ACCOUNTING & FINANCE

afaanz

Accounting & Finance

Discovering bank risk factors from financial statements based on a new semi-supervised text mining algorithm

Lu Wei[a,b], Guowen Li[a,b], Xiaoqian Zhu[a], Jianping Li[a,b]

[a]*Institutes of Science and Development, Chinese Academy of Sciences,*
[b]*University of Chinese Academy of Sciences, Beijing, China*

Abstract

This paper aims to comprehensively uncover bank risk factors from qualitative textual risk disclosures reported in financial statements, which contain a huge amount of information on bank risks. We propose a new semi-supervised text mining approach named naive collision algorithm to analyse the textual risk disclosures, which can more accurately identify bank risk factors compared with the typical unsupervised text mining approach. We identified 21 bank risk factors in total, which is far more than identified in previous studies. We further analyse the importance of each bank risk factor and how the importance of each risk factor changes over time.

Key words: Bank risk; Risk factor; Financial statements; Semi-supervised; Text mining

JEL classification: C80, G21

doi: 10.1111/acfi.12453

1. Introduction

Bank risk is a central issue affecting financial stability, as proved by the subprime crisis, during which bank risks had a first-order effect on financial stability (Laeven and Levine, 2009; Allen and Faff, 2012). Thus, bank risk

We sincerely thank the editor and reviewers for their very valuable and professional comments. This research has been supported by grants from the National Natural Science Foundation of China (71601178, 71425002) and the Youth Innovation Promotion Association of the Chinese Academy of Sciences (2012137, 2017200).

Please address correspondence to Jianping Li via email: ljp@casipm.ac.cn

management is of great importance to avoid future financial crises (Zhu *et al*., 2015; Berger *et al*., 2016; Zhu *et al*., 2019).

Economic theory tells us that bank risks are affected by various risk factors (Jarrow and Turnbull, 2000). Bank risks can be described as profits and losses due to movements in risk factors (Brown, 2012). Thus, uncertainties about risk profit and loss (P&L) are related to uncertainties about risk factors, and therefore risk factors can provide a logical answer to the risk P&L (Rosenberg and Schuermann, 2006). Risk factors can capture bank risks well if they can explain most of the P&L uncertainty. If they explain only a small fraction of the risk P&L uncertainty, the residuals will have to be examined to see whether many more risk factors exist that are worth investigating and including in the risk analysis (Alexander and Pezier, 2003; Breuer *et al*., 2010). Therefore, a comprehensive selection of risk factors is of the utmost importance for explaining risk P&L (Embrechts *et al*., 2013). Omitting risk factors may lead to bias in estimating bank risks.

Many studies have made contributions to the identification of bank risk factors (Grundke, 2010). Specifically, Jarrow and Turnbull (2000), Alexander and Pezier (2003), Medova and Smith (2005), Rosenberg and Schuermann (2006), Aas *et al*. (2007), Grundke (2009, 2010), Breuer *et al*. (2010), Kretzschmar *et al*. (2010) and Bellini (2013) determined specific macroeconomic risk factors, including interest rate, credit spread, exchange rate, equity market index, gross domestic product and many more. Further, local competition (Chari and Gupta, 2008), local culture (Li and Guisinger, 1992), the degree of regulatory, monetary and legal complexity (Berger *et al*., 2004), mergers and acquisitions (Alibux, 2007; Hagendorff and Keasey, 2009), the degree of economic and political instability (Brewer and Rivoli, 1990), the extent of market imperfections and asymmetric information problems (Buch and DeLong, 2004; Gleason *et al*., 2006) are also used as risk factors in measuring bank risks.

However, the existing literature is far from reaching an agreement on the identification of bank risk factors. The bank risk factors selected by various studies differ greatly. The reason is that the bank risk factors in existing studies are identified in one of two ways. The first is selecting bank risk factors based on the researchers' subjective judgement (Rosenberg and Schuermann, 2006). The other is selecting bank risk factors by summarising the bank risk factors identified in previous studies (Breuer *et al*., 2010). However, these two methods for identifying bank risk factors can only identify a limited range of bank risk factors. Many potential risk factors that also significantly affect bank risks have not yet been identified (Grundke, 2010). Almost no prior work can fully determine how many risk factors affect bank risks. Thus, a new way to comprehensively discover bank risk factors is needed.

Recently, managers and researchers have found that the qualitative textual risk disclosures reported in financial statements are important sources of corporate risk factors (Bao and Datta, 2014; Campbell *et al*., 2014; Wei *et al*.,

2019). Beginning in 2005, the US Securities and Exchange Commission (SEC) required public companies to add a separate section 1A in their SEC Form 10-K to discuss 'the most significant factors that make the offering speculative or risky' (SEC, 2005). Researchers have provided insight into analysing corporate textual risk disclosures to discover risk factors (Huang and Li, 2011; Bao and Datta, 2014; Dyer *et al*., 2017; Miller, 2017). Hence, using a text mining approach to analyse textual risk disclosures of the banking industry is a feasible way to comprehensively identify bank risk factors. However, the previous studies have not focused on analysing the risk information contained in the textual risk disclosures reported in financial statements of the banking industry.

Several text mining approaches for analysing the textual risk disclosures reported in financial statements have been proposed. Specifically, the multi-label categorical K-nearest neighbour (ML-CKNN) proposed by Huang and Li (2011) and the dictionary used by Campbell *et al*. (2014) need to predefine a list of risk factors reported in textual risk disclosures. In most cases, however, the risk factors might be hard to derive beforehand. Some important risk factor types may be left out from the predefined list of risk factors. Thus, it is hard for these text mining approaches to comprehensively identify all the risk factors. Bao and Datta (2014) proposed an unsupervised Sentence Latent Dirichlet Allocation (Sent-LDA) approach to comprehensively discover risk factors. However, the unsupervised algorithm makes researchers stay away from data, which may lead to biased classification results without appropriate human interpretation and adjustment (Miller, 2017).

Therefore, this paper aims to comprehensively and accurately identify bank risk factors from qualitative textual risk disclosures reported in financial statements. Compared with identifying bank risk factors based on researchers' subjective judgements or summarising of existing bank risk factors in previous studies, the textual risk disclosures that contain a huge amount of bank risk information can help us to realise the comprehensive identification of bank risk factors. A new semi-supervised text mining approach called the 'naive collision algorithm' is proposed to analyse the textual risk disclosures, which can more accurately identify bank risk factors than the typical unsupervised text mining approach. In the experiment, the proposed naive collision algorithm is used to identifying bank risk factors based on 59,418 headings of textual risk disclosures from 2,189 SEC Form 10-K filings of US commercial banks over the period 2010–2016. In addition, the importance of each risk factor and the annual change in importance of each risk factor are further analysed. Finally, the proposed semi-supervised naive collision algorithm and a typical unsupervised text mining approach are empirically compared in terms of accuracy.

The remainder of this paper is organised as follows. Section 2 introduces the proposed semi-supervised naive collision algorithm. Section 3 discusses the bank risk factors discovered. Section 4 concludes this paper.

2. The proposed semi-supervised naive collision algorithm

In this section, we introduce our proposed semi-supervised text mining algorithm, called the naive collision algorithm, for classifying textual risk disclosures. We observe that almost all risk factor disclosures consist of a summary heading and detailed explanations. However, text mining approaches applied both to headings and descriptions led to a worse performance in prior studies (Huang and Li, 2011). Thus, our proposed algorithm is designed for classifying headings of textual risk disclosures. Furthermore, in accordance with Bao and Datta (2014), we assume that all words in a heading are sampled from the same topic.

Based on the concept of the rule-based system (Bengio *et al.*, 2015), we develop our naive collision algorithm. Figure 1 gives the graphical representation of the proposed algorithm.

Specifically, the inputs of our proposed algorithm are the training set and test set containing feature vectors of textual risk factor headings. Since the contents of Form 10-K textual disclosures tend to be boilerplate in nature, the same information is presented by similar sentences (Dyer *et al.*, 2017). Thus, we apply the vector space model (VSM) to construct feature vectors by quantifying headings of textual risk disclosures, which can make headings with the same information have higher similarities (Turney and Pantel, 2010). In addition, by analysing the textual risk disclosures, we found that the meaning of a heading is expressed mainly by using nouns, so we select all nouns in a heading to form the feature vector. To implement our algorithm, feature vectors of textual risk factor headings need to be divided into a training set and several test sets. Here, we assume that the feature vectors are divided into one training set and n test sets to illustrate our algorithm.

