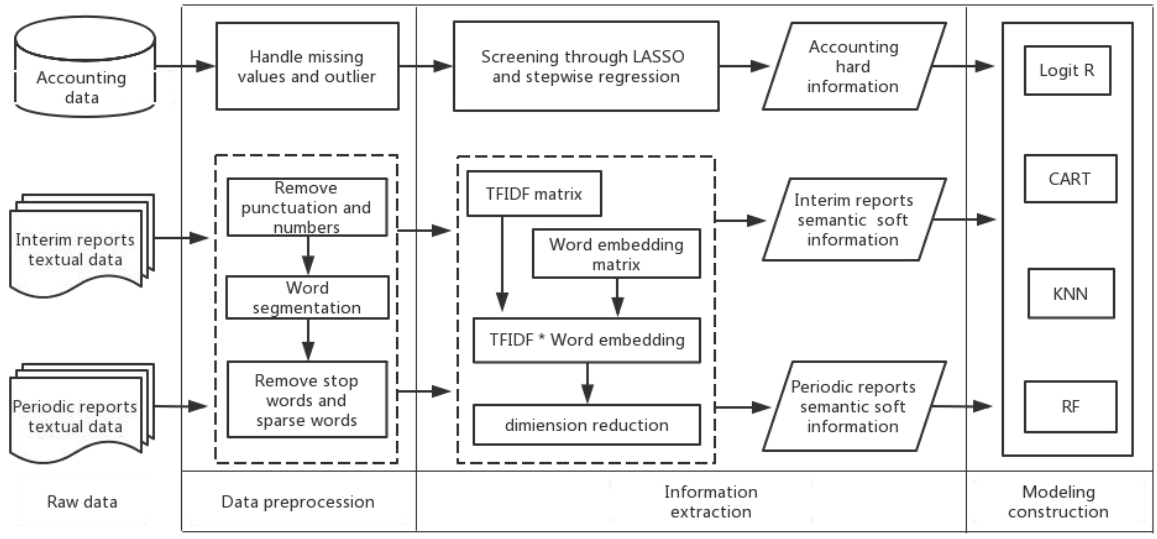

Figure1 Financial distress prediction framework for unlisted public firms

More specifically, Section 3.2 summarizes the accounting information list of unlisted public firms by reviewing relevant studies, and introduces the accounting information screening method in this paper. Section 3.3 proposes the semantic features extraction method and quantification process in this study. Section 3.4 introduces four financial distress prediction models and three evaluation criteria.

3.2 Accounting information selection

Scholars have proposed different accounting information variables as the basis of financial distress prediction, which can be roughly divided into two types. One is to select certain accounting information features purposefully based on some models[4,31]. The other one is to filter from many collected accounting information variables by summarizing previous studies[32-33]. In our research, we first compile a list of accounting information including profitability and quality of earnings, capital structure and solvency, operating capacity, growth capacity and cash flow, total 25 variables.

Then simplified and practical accounting information variables need to be screened out. Although there are many methods for feature selection, they usually need to a threshold value. In this research, we consider two different feature selection methods, namely Lasso and stepwise regression, which have the advantage of not requiring any parameters for the execution process.

3.3 Word embedding

As mentioned earlier, we need to encode unstructured text as a numerical representation to take them as inputs of models. A discrete representation model, One-hot, represents the text in binary terms with the dimension of each word being the size of vocabulary, however, it is easy to cause the problem of dimensionality curse, and words are independent from each other so that it cannot reflect sequence information. A distributed representation model, LDA topic model, is used to find representative topics from the text library and get the distribution of words on each topic, but are insufficient at forming a vector space structure for capturing semantic information. Word embedding models based on local context windows, such as Word2Vec, can learn the co-occurrence relation between words, and the word vectors are built on the premise of distributed assumption, but it is expected to create a specific vocabulary before training word vectors and lacks of dynamism.