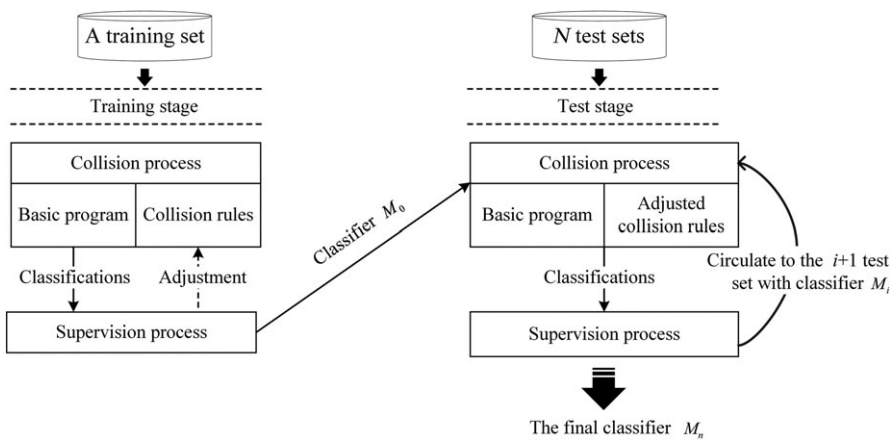

Figure 1 Graphical model of the proposed semi-supervised naive collision algorithm.

Having determined the training set and the test sets, we input them into the naive collision algorithm. It is clear from Figure 1 that the algorithm is divided into two stages and every stage is composed of two processes. Specifically, the collision process automatically classifies risk factors. The supervision process adjusts the automatic classifications with appropriate human interpretation. Thus, our proposed semi-supervised algorithm is characterised by combing the automatic collision process and the human supervision process. In the following, we describe the procedure of our proposed algorithm in detail.

First, in the training stage, the training set is inputted into the collision process to obtain the original automatic classifications via the basic program, under which different feature vectors are merged into the same classification when the similarities between them are beyond the set threshold. The specific pseudocode of the basic program is recorded in Table 1. Then, the original automatic classifications are analysed in the following supervision process. The supervision process consists of two subprocesses of supervision. The aim of the first subprocess of supervision is to adjust the collision rules to make the automatic classifications more accurate and convergent though expert knowledge. Then, we combine the adjusted collision rules with the basic program in

Table 1
The specific pseudo-code of the basic program in the collision process

Algorithm: the basic program of the collision process		
1:	**procedure** TAKE (v_0)	// Initialize the classifier, M
2:	$m_{v_0} \leftarrow \{v_0\}$	
3:	$M \leftarrow \{m_{v_0}\}$	
4:	**end procedure**	
5:	**procedure** COLLISION (V, M, ε)	// Input data set V, the initialized classifier M and
6:	**for** $v_i \in V$ **do**	Set the threshold value ε to collide vectors
7:	**for** m $\in M$ **do**	into the classifier
8:	**for** $v_j \in m$ **do**	
9:	$similarity_{i,j} \leftarrow \cos(v_i, v_j)$	// Compute the similarity of the vector and vectors
10:	**end for**	In the classifier
11:	**end for**	
12:	$similarity_{i,p} \leftarrow \max(similarity_{i,j})$	// Select the maximal similarity
13:	**if** $similarity_{i,p} \geq \varepsilon$ **then**	// Cluster according to the comparison of the
14:	$m_{v_p} \leftarrow m_{v_p} \cup \{v_i\}$	maximal similarity and the threshold value
15:	**else**	
16:	$m_{v_i} \leftarrow \{v_i\}$	
17:	**end if**	
18:	**end for**	
19:	**end procedure**	

the collision process to update the original automatic classifications. We repeat the first subprocess of supervision until the most suitable adjusted rules are developed. Having determined adjusted collision rules, we can get the most accurate and convergent automatic classifications by combing the adjusted collision rules and the basic program in the collision process.

The second subprocess of supervision is to merge the most accurate and convergent automatic classifications with the same topic. The core subprocess of supervision is labelling each classification to reflect its topic based on expert judgements. Overall, in the training stage, we obtain an original classifier, M_0, which contains the risk factor classification results and their corresponding feature vectors.

The original classifier M_0 obtained from the training stage is used as the input in the test stage. Feature vectors contained in the first test set are classified based on the classifier M_0 through the collision process by combing the basic program and adjusted collision rules. Then, the derived classifications are merged to arrive at the appropriate classifier M_1 through the supervision process. The enhanced classifier M_1 is further applied to the next test set to classify feature vectors. In this way, the classifier M_i is used as the input of the $i + 1$ test set to obtain the classifier M_{i+1} We repeat the collision and supervision processes n times in the test stage until all feature vectors contained in n text sets are gradually classified. Finally, we obtain the most enhanced classifier M_n, which provides the final risk factor classification results and the corresponding feature vector of each risk factor type.

In summary, training and test sets containing feature vectors obtained by VSM are used as inputs to our proposed naive collision algorithm. Following a training stage and a text stage, we eventually obtain the output of risk factor classifications. The characteristic of semi-supervision is reflected by an automated collision process and a human supervision process in each stage. Compared with the unsupervised Sent-LDA text analysis approach, it provides more accurate classifiers in identifying risk factors.

3. Empirical analysis

In this section, the proposed naive collision algorithm is implemented to discover bank risk factors by analysing the textual risk disclosures reported in section 1A of US commercial banks' SEC Form 10-K filings. We first introduce our data source and check our data quality. Then, we design the adjusted collision rules according to the characteristics of US commercial banks' textual risk disclosures and verify the convergence of the algorithm. Third, we discuss the classification results of bank risk factors and draw conclusions. Finally, we compare our proposed semi-supervised naive collision algorithm with the competing unsupervised text mining model in terms of predictive power and clustering quality.

3.1. Data preparation

Although the SEC required all filers to discuss company risk factors in one additional section (section 1A) in their Form 10-K in 2005 (SEC, 2005), the effectiveness of risk factor disclosures had been widely questioned. In 2010, the SEC issued comment letters asking filers to only include specific risk factors related to the individual company to enhance effectiveness (SEC, 2010; Zhu *et al.*, 2016). Thus, to obtain the more effective textual risk disclosures, we extract the summary headings of textual risk disclosures reported in section 1A of Form 10-K filings over the period 2010–2016 to prepare our data set.

To analyse bank risk factors, US commercial banks are selected based on an industrial classification (SIC) code list on the SEC's website (https://www.sec.gov/info/edgar/siccodes.htm). The SICs related to commercial banks are 6021, 6022 and 6029, whose corresponding industry names are national commercial banks, state commercial banks and commercial banks, NEC, respectively. Then, we can collect the names of US commercial banks and their corresponding Form 10-K filings from the EDGAR databases on the SEC's website.

To extract headings from textual risk disclosures and structure them as Excel files, we set rules based on the characteristics that textual risk factor headings are usually marked in italics, boldface or underlines in the Form 10-K. To verify the accuracy of data extraction based on our rules, we manually checked the structural headings. The results show that only 64 out of 3,000 headings are misextracted and the accuracy rate is as high as 97.87 percent, which indicates the robustness of our extraction rules. Eventually, we obtained 59,418 headings of textual risk disclosures from 2,189 Form 10-K filings of US commercial banks (27.14 headings per Form 10-K on average) over the period of 2010–2016.

3.2. Implementation of the proposed algorithm

Having collected textual risk factor headings, we use VSM to quantify them to form feature vectors. To identify annual bank risk factors, the feature vectors are divided into seven sets by year, which allows us to analyse the annual change in bank risk factors over the sample period. Specifically, feature vectors from 2010 are put into a training set. Six test sets contain feature vectors from 2011 to 2016.

After putting the training set into the training stage of our proposed algorithm, we set a higher threshold value of similarity $\varepsilon = 0.71$ under the basic program. The reasons why we set such a threshold are as follows. First, from the perspective of theory, a threshold of similarity larger than $\sqrt{2}/2$ can avoid the problem of over-clustering. For example, α, β, γ are feature vectors made up of nouns. w_1, w_2 are nouns that made up these three feature vectors. The w_1, w_2 numbers represent the number of times w_1, w_2 appeared in a heading of textual

risk disclosures, respectively. When $\alpha = (w_1 = 1, w_2 = 1)$, $\beta = (w_1 = 1, w_2 = 0)$ and $\gamma = (w_1 = 0, w_2 = 0)$, we can calculate that similarity(α, β) = $\sqrt{2}/2$, similarity(α, γ) = $\sqrt{2}/2$, and similarity(β, γ) = 0 based on the formula of similarity calculation. If the threshold of similarity is smaller than $\sqrt{2}/2$, α and β will be classified together, and α and γ will be classified together. Hence, α, β, γ will be classified into one classification, which leads to the problem that β and γ with a similarity of 0 are also classified into one classification. Thus, the threshold of similarity should be larger than $\sqrt{2}/2$ to avoid the problem of over-clustering ($\sqrt{2}/2 \approx 0.707$).

Second, from an empirical point of view, many pre-experiments were performed to determine the threshold of similarity. As we know, the higher the threshold of similarity, the higher the accuracy of automatic classification results. However, more human efforts are needed to merge a large number of automatic classification results (Turney and Pantel, 2010). Thus, the setting of the similarity threshold should balance the classification accuracy and human efforts. The results of pre-experiments show that by taking 0.71 as the threshold of similarity, we can obtain a relatively higher accuracy of automatic classification results, and the workload of the manual merging of classification results is also acceptable. Thus, the similarity threshold set by us is 0.71, which is larger than $\sqrt{2}/2$ ($\sqrt{2}/2 \approx 0.707$) and achieves a better balance between classification accuracy and human work.