In this study, we hope to train word vectors through a vocabulary containing almost common words to generate word embedding. We start with a recent language representation model called BERT (Devlin et al., 2018), which is a method of pre-trained text representation that can be used to extract high-quality language features from textual data to generate state-of-the-art prediction. We use the Bert model based on a 17960 Chinese token vocabulary, and set up parameters hidden-size as 768, filter-size as 4, num-filters as 256 and learning-rate as $1e^{-5}$. In the end, each word in each document are turned into 768 dimensional real-valued vectors.

3.4 Document embedding

Document vectors can usually be approximated by the arithmetic mean of word vectors, but in fact, the importance of each word in the document are different. For better representation of documents with 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 (TF-IDF) for each term in each current report document. TF-IDF is a statistical method used to assess the importance of a word to one document in a corpus. TF represents the frequency a word appears in a document. It's easy to understand, but it has an obvious disadvantage, namely, it gives higher weight to words that often appear but lack the distinguish power. IDF, as a measure of the general importance of a word, can effectively make up for the disadvantage. Given a word t and a document d, the TFIDF of t in d is calculated as:

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

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 $|\{d_j \in D : t_i \in d_j\}|$ is the number of documents where t_i appears.

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

$$V_d = tfidf_1 * v_1 + tfidf_2 * v_2 + ... + tfidf_n * v_n$$

Finally, the document embedding with 768 dimensions is generated.

3.5 Dimension reduction

Due to the limitation of our sample size, if such high-dimensional features are substituted into the prediction model, it is easy to cause overfitting problem. Therefore, we adopt a common dimension reduction technology, principal component analysis (PCA), to reduce the dimension of document embedding. As an advantage, PCA maps high-dimensional coordinates to low-dimensional coordinates, which can compress data while minimizing information loss.

The criterion for selecting a low-dimensional coordinate is to find the eigenvalues and eigenvectors of the covariance matrix, the eigenvectors representing the coordinate system, and the eigenvalues representing the lengths mapped to the new coordinates. Based on the common practice in dimension reduction, We use the

covariance ratio to retain 85% of the information, then get the dimensions indirectly. The basic formula is as follows.

$$\sqrt{\sum_{i=1}^{k}\delta_i^{\,2} \Big/ \sum_{i=1}^{n}\delta_i^{\,2}} \approx 85\%$$

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

3.6 Prediction models of financial distress and evaluation criteria of performance

Financial distress prediction can be regarded as a dichotomy problem, which is divided into financial distress firms and normal firms. In this paper, we use four common dichotomies, 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 is assumed that there is 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, then the probability threshold is set for classification. In CART model, the records with the same target attribute value are divided recursively into binary nodes to obtain binary classification results. The KNN model determines the classification of the test set 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 Bagging method. It first samples the training set according to bootstrap method, then builds CART decision tree for each sample subset, finally syntheses multiple decision trees to obtain the final classification result.

The area under ROC curve (AUC), h-measure (HM) and Kolmogorov-Smirnov (KS) are selected to judge the prediction performance of each model. AUC index reflects the comprehensive discriminability of the model to the financial distressed firms and normal firms. HM index sets the model loss of classification error based on bata distribution, which can overcome the deficiency of loss function change of AUC index. KS index is the maximum difference between the cumulative distribution of financial distressed firms and normal firms predicted by the model, which reflects the model's ability to distinguish between the two types of samples. Model effect is better with larger value of these three indexes.

4 Empirical Evaluation

4.1 Samples

In order to exclude the influence of heterogeneity of different industries, we select the second largest industry

in NEEQ market as the representative, and take all information technology service firms in 2018-2019 as the samples of financial distress prediction for unlisted public firms. All the selected samples joined in the NEEQ market before 2016. The firms with special treatment (ST) due to negative net assets at the end of last year are taken as samples of financial distress, and the remaining non-ST firms with continuous operation are taken as normal samples. Finally, the collected samples include 36 financial distress samples in 2018, 33 financial distress samples in 2019, and 1,128 normal samples in 2018-2019, and the financial distress ratio is 5.764%.