Setting the similarity threshold at 0.71, we obtain an original automatic classifier with 2,491 bank risk factor classifications through the collision process. Then, in the first subsupervision process, we check over feature vectors of each classification contained in the automatic classifier to set adjusted collision rules to improve the accuracy of the classifications. By repeating the above collision process and the first subsupervision process three times, we found four problems, for each of which we set a corresponding collision rule to address the problem. The four adjusted collision rules are summarised in Table 2.

First, we found that some common words contained in the feature vectors, such as *risk* and *company*, are not representative of bank risks. Additionally, the company names and some special symbols are also not representative of bank risks. Thus, we record these words, including company names, special symbols and common words in a stop-words file and extract them from the feature vectors. The corresponding rule is rule 'a' in Table 2, whose aim is to extract words that are not representative of bank risks from the inputted feature vectors.

Second, for the automatic classifications that are not satisfactory, we check the original summary headings of textual risk disclosures. We find that verbs and adjectives in these headings are informative for explaining the meanings of headings. However, as discussed in Section 2, we select all nouns of a heading to form the feature vector because the meaning of a heading is expressed mainly using nouns by analysing the headings of textual risk disclosures. Thus, the informative verbs and adjectives are extracted from feature vectors, which lead to worse automatic classifications. To overcome this problem, we transfer

Table 2
The adjusted collision rules for improving the accuracy of classifications

Adjusted collision rules
a. Write down the names of companies, some special symbols and common words into the stop-words file. Common words are words appearing in many headings while are not representative of bank risk factors, such as *risk*, *company*, and so on.
b. Transfer some informative verbs and adjectives (e.g. compete and technological) into corresponding nouns.
c. To generate more convergent classification results, if the dimension of the feature vector v_i is equal to 2 or 3, reduce the dimension of v_j to the same dimension as v_i when computing similarity$_{i,j}$, where j varies from 1 to $i-1$.
d. Due to setting the rule c, the higher dimension v_i is capable of capturing lower dimension vectors. The misextracted text almost has a higher dimension, which leads to less accuracy of classification results. To address this problem, if the dimension of v_i is greater than or equal to 9, regard it as dirty data and drop it.

informative verbs and adjectives (e.g., compete and technological) into corresponding nouns (rule 'b'). By doing this, we can obtain better automatic classification results.

Third, by analysing the feature vectors, we find that in most cases, the number of nouns contained in each feature vector ranges from two to eight. For the feature vectors that contain two or three nouns, it is almost impossible to merge them together with other feature vectors that contain more than three nouns according to the basic program. This may lead to a worse convergence of automatic classification results with too many classifications. Thus, to generate more convergent classification results, we set rule 'c', according to which if the dimension of the feature vector v_i is equal to two or three, reduce the dimension of v_j to the same dimension of v_i when computing similarity$_{i,j}$, where j varies from 1 to $i-1$.

Finally, we set rule 'd' to address the problem of the ability of misextracted text to capture headings of textual risk disclosures being too strong. In Section 3.1, we show that there are minimal proportions of text extracted from the Form 10-K filings that are misextracted, which is quite common when extracting specific text from a file. For example, Bao and Datta (2014) had some misextracted content. By analysing the misextracted text, we find that the misextracted text is usually a paragraph or a long sentence, so the corresponding feature vector has many nouns. Due to rule 'c', the higher dimension of the feature vector v_j can capture feature vectors with lower dimensions. Thus, many headings of textual risk disclosures may be clustered together with the misextracted text. Hence, to guarantee the accuracy of the classification, if the dimension of v_j is greater than or equal to nine, v_j is probably formed by the misextracted text, so we regard it as dirty data and drop it.

Based on the basic program and the adjusted collision rules in the collision process, we obtain the original classifier M_0 with 1,796 bank risk factor

classifications and their corresponding feature vectors. Then, in the second subsupervision process, we manually label classification results using the human experts' domain knowledge. There are some automatic labelling methods (Mei *et al*., 2007), but these are not suitable in cases where such labelling requires domain knowledge. In fact, in most topic model research, it is customary to manually label topics, ensuring high labelling quality (Chang *et al*., 2009). Huang and Li (2011) design a manual labelling procedure that makes use of human experts' domain knowledge. Bao and Datta (2014) adopt the manual labelling procedure designed by Huang and Li (2011) to label risk factors. Thus, here we also use the manual labelling procedure designed by Huang and Li (2011) to give automatic classification results meaningful names.

Specifically, four persons on our research team undertook the work of labelling these 1,796 automatic classifications. Their research field is bank risk management; therefore, they have domain knowledge for giving meaningful label names to categorise the corresponding classification results. Before labelling, they are trained on a number of real labelled examples of each risk factor. Each person labels 898 out of 1,796 classifications and each classification is labelled by two persons. Thus, based on the four-person domain knowledge of bank risk, we manually label the 1,796 classifications with meaningful names. Furthermore, during the process of manual labelling, we found that the classification results are representative of risk factors and are thus easier to label. For each half of the 1,796 classifications, the two persons reach a relatively high degree of agreement on labels, which suggests that the classification results obtained through the collision process of our proposed naive collision algorithm are satisfactory. Finally, after merging classifications with the same label, the 1,796 original automatic classifications are merged into 20 types of bank risk factors with summary labels.

In the test stage, holding the similarity threshold value constant, we use the original classifier M_0 obtained from the preceding training stage to classify feature vectors contained in the following year 2011 test set. Through the collision process and supervision process, we obtain an enhanced classifier M_1. We then apply M_1 to the 2012 test set to obtain a more enhanced classifier M_2. We repeat the above steps until the last 2016 test set is classified to generate the most enhanced final classifier M_6, which contains the final classification result of 21 bank risk factors.

Obviously, as the classifier is gradually enhanced with a growing number of feature vectors included in each classification, the classification results of seven data sets (including one training set and six test sets) via the automatic collision process become increasingly convergent with fewer and fewer automatic classifications. At the same time, the automatic collision process needs a longer running time. From Figure 2, it is clear that the number of classifications through the collision process experienced a downward tendency by decreasing from 1,796 to 160. In contrast, the time required for the automatic classification increased from 1.6 to 6.9 hours.

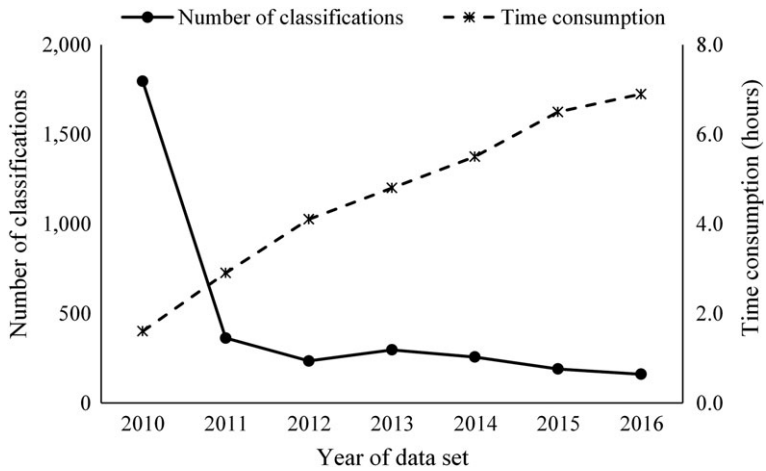

Figure 2 Evolution of classification results and time consumption in the collision process.

3.3. *Results of the identification of bank risk factors*

As presented in Section 3.2, we discover 21 bank risk factors by implementing our proposed naive collision algorithm, including the regulation, strategy, management operation, macroeconomic factors, loan loss, asset value fluctuation, capital availability, financial expense, competition, development of borrowers, reputation, product and service, merger and acquisition, accounting standard, financial institutions interaction, political environment, third-party cooperation, liability obligation, disaster, country credit rating and other risk factors. These 21 bank risk factors are visualised using word clouds in Figure 3, where the font size corresponds to the probability of the word occurring in the risk factor type. For each type of bank risk factor, we give a brief description in Table 3 along with a specific example.

Several previous studies also analysed risk factor disclosures to classify risk factors reported in Form 10-K. They use Form 10-K filings from all industries, and thus the identified risk factors are general and exposed by various industries. Unlike previous studies, we just collect risk factor disclosures from commercial banks' Form 10-K filings to discover bank risk factors. Thus, here we make a comparison between our classification results for bank risk factors and general risk factors identified in other studies (summarised in Table 4) to identify which risk factors are unique to the banking industry.

From Table 4, we can see that Huang and Li (2011), Mirakur (2011), Bao and Datta (2014) and Campbell *et al.* (2014) have found various general risk factors and labelled them based on their judgements. Risk factors with the same content may be labelled with different names in different studies. Thus, we compare the bank risk factors discovered by us with general risk factors

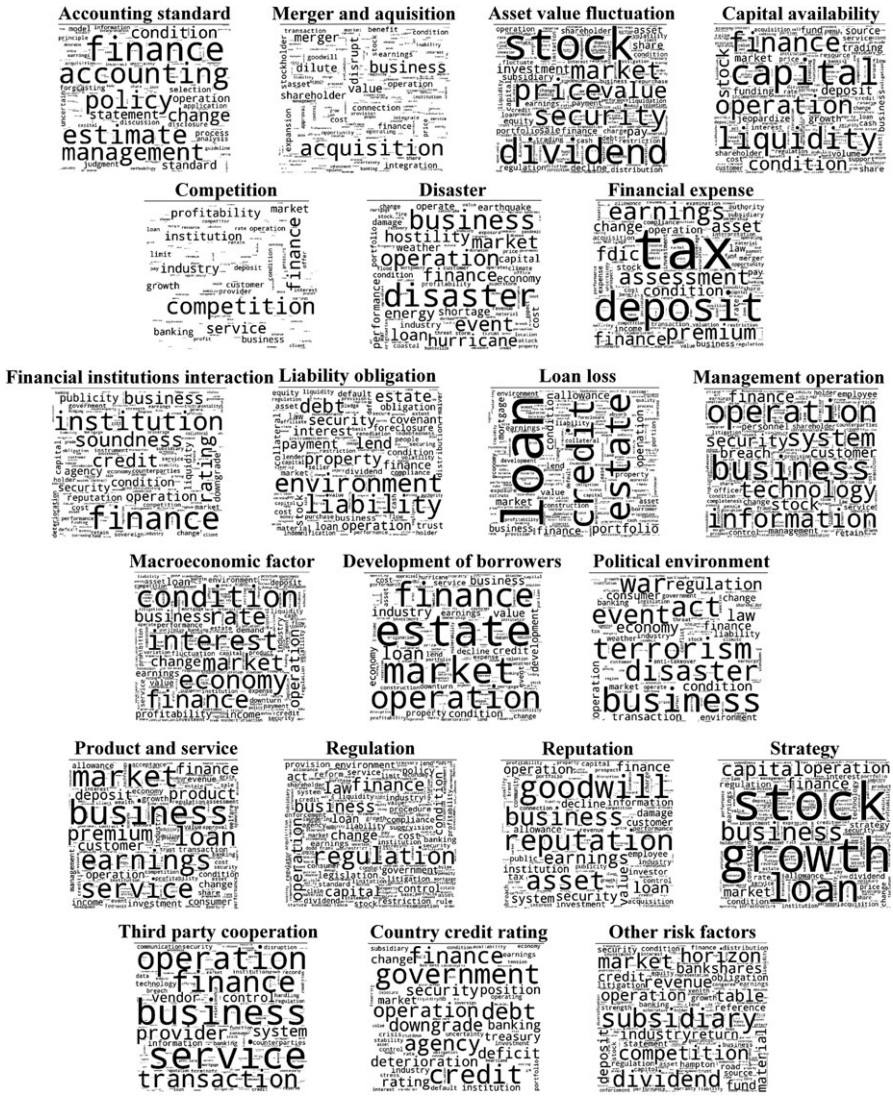

Figure 3 The word clouds of 21 bank risk factors identified based on our proposed algorithm.

identified by other researchers in terms of content. We find that 17 of the 21 identified bank risk factors are general and can affect the risks in various industries. Only four risk factors are unique to the banking industry. In other words, four risk factors identified by us, including financial institutions' interaction, loan loss, development of borrowers and country credit rating, are only experienced by commercial banks.

Table 3
Risk factor categorisations and definitions

	Risk factors	Definitions and examples
1	Accounting standard	The risk factor that relates to the requirements of accounting standards and the change in them. For example, 'Changes in accounting standards could materially impact the Corporation's financial statements.'
2	Merger and acquisition	Risks arise from the process of merger and acquisition. For example, 'We will face risks with respect to any future expansion and acquisitions or mergers.'
3	Asset value fluctuation	Potential losses resulting from asset value fluctuations in market investment activities. For example, 'Declines in asset values may result in impairment charges and may adversely affect the value of the Company's investments, financial performance, and capital.'
4	Capital availability	Risk factor related to the availability of support liquidity through funding or assets liquidity with a fair value. For example, 'Capital resources and liquidity are essential to our businesses and could be negatively impacted by disruptions in our ability to access other sources of funding.'
5	Competition	Fierce competition among commercial banks affects banks' profitability and further affects banks' risk-taking behaviours. For example, 'The corporation operates in a highly competitive industry and market area.'
6	Disaster	Huge damage caused by disasters, such as hurricanes and earthquakes. For example, 'Severe weather, natural disasters or other climate change-related matters could significantly impact our business.'
7	Financial expense	Financial expenses influence bank profitability, including tax payment, deposit premium expense to the Federal Deposit Insurance Corporation (FDIC) and so on. For example, 'Increases in FDIC insurance premiums may have a material adverse effect on our results of operations.'
8	Financial institutions interaction	Business interactions between banks and other financial industries make risk contagion possible. For example, 'We may be adversely affected by the soundness of other financial institutions.'
9	Liability obligation	Risk factor related to various liability obligations taken by banks, appearing as environment obligation, debt, payment and so on.

(continued)

Table 3 (continued)

	Risk factors	Definitions and examples
		For example, 'Our obligation to make interest payments to the trust on the debentures is subordinated to existing liabilities or additional debt we may incur.'
10	Loan loss	Potential credit losses related to the credit qualities and default of borrowers. For example, 'Our consumer loans generally have a higher degree of risk of default than our other loans.'
11	Management operation	This includes key factors that are important for daily robust operation, such as personnel, system, information, internal control, technology, litigation, agency problem, etc. For example, 'We rely heavily on our management team, and the unexpected loss of key management may adversely affect our operations.'
12	Macroeconomic factor	Changes in macroeconomic conditions, such as interest rate, equity index or foreign exchange have a huge impact on banks' activities and therefore may lead to potential risk losses. For example, 'Our results of operations may be adversely affected by changes in national and/or local economic conditions.'
13	Development of borrowers	Developments in other industries, such as energy, car, agriculture and real estate have major implications for the banks' deposit, loan and investment business. For example, 'Real estate lending in our core Texas markets involves risks related to a decline in value of commercial and residential real estate.'
14	Political environment	The stability of the political environment is important for a bank's operations. Terrorism and war will lead to huge damage. For example, 'Acts or threats of terrorism and political or military actions by the United States or other governments could adversely affect general economic industry conditions.'
15	Product and service	Popular financial products and services provided by banks are sources of profit. For example, 'New lines of business or new products and services may subject us to additional risks. A failure to successfully manage these risks may have a material adverse effect on our business.'
16	Regulation	Regulatory laws, policies, acts affect the risk-taking behaviour of banks. For example, 'The company operates in a highly regulated environment and may be adversely impacted by changes in laws and regulations.'
17	Reputation	A good reputation is helpful for profitability while a bad reputation puts a bank at high risk. For example, 'An impairment in the value of our goodwill could negatively impact our earnings and capital.'
18	Strategy	Decisions on a bank's future growth, including expansion, market share, internationalisation, etc. For example, 'Our growth strategy includes risks that could have an adverse effect on financial performance.'

(continued)

Table 3 (continued)

	Risk factors	Definitions and examples
19	Third-party cooperation	Risk factor associated with the process of cooperation with peers, upstream enterprises and downstream enterprises. For example, 'The Corporation is subject to risk from the failure of third-party vendors.'
20	Country credit rating	The government credit rating affects financial markets and therefore affects ban risks. For example, 'Recent and/or future US credit downgrades or changes in outlook by major credit rating agencies may have an adverse effect on financial markets, including financial institutions and the financial industry.'
21	Other risk factors	Risk factors that cannot be categorised into any of the other 20 risk factor types are labelled as 'other risk factors'. For example, 'Risk factors that may affect future results.'

Table 4
The summary of risk factor classifications in four previous studies

	Studies			
Risk factor	Huang and Li (2011)	Mirakur (2011)	Campbell *et al.* (2014)	Bao and Datta (2014)
1	Financial condition risks	Accounting	Financial	Human resources
2	Restructure risks	Acquisitions	Other idiosyncratic	Intellectual property licensing
3	Funding risks	Calamities	Legal and regulatory	Product defects lawsuits
4	Merger & acquisition risk	Capital expenditures	Other systematic	Regulation changes
5	Regulation risks	Capital structure	Tax	Catastrophes input prices
6	Catastrophes risks	Cash		Volatile stock price
7	Shareholder's interest risks	Competition		Shareholder's interest
8	Macroeconomic risks	Contracts		Macroeconomic cyclical industry
9	International risks	Credit risk		Cost risks
10	Intellectual property risks	Customer concentration		Rely on large customers
11	Potential defects in products	Distribution		Competition
12	Potential/ongoing lawsuits	Government		Volatile stock price
13	Infrastructure risks	Industry		Debt risks
14	Disruption of operations	Insurance		Funding
15	Human resource risks	Intellectual property		Financial condition risks
16	Licensing related risks	International		Property
17	Suppliers risks	Inventory		Investment
18	Input prices risks	Investments		Regulation changes
19	Rely on few large customers	Key personnel		Tax risks
20	Competition risks	Labor		International risks

(continued)

Table 4 (continued)

	Studies			
Risk factor	Huang and Li (2011)	Mirakur (2011)	Campbell *et al.* (2014)	Bao and Datta (2014)
21	Industry is cyclical	Legal		Credit risks
22	Volatile demand and results	Macro		Volatile demand product introduction
23	Volatile stock price risks	Marketing		Suppliers
24	New product introduction risks	Operations		Accounting risks
25	Downstream risks	Regional		Production introduction
26		Solvency		Downstream
27		Stock price		Infrastructure
28		Suppliers		Credit risks
29		Takeovers		Acquisition restructure
30				Infrastructure operation disruption

3.4. Analysis of bank risk factors

Having discovered 21 bank risk factors, we further analyse which risk factors have stronger effects on bank risks. In estimating bank risks, selecting several important risk factors instead of putting all risk factors into models will make bank risk measurement more efficient.