4.2 Experimental design

Considering the timeliness of current reports, we take half a year as the time period, and collect data sources including accounting data, MD&A textual data in periodic reports, and textual data in current reports between annual and semi-annual report.

We use the data in the half year before the release of Year t-1 annual report to predict whether a firm will get into financial distress in Year t. If a firm is identified as ST in Year 2019, its information of annual report in Year 2018 would be abnormal. So we take half a year as the time period, and collect the accounting information at mid-2018, the soft information of MD&A text in Year 2018 semi-annual report, and the soft information of current reports between the Year 2018 semi-annual report and the Year 2018 annual report to predict whether a firm is in financial distress in Year 2019.

Before the experiment, the data set should be divided into train set and test set, and then we can observe the prediction effects of models on the test set. Different scholars have different partitioning methods on different data sets, such as leave-one method, random proportion (e.g. 6:4, 7:3, 8:2, 9:1) , time window (e.g., data before Year 2018 as training set , data after Year 2018 as test set), and K-fold cross validation. The K-fold cross validation can repeatedly use randomly generated subsamples for training and verification at the same time, which largely avoids the limitation of insufficient or excessive training. Therefore, we make 10 times of 10-fold cross-validation based on the same data set and take the average value of 100 results as the final prediction result of each model.

4.3 Processing of accounting data

Our accounting data comes from China Stock Market & Accounting Research (CSMAR) Database and Choice Database. Its collection and screening mainly include four steps. Firstly, the original accounting information of all samples is collected as comprehensively as possible, including 25 financial features that reflect the firms' profitability, operation ability, growth ability and debt paying ability etc., and the samples and features with a deletion rate of more than 30% are deleted. Secondly, a boxplot is drawn to find and delete outliers of each feature, and then the missing values and deleted outliers are filled in with k-nearest neighbor method. Finally, the processed

features are screened by LASSO and stepwise regression, and then the screened financial features are substituted into the prediction model. By comparing the prediction results, it can be found that the stepwise regression method was superior to LASSO. The remaining 12 accounting features after screening by stepwise regression method are shown in Table 1.

Table 1 Description of accounting features

Code	Features	Description
X_1	Base earnings per share	Net profit/total equity
X_2	Capital reserve per share	Capital reserve/total equity
X_3	Undistributed profit per share	Undistributed profits/total equity
X_4	Logging of operating income	LN(operating income)
X_5	Net profit growth rate	(Current net profit/priorperiod net profit)-1
X_6	Return on equity	Net profit per share/net assets per share
X_7	Return on total asset	Pre-tax profits/average total assets
X_8	Net interest rate on sales	Net interest rate/sales revenue
X_9	Total asset turnover	Main business income/total assets
X_{10}	Turnover of accounts receivable	Net operating income/ receivables
X_{11}	Asset-liability ratio	Total liabilities/total assets
X_{12}	Current ratio	Current assets/current liabilities

4.4 Processing of textual data

We download current reports of 1197 samples in the time window from NEEQ website, the total number is 12798. Similarly, we download periodic reports and grab the part of MD&A in them. Then we carry on the stage of preprocessing the MD&A texts and current reports texts, where three steps are included. Firstly, we remove punctuation and numbers by writing the regular expressions. Secondly, we create a custom dictionary with 1398 words that have specific meanings and are indivisible. And based on it, we use jieba natural language processing tool to cut the sentences into a series of separate words. Thirdly, 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, namely, 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 6070 words by removing duplicates for all words in all documents.

After preprocessing, each word in each document is included in the word list produced finally, we then carry on the stage of extracting semantic features and four steps are included. Firstly, we construct a document-term matrix (DTM), and each entry in the DTM is the term frequency inverse document frequency (TF-IDF) for each term in each current report document. Secondly, we adopt the word embedding algorithm Bert to capture the semantic meaning from current reports, and turn terms into real-valued vectors. Thirdly, we multiply the TF-IDF of each term in each document by its word vector to get the real-valued vector of each document. Finally, we use principal component analysis (PCA) to reduce the dimension of document vectors.