3.4.1. Top most important bank risk factors

The more frequent the disclosure of a risk factor, the more textual risk factor headings are classified into a risk factor, and the more attention paid by commercial banks to this risk factor, which further indicates that this risk factor is more important for commercial banks. Thus, the importance ratio, which is used to measure the importance of each risk factor, is calculated based on the principle that the annual disclosure frequency of each risk factor can reflect the importance of the risk factor. The annual disclosure frequency of a risk factor is the number of headings classified into the risk factor of each sample year. Thus, we calculate the importance ratio by dividing the number of textual risk factor headings classified into a risk factor by the total number of textual risk factor headings. Formally,

$$\text{Important ratio}_{i,t} = \frac{F_{i,t}}{TF_t} = \frac{F_{i,t}}{\sum_{t=1}^{21} F_{i,t}}, \tag{1}$$

where Important ratio$_{i,t}$ and $F_{i,t}$ stand for the importance ratio and the disclosure frequency of bank risk factor i at year t, respectively. TF_t represents the total disclosure frequency of all 21 bank risk factors at year t, which is the sum of annual disclosure frequencies of 21 bank risk factors at year t.

In addition to the annual importance ratio, we can also calculate the total importance ratio of a risk factor during a period, which represents the importance of the risk factor during the sample period. The total importance ratio is written as

$$\text{Total important ratio}_i = \frac{\sum_{t=1}^{T} F_{i,t}}{\sum_{t=1}^{T} TF} = \frac{\sum_{t=1}^{T} F_{i,t}}{\sum_{t=1}^{T}\sum_{i=1}^{21} F_{i,t}}, \tag{2}$$

where Total important ratio$_i$ stands for the total importance ratio of the risk factor i during the sample period T.

Overall, the importance ratio is used to measure the importance of each risk factor. The larger value of the importance ratio corresponds to the higher level

of importance of the risk factor and vice versa. The calculated total importance ratios of these 21 bank risk factors over the sample period 2010–2016 are summarised in Table 5.

We sort the 21 identified bank risk factors from high to low values of importance ratios in Table 5. To select the top most important bank risk factors, we take the cumulative importance ratio of 80 percent as the boundary. It is clear that the cumulative importance ratio of the top 8 of the 21 bank risk factors, including regulation (16.64 percent), strategy (14.89 percent), management operation (13.02 percent), macroeconomic factor (10.91 percent), loan loss (9.38 percent), asset value fluctuation (7.22 percent), capital availability (6.01 percent) and financial expense (3.21 percent), account for over 80 percent (81.28 percent), which indicates that these eight bank risk factors are important for bank risk measurement. The cumulative importance of the remaining 13 bank risk factors only accounts for <20 percent. In other words, although the number of important risk factors accounts for only 38.10 percent of all risk factors, their cumulative importance is up to 81.28 percent. Using these top 8 most important risk factors in bank risk modelling can explain the majority of risk P&L, which makes bank risk measurement more efficient. Of course, researchers can set the boundary of the cumulative importance ratio based on their own research needs to select important risk factors for measuring bank risks.

Table 5
Importance ratios of 21 bank risk factors based on all samples from 2010 to 2016

Risk factor	Importance ratio (%)	Cumulative importance ratio (%)
Regulation	16.64	16.64
Strategy	14.89	31.53
Management operation	13.02	44.55
Macroeconomic factor	10.91	55.46
Loan loss	9.38	64.84
Asset value fluctuation	7.22	72.06
Capital availability	6.01	78.07
Financial expense	3.21	81.29
Competition	2.73	84.02
Development of borrowers	2.71	86.72
Reputation	2.32	89.04
Product and service	1.93	90.97
Merger and acquisition	1.58	92.54
Accounting standard	1.40	93.95
Other risk factors	1.39	95.33
Financial institutions interaction	1.38	96.71
Political environment	1.23	97.94
Third-party cooperation	1.06	99.00
Liability obligation	0.48	99.48
Disaster	0.36	99.84
Country credit rating	0.16	100.00

Moreover, from the top 8 most important risk factors, we can see that the top 3, i.e. regulation, strategy and management operation, whose cumulative importance (44.55 percent) is approaching 50 percent, are risk factors affecting non-financial risks. The risk factors affecting financial risks (i.e. credit and market risks), including macroeconomic factor, loan loss, asset value fluctuation and capital availability, are less important. However, researchers and regulators take credit and market risks, which essentially are financial risks, as the two main risk types faced by banks (Basel Committee on Banking Supervision, 2006; Li *et al.*, 2015, 2018). Thus, by analysing the textual risk disclosures, we found that bankers are aware of the importance of non-financial risks. In future, more attention should be paid to the management of non-financial risks.

Furthermore, we want to study whether the important risk factors for each year are the same. The annual importance ratio can reflect the annual importance of a risk factor. Thus, according to Equation (1), we calculate the annual importance ratios of 21 bank risk factors from 2010 to 2016. We still take the cumulative importance ratio of 80 percent as the boundary to select the most important bank risk factors. The annual top 8 most important risk factors from 2010 to 2016 are summarised in Table 6 for comparison.

From Table 6, it is clear that regulation, management operation, macroeconomic factors, loan loss, asset value fluctuation and capital availability were always important risk factors for commercial banks from 2010 to 2016. It is noteworthy that the importance of strategy increased from outside of the top 8 to rank 6, rank 3 and rank 1, which indicates that strategy is increasingly important for commercial banks' risk profiles from 2010 to 2016. In contrast, the development of borrowers became less important and fell out of the top 8 except in 2010. Additionally, from 2010 to 2014, the risk factor type of financial expense ranked approximately 8th. However, in 2015 and 2016 it was replaced by reputation, which suggests that the importance of reputation gradually outperformed that of financial expense.

3.4.2. Change in the importance of bank risk factors

By analysing whether the important risk factors for each year are the same, we find that the annual importance of risk factors changed over the sample period. Thus, we give a further analysis of the change of importance of each risk factor type over the period 2010–2016. As discussed in Section 3.4.1, the number of risk factor headings classified into a risk factor can reflect the change in attention paid by commercial banks to the risk factor. Thus, we analyse the change in the importance of each risk factor by comparing the number of risk factor headings classified into a risk factor from 2010 to 2016.

Based on the number of risk factor headings classified into a risk factor summarised in Table 7, we can determine whether some risk factors appeared or disappeared in later years. If the number of headings is zero, this indicates

Table 6
Annual rankings of top 8 most important bank risk factors from 2010 to 2016

	2010	2011	2012	2013	2014	2015	2016
Regulation (%)	1 (15.03)	1 (18.44)	1 (16.54)	2 (16.44)	2 (17.85)	2 (16.05)	2 (16.12)
Management operation (%)	2 (13.88)	2 (13.76)	2 (13.31)	3 (13.03)	3 (12.55)	3 (12.47)	3 (11.91)
Macroeconomic factor (%)	3 (13.67)	3 (11.82)	4 (11.19)	4 (10.31)	4 (9.46)	4 (9.71)	4 (9.71)
Loan loss (%)	4 (11.70)	4 (10.36)	5 (9.23)	5 (9.45)	5 (8.72)	5 (7.94)	5 (7.80)
Asset value fluctuation (%)	5 (10.97)	5 (8.47)	6 (6.40)	6 (6.19)	6 (5.89)	6 (5.93)	6 (5.94)
Capital availability (%)	6 (7.79)	7 (6.34)	7 (6.02)	7 (5.73)	7 (5.75)	7 (5.14)	7 (5.00)
Financial expense (%)	7 (4.17)	8 (3.83)	8 (3.75)	8 (2.93)	8 (2.64)	–	–
Development of borrowers (%)	8 (4.02)	–	–	–	–	–	–
Strategy (%)	–	6 (7.22)	3 (12.45)	1 (17.43)	1 (19.90)	1 (23.37)	1 (23.50)
Reputation (%)	–	–	–	–	–	8 (2.59)	8 (2.59)

Values in parentheses represent the important ratio of the corresponding bank risk factor type. Other than 2014, in which the top 8 most important risk factors account for 78.89%, the top 8 most important risk factors in other years account for more than 80%.

that the corresponding risk factor cannot be identified in that year. We can see from Table 7 that the number of headings classified into the risk factor 'Country credit rating' is zero in 2010, which indicates that 'Country credit rating' appeared beginning in 2011. Additionally, except for the 'Country credit rating' risk factor being zero in 2010, no other value of zero appears in the table, which indicates that no other risk factors appeared or disappeared in later years.