5 Results and analysis

5.1 Prediction performance of semantic features in current reports

In order to test whether semantic features in current reports can improve the effect of financial distress prediction, we compare two groups of experiments. The first group contains only accounting information. The second group contains accounting information and semantic features in current reports. Features extracted from each kind of information are substituted into LR, CART, KNN and RF models respectively for 10 times of ten-fold cross validation, and the average value of 100 results is taken as the final prediction result of each model. AUC, KS and H-measure are used to judge the prediction performance of each model, and the mean value and 95% confidence interval of the three evaluation standards are shown in Table 2.

Table2 Prediction performance of semantic features in current reports

	Models	AUC		KS		H-measure	
Accounting	LR	0.8529	(0.8352,0.8705)	0.7033	(0.6735,0.7330)	0.6111	(0.5775,0.6448)
	CART	0.7207	(0.6979,0.7435)	0.4943	(0.4577,0.5309)	0.4145	(0.3777,0.4513)
	KNN	0.7596	(0.7383,0.7809)	0.5395	(0.5020,0.5769)	0.4334	(0.3965,0.4704)
	RF	0.8710	(0.8546,0.8874)	0.7105	(0.6792,0.7418)	0.6213	(0.5864,0.6563)
Accounting+	LR	0.8939	(0.8790,0.9088)	0.7450	(0.7195,0.7705)	0.6570	(0.6275,0.6866)
Current-semantic	CART	0.7876	(0.7641,0.8110)	0.5988	(0.5632,0.6345)	0.4720	(0.4320,0.5120)
	KNN	0.8295	(0.8111,0.8478)	0.6594	(0.6256,0.6933)	0.5630	(0.5281,05979)
	RF	0.9357	(0.9264,0.9449)	0.8188	(0.7972,0.8405)	0.7235	(0.6965,0.7504)

We can see from the Table2 that adding the extracted semantic features significantly improve predictive performance in terms of every performance measure over the accounting features. The AUC mean of LR, CART, KNN and RF increase by 0.041, 0.0304, 0.0699 and 0.0647 respectively, and the average increase of the four models is 0.0515. The KS mean of LR, CART, KNN and RF increase by 0.0417, 0.0708, 0.1199 and 0.1083 accordingly, and the average increase of the four models is 0.0852. And the H-measure mean of LR, CART, KNN and RF increase by 0.0459, 0.0555, 0.1296 and 0.1022 accordingly, and the average increase of the four models is 0.0833.

5.2 Comparison with semantic features in periodic reports

We also conduct one experiment to test the prediction performance of semantic features in MD&A of periodic reports, which contains accounting information and semantic features in periodic reports. The performances are shown in Table 3.

Table3 Comparison between semantic features in current reports and that in periodic reports

	Models	AUC		KS		H-measure	
Accounting	LR	0.8529	(0.8352,0.8705)	0.7033	(0.6735,0.7330)	0.6111	(0.5775,0.6448)
	CART	0.7207	(0.6979,0.7435)	0.4943	(0.4577,0.5309)	0.4145	(0.3777,0.4513)
	KNN	0.7596	(0.7383,0.7809)	0.5395	(0.5020,0.5769)	0.4334	(0.3965,0.4704)
	RF	0.8710	(0.8546,0.8874)	0.7105	(0.6792,0.7418)	0.6213	(0.5864,0.6563)
Accounting+	LR	0.8939	(0.8790,0.9088)	0.7450	(0.7195,0.7705)	0.6570	(0.6275,0.6866)
Current-semantic	CART	0.7876	(0.7641,0.8110)	0.5988	(0.5632,0.6345)	0.4720	(0.4320,0.5120)
	KNN	0.8295	(0.8111,0.8478)	0.6594	(0.6256,0.6933)	0.5630	(0.5281,05979)
	RF	0.9357	(0.9264,0.9449)	0.8188	(0.7972,0.8405)	0.7235	(0.6965,0.7504)
Accounting+	LR	0.8325	(0.8122,0.8528)	0.6767	(0.6463,0.7070)	0.5786	(0.5445,0.6127)
Periodic-semantic	CART	0.7511	(0.7261,0.7761)	0.5651	(0.5275,0.6027)	0.4700	(0.4320,0.5080)
	KNN	0.7565	(0.7371,0.7759)	0.5379	(0.5055,0.5703)	0.3721	(0.3392,0.4050)
	RF	0.8526	(0.8363,0.8688)	0.6879	(0.6607,0.7151)	0.5702	(0.5387,0.6017)