Furthermore, we also analyse the trends in the importance of bank risk factors over the sample period according to the number of risk factor headings classified into a risk factor summarised in Table 7. Specifically, these 21 bank risk factors can be divided into three categories of more important, stable and less important. The specific change in trends of the importance of bank risk factors in these three categories are visually shown in Figures 4–6, respectively.

From Figure 4, it is easy to see that the importance of strategy, political environment and reputation experienced an increasing tendency. As recorded in Table 7, the number of risk factor headings classified into strategy, political environment and reputation increased from 287, 38 and 116 to 1904, 114 and 210, respectively. The sharp increase in strategy made it become the second most important bank risk factor type for commercial banks. Thus, in addition to the top most important risk factors, researchers and bankers

Table 7
Number of risk factor headings classified into each of the 21 risk factors from 2010 to 2016

		2010	2011	2012	2013	2014	2015	2016
Category	Risk factor	Number of headings						
More important	Strategy	287	661	986	1,363	1,616	1,877	1,904
	Political environment	38	98	145	127	101	93	114
	Reputation	116	203	219	188	211	208	210
Stable	Product and service	135	229	165	150	142	143	160
	Accounting standard	130	103	98	122	116	122	128
	Merger and acquisition	144	126	128	125	121	133	143
	Country credit rating	0	15	15	19	17	12	16
	Third-party cooperation	113	94	75	82	82	73	102
	Regulation	1,387	1,688	1,310	1,285	1,450	1,289	1,306
	Disaster	43	33	25	33	24	25	26
	Other risk factors	117	114	147	121	110	109	91
Less important	Asset value fluctuation	1,013	775	507	484	478	476	481
	Capital availability	719	580	477	448	467	413	405
	Loan loss	1,080	948	731	739	708	638	632
	Management operation	1,281	1,260	1,054	1,019	1,019	1,002	965
	Macroeconomic factor	1,262	1,082	886	806	768	780	787
	Competition	343	286	231	187	201	178	167
	Financial expense	385	351	297	229	214	195	205
	Financial institutions interaction	174	178	169	80	76	60	70
	Liability obligation	93	33	26	28	31	38	29
	Development of borrowers	371	297	229	183	169	169	162

should pay more attention to these three risk factors that have become increasingly important.

Bank risk factors whose significances were relatively stable from 2010 to 2016 are illustrated in Figure 5. We can see that the changes in the importance of eight risk factors, including product and service, other risk factors, disaster, regulation, third-party cooperation, country credit rating, merger and acquisition and accounting standard, are slight, without an obvious upward or downward trend. Thus, the significance levels of these eight risk factors were relatively stable over the period 2010–2016 for commercial banks.

As shown in Figure 6, ten bank risk factors experienced a decreasing tendency in importance from 2010 to 2016. Specifically, from Table 7, the number of risk factor headings classified into asset value fluctuation, capital availability, loan loss, management operation, macroeconomic factor, competition financial expense, financial institutions interaction, liability obligation and development of borrowers declined from 1,013, 719, 1,080, 1,281, 1,262, 343, 385, 174, 93 and 371 to 481, 405, 632, 965, 787, 167, 205, 70, 29 and 162, respectively. These declining trends in the number of risk factor headings

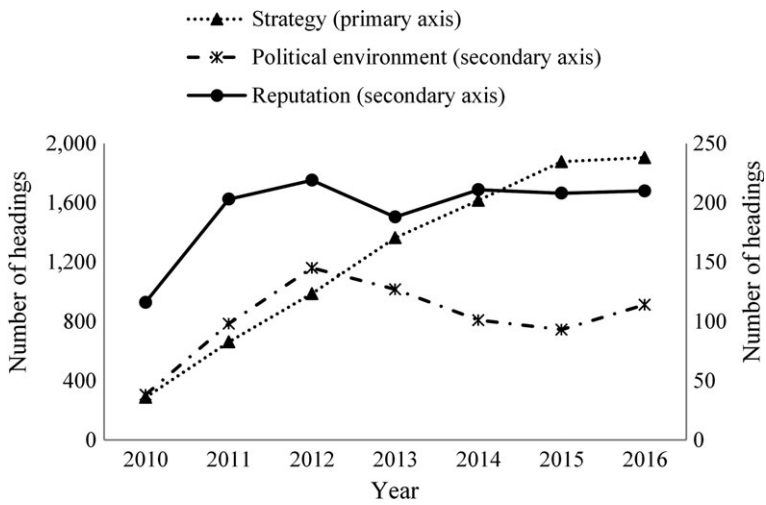

Figure 4 Three risk factors that became more important from 2010 to 2016.

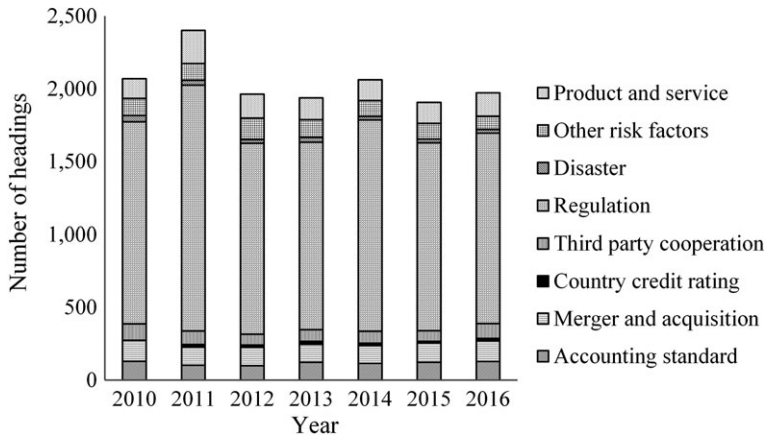

Figure 5 Eight risk factors whose importance remained stable from 2010 to 2016.

classified into these ten bank risk factors show that the attention paid by US commercial banks to the impacts of these ten risk factors on bank risks gradually decreased, which further indicates that these ten bank risk factors became less important from 2010 to 2016.

Overall, the importance of each bank risk factor is changing over time. Strategy, political environment and reputation are three bank risk factors that have become increasingly important. The Basel Committee on Banking Supervision suggests that banks' business activities inevitably produce a variety of different types of risk, including credit, market, operational, liquidity, legal, strategic, country and reputational risk (Basel Committee on Banking

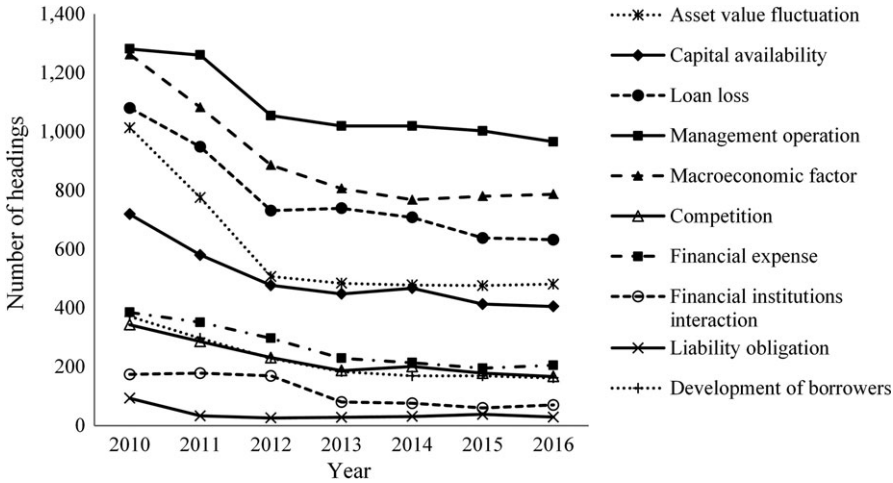

Figure 6 Ten risk factors that experienced a declining trend in importance from 2010 to 2016.

Supervision, 2006). However, the key issues of bank risk management are the measurements of credit, market and operational risks in previous studies (Li *et al.*, 2015). The measurements for increasingly important strategic, political and reputational risks have not been paid enough attention. Thus, in the future, researchers need to put more emphasis on measuring the strategic, political and reputational risks.

3.5. Comparison with the unsupervised text mining approach

To show that using our proposed naive collision algorithm can result in more accurate classifications of bank risk factors, we made an empirical comparison between it and the unsupervised text mining approach called Sent-LDA in terms of clustering quality and predictive power (Bao and Datta, 2014). The predictive power and clustering quality are evaluated using two commonly used measures of precision and recall (Powers, 2011).

Since random sampling is most appropriate for obtaining a representative sample (Grimmer and Stewart, 2013), in accordance with Bao and Datta (2014), who made an empirical comparison between Sent-LDA and LDA, we randomly select 3,000 out of a total of 59,418 textual risk factor headings for labelling. As discussed in Section 3.2, four persons on our research team manually labelled these 3,000 headings with meaningful names. Specifically, each person labelled 1,500 out of 3,000 headings and each heading was labelled by two people. To ensure consistency, we only retained the headings whose labels were agreed upon by both annotators. This led to a set of 2,653 examples. Then, the headings with the same label were merged together as one classification. Thus, by manually labelling the headings of textual risk

disclosures, we obtained the most accurate classification result of bank risk factors, which is regarded as the benchmark to measure the classification accuracies of our proposed naive collision algorithm and the Sent-LDA text mining approach. For convenience, the classification results obtained through manually labelling the 2,653 headings are called manual bank risk factors.