From Table3, we can see that the predictive abilities of semantic features in current reports are superior to that in periodic reports. The AUC mean of LR, CART, KNN and RF is over 0.0614, 0.0365, 0.073, and 0.0831 respectively. In addition, we found that after incorporating semantic features in period reports on the basis of accounting information, the performance only rises in CART model, while falls in LR, KNN and RF model. The

AUC value of CART increases from 0.7207 to 0.7511.

5.3 Comparison with topic features

After confirming that the semantic features in current reports indeed contributed to the improvement of prediction performance, we further compared the word embedding method with the LDA topic model. The LDA model is an unsupervised machine learning technology. It can be used to identify topic information hidden in document set or corpus. The performances are shown in Table 4.

A Table4 Comparison between semantic features and topic features

	Models	AUC		KS		H-measure	
Accounting+	LR	0.8939	(0.8790,0.9088)	0.7450	(0.7195,0.7705)	0.6570	(0.6275,0.6866)
Current-semantic	CART	0.7876	(0.7641,0.8110)	0.5988	(0.5632,0.6345)	0.4720	(0.4320,0.5120)
	KNN	0.8295	(0.8111,0.8478)	0.6594	(0.6256,0.6933)	0.5630	(0.5281,05979)
	RF	0.9357	(0.9264,0.9449)	0.8188	(0.7972,0.8405)	0.7235	(0.6965,0.7504)
Accounting+	LR	0.8711	(0.8550,0.8872)	0.7261	(0.7004,0.7519)	0.6340	(0.6040,0.6641)
Current-topic	CART	0.7255	(0.7015,0.7496)	0.5238	(0.4894,0.5582)	0.4248	(0.3894,0.4603)
	KNN	0.8491	(0.8333,0.8649)	0.6981	(0.6686,0.7276)	0.5791	(0.5473,0.6108)
	RF	0.9191	(0.9092,0.9290)	0.7750	(0.7517,0.7982)	0.6720	(0.6429,0.7012)

B

	Models	AUC		KS		H-measure	
Accounting+	LR	0.8325	(0.8122,0.8528)	0.6767	(0.6463,0.7070)	0.5786	(0.5445,0.6127)
Periodic-semantic	CART	0.7511	(0.7261,0.7761)	0.5651	(0.5275,0.6027)	0.4700	(0.4320,0.5080)
	KNN	0.7565	(0.7371,0.7759)	0.5379	(0.5055,0.5703)	0.3721	(0.3392,0.4050)
	RF	0.8526	(0.8363,0.8688)	0.6879	(0.6607,0.7151)	0.5702	(0.5387,0.6017)
Accounting+	LR	0.8309	(0.8099,0.8519)	0.6768	(0.6463,0.7072)	0.5651	(0.5306,0.5996)
Periodic-semantic	CART	0.7468	(0.7416,0.7520)	0.5524	(0.5272,0.5976)	0.4627	(0.4054,0.5200)
	KNN	0.7206	(0.7007,0.7405)	0.4931	(0.4617,0.5244)	0.3065	(0.2754,0.3377)
	RF	0.8521	(0.8452,0.8590)	0.6829	(0.6758,0.6900)	0.5666	(0.5436,0.5896)

As shown in Table4, the topic features also contribute to performance improvement over the accounting features, but the extracted semantic features led to bigger improvement. The performances of semantic features in current reports are better than the topic features in current reports for LR, CART and RF model. The AUC mean of

LR, CART and RF is over 0.0228, 0.0621 and 0.0166 respectively. And the semantic features in periodic reports show better performance than topic features in them.