Having obtained the manual bank risk factors by determining the manual labels of 2,653 sample headings, we then identified the bank risk factors using the proposed naive collision algorithm and the Sent-LDA approach, respectively. The classifications obtained via the proposed naive collision algorithm were called semi-supervised naive collision (SSNC) bank risk factors and the labels given to them were called SSNC labels. The classifications obtained via the Sent-LDA approach were called Sent-LDA bank risk factors and the labels given to them were named Sent-LDA labels. The SSNC labels and Sent-LDA labels were collectively called the text mining labels for convenience.

By observing the headings classified into one text mining bank risk factor, SSNC or the Sent-LDA bank risk factor, the headings whose manual labels were consistent with the text mining label of the bank risk factor indicate that these headings contained in the text mining bank risk factor are correct. Thus, for one SSNC or Sent-LDA bank risk factor, the precision rate is used to reflect the proportion of headings whose manual labels are consistent with the text mining label of the bank risk factor. Formally,

$$PR_i = \frac{M_i}{N_i}, \tag{3}$$

where PR_i denotes the precision rate of text mining bank risk factor i. M_i represents the number of headings whose manual labels are consistent with the label of the text mining bank risk factor i in a text mining bank risk factor i. N_i stands for the number of all of the headings classified into the text mining bank risk factor i. The larger PR_i is, the higher the proportion of the correct headings contained in one text mining bank risk factor, and the higher the accuracy of the classification result obtained through the text mining approach will be.

From the perspective of manual labels of headings, headings with the same manual label may be classified into different text mining bank risk factors. Headings are classified into the correct classification when the label of the text mining bank risk factor is consistent with the manual label of headings. Thus, for a manual bank risk factor, the recall rate is used to measure the proportion of headings that are classified into the correct text mining bank risk factor. The recall rate is written as

$$RR_j = \frac{W_j}{K_j}, \tag{4}$$

where RR_j denotes the recall rate of the manual bank risk factor j. W_j represents the number of headings that are classified into the correct text

mining bank risk factor in a manual bank risk factor j. K_j denotes the number of all of the headings contained in the manual bank risk factor j. The larger RR_j is, the higher the proportion of headings that are classified into the correct text mining bank risk factor, and the higher the accuracy of the classification result.

Overall, the two measures of precision and recall can be used to measure the accuracy of the classification results. Through the confusion matrix, we can calculate the precision rate and recall rate for each of the identified 21 bank risk factors according to Equations (3) and (4). Then, by averaging the precision rate and recall rate of the identified 21 bank risk factors, we obtain the averaged precision rate and averaged recall rate, respectively. The precision rate and recall rate of each bank risk factor are used to reflect the classification accuracy of the corresponding bank risk factor. The averaged precision rate and averaged recall rate reflect the averaged level of classification accuracy using the text mining approach to classify bank risk factors.

As we want to make an empirical comparison in terms of clustering quality and predictive power, we need to implement two comparison experiments between our proposed semi-supervised naive collision algorithm and the unsupervised Sent-LDA approach. Specifically, the selected 3,000 headings are used for comparing the clustering quality. In other words, we want to compare the performance of our proposed semi-supervised text mining approach and the unsupervised Sent-LDA text mining approach in classifying these 2,653 headings of textual risk disclosures. The predictive power evaluates how well a model performs when predicting unobserved documents. Thus, by removing the selected 3,000 headings from the total of 59,418 headings of textual risk disclosures, we obtain the classifiers of the naive collision algorithm and the Sent-LDA approach using the remaining 56,418 headings. Then, we apply the SSNC and Sent-LDA classifiers into the selected 3,000 headings to compare the predictive power by classifying the selected 3,000 headings.

The confusion matrixes are summarised in Tables A1, A2 and A3 in the Appendix. The naive collision algorithm produces the same classification results in the two experiments of clustering quality and predictive power; therefore, these two experiments produce the same confusion matrix (Table A1). Tables A2 and A3 record the confusion matrixes of the unsupervised Sent-LDA text mining approach in the experiments of clustering quality and predictive power, respectively. The calculated precision rate and recall rate are summarised in Table 8.

From Table 8, both the precision rate and recall rate indicate that our proposed semi-supervised naive collision algorithm performs better than the unsupervised Sent-LDA approach in terms of both clustering quality and predictive power. Specifically, when evaluating the clustering quality, the averaged precision rate of our proposed naive collision algorithm is up to 75.41 percent, which is far higher than that of the Sent-LDA approach (19.53 percent). The averaged recall rate of our proposed naive collision algorithm (76.05 percent) is also far higher than that of the Sent-LDA approach (19.93

percent). Thus, the higher averaged precision rate and the higher averaged recall rate show that our proposed naive collision algorithm can obtain more accurate classification results of bank risk factors. In the experiment of evaluating predictive power, compared with the averaged precision rate (26.54 percent) and averaged recall rate (27.61 percent) of the Sent-LDA approach, the averaged precision rate (75.41 percent) and averaged recall rate (76.05 percent) of our proposed naive collision algorithm are far higher. Thus, our proposed semi-supervised text mining approach performs better in terms of predictive power.

Furthermore, the precision rate and recall rate of our proposed semi-supervised text mining approach are the same in two experiments of clustering quality and predictive power, which indicates that our proposed semi-supervised text mining approach can achieve the same highly accurate classification result, whether used for a small sample or a large sample. However, the unsupervised Sent-LDA approach performs worse with a small sample. The reason is that the classification accuracy of the unsupervised Sent-LDA depends on the size of the sample. The larger the sample size, the more accurate the classifications. Thus, Bao and Datta (2014) obtained a relatively accurate classification result by using the Sent-LDA approach to deal with a large sample with 322,287 headings.

However, in our experiments, the size of the sample is small, which may be the reason that the classification result obtained by the Sent-LDA is less precise. In particular, because we obtain the Sent-LDA classifier based on 56,418 headings in the experiment on evaluating predictive power, rather than 3,000 headings used in the experiment on clustering quality, the averaged precision rate (26.54 percent) and averaged recall rate (27.61 percent) in the experiment on predictive power are larger than the averaged precision rate (19.53 percent) and the averaged recall rate (19.93 percent) in the experiment on clustering quality, which indicates that the Sent-LDA performs even worse in the experiment on clustering quality with a smaller sample.

To summarise, the far greater precision rate and recall rate show that our proposed semi-supervised approach performs better in terms of clustering quality and predictive power. Furthermore, the classification accuracy of our proposed semi-supervised text mining approach is not affected by the size of the sample. Even based on a small sample, our proposed semi-supervised text mining approach can produce the same accurate classification result as the large sample.

4. Conclusion

The identification of bank risk factors in existing literature depends mainly on the researchers' subjective judgements or summarising of previously identified bank risk factors, which lead to a problem that the identified bank risk factors are limited. The main contribution of this paper is to comprehensively discover bank risk factors from qualitative textual risk disclosures reported in financial

Table 8
The precision and recall rate of the semi-supervised and the unsupervised text mining approaches in terms of clustering quality and predictive power

	Semi-supervised		Unsupervised			
			Precision rate		Recall rate	
Bank risk factors	Precision rate Clustering quality/ Predictive power (%)	Recall rate Clustering quality/ Predictive power (%)	Clustering quality (%)	Predictive power (%)	Clustering quality (%)	Predictive power (%)
Accounting standard	97.32	93.97	23.42	77.78	22.41	78.45
Merger and acquisition	99.18	90.30	20.87	20.00	17.91	17.16
Asset value fluctuation	89.12	66.84	18.68	38.38	8.67	19.39
Capital availability	91.37	72.57	23.01	21.30	14.86	13.14
Competition	100.00	80.32	33.70	33.95	16.49	29.26
Disaster	100.00	65.08	32.94	60.00	44.44	64.29
Financial expense	75.36	62.28	42.86	48.68	50.30	55.09
Financial institutions interaction	70.79	74.12	22.75	18.29	56.47	55.29
Liability obligation	90.43	82.52	20.51	20.30	23.30	26.21
Loan	63.12	61.38	18.09	16.28	11.72	9.66
Management operation	89.04	69.89	21.78	30.33	11.83	19.89
Macroeconomic factor	83.44	70.79	25.69	19.17	20.79	12.92
Other industry development	74.82	70.27	24.37	25.74	19.59	23.65
Other risk factors	17.02	25.40	7.48	7.29	17.46	11.11
Political environment	29.49	100.00	3.27	2.25	10.87	4.35
Product and service	48.10	78.35	6.25	12.63	6.19	12.37
Regulation	78.00	64.29	37.96	26.57	22.53	20.88
Reputation	67.88	81.58	12.00	28.28	13.16	24.56
Strategy	45.14	92.86	3.70	9.45	8.57	17.14
Third-party cooperation	74.02	96.91	1.82	34.38	2.06	34.02
Country credit rating	100.00	97.30	8.97	6.20	18.92	21.62
Averaged	75.41	76.05	19.53	26.54	19.93	27.16

statements, which contain a huge amount of information on bank risks. To analyse textual risk disclosures, a new semi-supervised text mining approach called the naive collision algorithm is proposed. Compared with the typical unsupervised Sent-LDA text mining approach, we empirically prove that our proposed semi-supervised naive collision algorithm can deliver a more accurate classification of bank risk factors in terms of clustering quality and predictive power.