5.4 Prediction performance using SMOTE

In order to avoid the possible effects of unbalanced data, we further test the robustness of the above experimental results by using SMOTE to process the imbalance samples. The performances are shown in Table 5.

Table5 Prediction performance using SMOTE

	Models	AUC		KS		H-measure	
Accounting	LR	0.8340	(0.8139,0.8542)	0.6877	(0.6578,0.7177)	0.5920	(0.5588,0.6252)
	CART	0.7918	(0.7682,0.8154)	0.6089	(0.5759,0.6420)	0.4807	(0.4447,0.5166)
	KNN	0.7697	(0.7513,0.7881)	0.5621	(0.5306,0.5937)	0.3898	(0.3558,0.4238)
	RF	0.8650	(0.8506,0.8794)	0.7108	(0.6826,0.7389)	0.5976	(0.5653,0.6299)
Accounting+	LR	0.8843	(0.8666,0.9020)	0.7493	(0.7242,0.7744)	0.6559	(0.6275,0.6842)
Current-semantic	CART	0.8141	(0.7902,0.8380)	0.6834	(0.6526,0.7141)	0.5408	(0.5070,0.5746)
	KNN	0.8960	(0.8831,0.9090)	0.7762	(0.7531,0.7994)	0.6498	(0.6230,0.6765)
	RF	0.9389	(0.9323,0.9456)	0.8331	(0.8157,0.8505)	0.7301	(0.7071,0.7532)
Accounting+	LR	0.8486	(0.8289,0.8684)	0.7186	(0.6925,0.7447)	0.6173	(0.5878,0.6468)
Current-topic	CART	0.8144	(0.7926,0.8361)	0.6756	(0.6462,0.7049)	0.5206	(0.4873,0.5539)
	KNN	0.8853	(0.8749,0.8958)	0.7697	(0.7516,0.7878)	0.6024	(0.5808,0.6240)
	RF	0.9367	(0.9293,0.9441)	0.8284	(0.8097,0.8471)	0.7220	(0.6975,0.7464)
Accounting+	LR	0.8244	(0.8039,0.8449)	0.6808	(0.6505,0.7111)	0.5837	(0.5502,0.6171)
periodic-semantic	CART	0.7903	(0.7680,0.8126)	0.6003	(0.5659,0.6347)	0.4658	(0.4281,0.5034)
	KNN	0.7967	(0.7769,0.8164)	0.5977	(0.5649,0.6306)	0.4290	(0.3947,0.4632)
	RF	0.8683	(0.8525,0.8840)	0.7222	(0.6948,0.7496)	0.6100	(0.5772,0.6428)
Accounting+	LR	0.8245	(0.8047,0.8443)	0.6579	(0.6272,0.6887)	0.5607	(0.5258,0.5955)
Periodic-semantic	CART	0.7846	(0.7613,0.8079)	0.5991	(0.5638,0.6344)	0.4584	(0.4206,0.4962)
	KNN	0.7524	(0.7336,0.7713)	0.5381	(0.5069,0.5694)	0.3519	(0.3211,0.3828)
	RF	0.8545	(0.8374,0.8716)	0.7003	(0.6716,0.7290)	0.5872	(0.5545,0.6199)

As can be seen from the Table 5. First, when incorporate semantic features in current reports on the basis of accounting information, the prediction effect is obviously improved. Second, the topic features in current reports

also contributes to performance improvement over the accounting information, but is inferior to semantic features in current reports. However, neither semantic features nor the topic features in periodic reports could improve the prediction performance. These results are consistent with the experimental results not using SMOTE.

5.5 Discussion

In the one hand, as our empirical evaluation shows the unavailability of the periodic reports in financial distress for Chinese unlisted firms, this shows a different conclusion from related research for listed companies. Meng (2017) divided MD&A in the periodic report into the review part and the prospect part and found that only the information content in the prospect part can significantly reduce the risk of bankruptcy. Observing our samples, most of China's unlisted public companies do not disclose the contents of the prospect part, which may be the main reason about useless of periodic reports.

In the other hand, our empirical evaluation also shows that both semantic features and the topic features in current reports could improve the performance of financial distress prediction for China's unlisted public firms. This could provide strong support for investors and creditors to make investment decision and avoid investment risk by combining with the soft information in current reports before making a decision.

6 Conclusion

Aiming at the problem of insufficient information and information lag in unlisted public firms, semantic features in current reports is used to predict financial distress for the first time in our research. We extract and quantify effective semantic features hidden in current reports, and construct prediction models to test predictive ability of the information. At the same time, the prediction performances are compared by using semantic features in current reports or in periodic reports. The results show that: whether or not using SMOTE, both semantic features and the topic features in current reports contribute to performance improvement over the accounting information.While neither semantic features nor the topic features in periodic reports could improve the prediction performance.

Our study still has several points that need to be improved in further study. First, the jieba word segmentation tool has some drawbacks, which may affect the experimental results. More advanced word segmentation techniques could be considered in future research. Second, this paper treats financial distress prediction as a dichotomy problem. However, in fact, there are different degrees about the financial distress. Future research needs to explore prediction models of financial distress with different degrees, such as mild, moderate and bankruptcy.

References

[1] Bushee B J, Leuz C. Economic consequences of SEC disclosure regulation: evidence from the OTC bulletin board[J]. Journal of accounting and economics, 2005, 39(2): 233-264.

[2] Zabri S M, Ahmad K, Wah K K. Corporate governance practices and firm performance: Evidence from top 100 public listed companies in Malaysia[J]. Procedia Economics and Finance, 2016, 35: 287-296.

[3] He W P. The regulation of securities markets in China[M]. Palgrave Macmillan, 2018.

[4] Altman E I, Iwanicz‐Drozdowska M, Laitinen E K, et al. Financial distress prediction in an international context: A review and empirical analysis of Altman's Z‐score model[J]. Journal of International Financial Management & Accounting, 2017, 28(2): 131-171.

[5] Doumpos M, Niklis D, Zopounidis C, et al. Combining accounting data and a structural model for predicting credit ratings: Empirical evidence from European listed firms[J]. Journal of Banking & Finance, 2015, 50: 599-607.

[6] Chen X, Wang X, Wu D D. Credit risk measurement and early warning of SMEs: An empirical study of listed SMEs in China[J]. Decision Support Systems, 2010, 49(3): 301-310.

[7] Tsai M F, Wang C J. On the risk prediction and analysis of soft information in finance reports[J]. European Journal of Operational Research, 2017, 257(1): 243-250.

[8] Cecchini M, Aytug H, Koehler G J, et al. Making words work: Using financial text as a predictor of financial events[J]. Decision Support Systems, 2010, 50(1): 164-175.

[9] Jiang C, Wang Z, Wang R, et al. Loan default prediction by combining soft information extracted from descriptive text in online peer-to-peer lending[J]. Annals of Operations Research, 2018, 266(1): 511-529.

[10] Mai F, Tian S, Lee C, et al. Deep learning models for bankruptcy prediction using textual disclosures[J]. European Journal of Operational Research, 2019, 274(2): 743-758.

[11] Davis A K, Tamasweet I. Managers' Use of Language Across Alternative Disclosure Outlets: Earnings Press Releases Versus MD&A[J]. Contemporary Accounting Research, 2011, 29(3): 804-837.

[12] Mcmullin J L, Miller B P, Twedt B J, et al. Increased mandated disclosure frequency and price formation: evidence from the 8-K expansion regulation[J]. Review of Accounting Studies, 2019, 24(1): 1-33.

[13] Van Buskirk A. Disclosure Frequency and Information Asymmetry[J]. Review of Quantitative Finance and Accounting, 2012, 38(4): 411-440.

[14] Defond M L, Hung M, Trezevant R, et al. Investor Protection and the Information Content of Annual Earnings Announcements: I nternational Evidence[J]. Journal of Accounting and Economics, 2007, 43(1): 37-67.

[15] Lerman A, Livnat J. The new Form 8-K disclosures[J]. Review of Accounting Studies, 2010, 15(4): 752-778.

[16] Wang Z, Jiang C, Zhao H, et al. Mining Semantic Soft Factors for Credit Risk Evaluation in Peer-to-Peer Lending[J]. Journal of Management Information Systems, 2020, 37(1): 282-308.

[17] Ross S A, Westerfield R W, Jaffe J F, Corporate finance, second ed.[M]. Homewood IL, 1999.

[18] Koh S K, Durand R B, Dai L, et al. Financial distress: Lifecycle and corporate restructuring[J]. Journal of Corporate Finance, 2015, 33: 19-33.

[19] Tinoco M H, Wilson N. Financial distress and bankruptcy prediction among listed companies using accounting, market and macroeconomic variables[J]. International Review of Financial Analysis, 2013, 30: 394-419.

[20] Liang D, Tsai C F, Wu H T. The effect of feature selection on financial distress prediction[J]. Knowledge-Based Systems, 2015, 73: 289-297.

[21] Sun J, Fujita H, Chen P, et al. Dynamic financial distress prediction with concept drift based on time weighting combined with Adaboost support vector machine ensemble[J]. Knowledge-Based Systems, 2017, 120: 4-14.

[22] Geng R, Bose I, Chen X. Prediction of financial distress: An empirical study of listed Chinese companies using data mining[J]. European Journal of Operational Research, 2015, 241(1): 236-247.

[23] Loughran T, Mcdonald B. When is a Liability not a Liability? Textual Analysis, Dictionaries, and 10-Ks[J]. Journal of Finance, 2011, 66(1): 35-65.

[24] Davis A K, Piger J M, Sedor L M. Beyond the numbers: Measuring the information content of earnings press release language[J]. Contemporary Accounting Research, 2012, 29(3): 845-868.

[25] Altman E I. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy[J]. The journal of finance, 1968, 23(4): 589-609.

[26] Merton R C. On the pricing of corporate debt: The risk structure of interest rates[J]. The Journal of finance, 1974, 29(2): 449-470.

[27] Crosbie P, Bohn J. Modeling default risk [R]. SAN FRANCISCO, KMV, LLC, 2003.

[28] Olson D L, Delen D, Meng Y. Comparative analysis of data mining methods for bankruptcy prediction[J]. Decision Support Systems, 2012, 52(2): 464-473.

[29] Ruibin G, Indranil B, Xi C. Prediction of financial distress: An empirical study of listed Chinese companies using data mining[J]. European Journal of Operational Research, 2015, 241(9): 236-247.

[30] Barboza F, Kimura H, Altman E. Machine learning models and bankruptcy prediction[J]. Expert Systems with Applications, 2017, 83: 405-417.

[31] Shumway T. Forecasting bankruptcy more accurately: A simple hazard model[J]. The Journal of Business, 2001,74(1):101-124.

[32] Hosaka T. Bankruptcy prediction using imaged financial ratios and convolutional neural networks[J]. Expert Systems with Applications, 2019, 117: 287-299.

[33] Liang D, Lu C C, Tsai C F, et al. Financial ratios and corporate governance indicators in bankruptcy prediction: A comprehensive study[J]. European Journal of Operational Research, 2016, 252(2): 561-572.