In the experiment, by using our proposed semi-supervised naive collision algorithm, we identified 21 bank risk factors in total based on 59,418 textual risk factor headings collected from 2,189 US commercial banks' Form 10-K filings from 2010 to 2016. Furthermore, we analyse the importance of each risk factor based on the disclosure frequency. Our findings and their practical implications for researchers and regulators are as follows. First, the 21 bank risk factors identified from textual risk disclosures are far more than the bank risk factors identified in previous studies. Using these 21 bank risk factors can more comprehensively reflect the risk condition of the banking industry, and lays a fundamental foundation for further bank risk measurement. Second, the findings of risk factor importance are different from the previous views in two aspects. Traditionally, credit and market risks, the two financial risks, are considered as the main risk types faced by banks. Thus, practitioners, researchers and regulators have paid more attention to their management. However, by ranking bank risk factors according to their importance, we found that the top 3 most important risk factors whose cumulative importance approached 50 percent are non-financial risk factors, i.e. regulation, strategy and management operation. Moreover, by analysing the annual change in risk factor importance, our empirical results show that the risk factors of strategic, political and reputation have become increasingly important, which reflects that practitioners and regulators should put more emphasis on the management of these risk factors.

This study is not without limitations. A common limitation of text-based analysis is that only the information contained in the text can be analysed. The bank risk factors that have not been disclosed in financial statements cannot be identified in this paper.

Acknowledgments

This work was supported by grants from the National Natural Science Foundation of China (71601178, 71425002) and the Youth Innovation Promotion Association of Chinese Academy of Sciences (2012137, 2017200). We sincerely thank the editor and reviewers for their very valuable and professional comments.

References

Aas, K., X. K. Dimakos, and A. Oksendal, 2007, Risk capital aggregation, *Risk Management* 9, 82–107.

Alexander, C., and J. Pezier, 2003, On the aggregation of market and credit risks, ISMA Centre Discussion Papers in Finance No. 2003-13 (University of Reading).
Alibux, A., 2007, Cross-border mergers and acquisitions in the European banking sector, Dissertation (Erasmus University of Rotterdam).
Allen, D., and R. Faff, 2012, The global financial crisis: some attributes and responses, *Accounting and Finance* 52, 1–7.
Bao, Y., and A. Datta, 2014, Simultaneously discovering and quantifying risk types from textual risk disclosures, *Management Science* 60, 1371–1391.
Basel Committee on Banking Supervision, 2006, *International Convergence of Capital Measurement and Capital Standards: A Revised Framework* (Bank for International Settlements, Basel).
Bellini, T., 2013, Integrated bank risk modeling: a bottom-up statistical framework, *European Journal of Operational Research* 230, 385–398.
Bengio, Y., I. J. Goodfellow, and A. Courville, 2015, Deep learning, *Nature* 521, 436–444.
Berger, A., C. Buch, G. Delong, and R. DeYoung, 2004, Exporting financial institutions management via foreign direct investment mergers and acquisitions, *Journal of International Money and Finance* 23, 333–366.
Berger, A. N., S. El Ghoul, O. Guedhami, and R. A. Roman, 2016, Internationalization and bank risk, *Management Science* 63, 2283–2301.
Breuer, T., M. Jandacka, K. Rheinberger, and M. Summer, 2010, Does adding up of economic capital for market and credit risk amount always to conservative risk estimates?, *Journal of Banking and Finance* 34, 703–712.
Brewer, T., and P. Rivoli, 1990, Politics and perceived country creditworthiness in international banking, *Journal of Money, Credit and Banking* 22, 357–369.
Brown, S. J., 2012, Quantitative measures of operational risk: an application to funds management, *Accounting and Finance* 52, 1001–1011.
Buch, C., and G. DeLong, 2004, Cross-border bank mergers: what lures the rare animal?, *Journal of Banking and Finance* 28, 2077–2102.
Campbell, J. L., H. C. Chen, D. S. Dhaliwal, H. M. Lu, and L. B. Steele, 2014, The information content of mandatory risk factor disclosures in corporate filings, *Review of Accounting Studies* 19, 396–455.
Chang, J., J. L. Boyd-Graber, S. Gerrish, C. Wang, and D. M. Blei, 2009, Reading tea leaves: how humans interpret topic models, in: Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, A. Culotta, eds., *Advances in Neural Information Processing Systems* (Curran Associates, New York), 288–296.
Chari, A., and N. Gupta, 2008, Incumbents and protectionism: the political economy of foreign entry liberalization, *Journal of Financial Economics* 88, 633–656.
Dyer, T., M. Lang, and L. Stice-Lawrence, 2017, The evolution of 10-K textual disclosure: evidence from Latent Dirichlet Allocation, *Journal of Accounting and Economics* 64, 221–245.
Embrechts, P., G. Puccetti, and L. Rüschendorf, 2013, Model uncertainty and VaR aggregation, *Journal of Banking and Finance* 37, 2750–2764.
Gleason, K., I. Mathur, and R. Wiggins III, 2006, The use of acquisitions and joint ventures by US banks expanding abroad, *Journal of Financial Research* 29, 503–522.
Grimmer, J., and B. M. Stewart, 2013, Text as data: The promise and pitfalls of automatic content analysis methods for political texts, *Political Analysis* 21, 267–297.
Grundke, P., 2009, Importance sampling for integrated market and credit portfolio models, *European Journal of Operational Research* 194, 206–226.
Grundke, P., 2010, Top-down approaches for integrated risk management: how accurate are they?, *European Journal of Operational Research* 203, 662–672.
Hagendorff, J., and K. Keasey, 2009, Post-merger strategy and performance: evidence from the US and European banking industries, *Accounting and Finance* 49, 725–751.

Huang, K. W., and Z. L. Li, 2011, A multilabel text classification algorithm for labeling risk factors in SEC form 10-K, *ACM Transactions on Management Information Systems (TMIS)* 2, 1–19.

Jarrow, R. A., and S. M. Turnbull, 2000, The intersection of market and credit risk, *Journal of Banking and Finance* 24, 271–299.

Kretzschmar, G., A. J. McNeil, and A. Kirchner, 2010, Integrated models of capital adequacy – why banks are undercapitalized, *Journal of Banking and Finance* 34, 2838–2850.

Laeven, L., and R. Levine, 2009, Bank governance, regulation and risk taking, *Journal of Financial Economics* 93, 259–275.

Li, J., and S. Guisinger, 1992, The globalization of service multinationals in the 'triad' regions: Japan, Western Europe and North America, *Journal of International Business Studies* 23, 675–696.

Li, J., X. Zhu, C. F. Lee, D. Wu, J. Feng, and Y. Shi, 2015, On the aggregation of credit, market and operational risks, *Review of Quantitative Finance and Accounting* 44, 161–189.

Li, J., L. Wei, C. F. Lee, X. Zhu, and D. Wu, 2018, Financial statements based bank risk aggregation, *Review of Quantitative Finance and Accounting* 50, 673–694.

Medova, E. A., and R. G. Smith, 2005, A framework to measure integrated risk, *Quantitative Finance* 5, 105–121.

Mei, Q. Z., X. H. Shen, and C. X. Zhai, 2007, Automatic labeling of multinomial topic models, in: P. Berkhin, R. Caruana, X. Wu, eds., Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining (ACM, New York), 490–499.

Miller, G. S., 2017, Discussion of 'The evolution of 10-K textual disclosure: evidence from Latent Dirichlet Allocation', *Journal of Accounting and Economics* 64, 246–252.

Mirakur, Y., 2011, Risk disclosure in SEC corporate filings, Working paper (University of Pennsylvania, Philadelphia, PA).

Powers, D. M. W., 2011, Evaluation: from precision, recall and F-Factor to ROC, informedness, markedness & correlation, *Journal of Machine Learning Technologies* 2, 37–63.

Rosenberg, J. V., and T. A. Schuermann, 2006, General approach to integrated risk management with skewed, fat-tailed risks, *Journal of Financial Economics* 79, 569–614.

SEC, 2005, Securities and Exchange Commission final rule, release no. 33-8591(FR-75). Available at: http://www.sec.gov/rules/final/33-8591.pdf.

SEC, 2010, Annual report pursuant to section 13 or 15(d) of the Securities Exchange Act of 1934, general instructions. Available at: http://www.sec.gov/about/forms/form10-k.pdf.

Turney, P. D., and P. Pantel, 2010, From frequency to meaning: vector space models of semantics, *Journal of Artificial Intelligence Research* 37, 141–188.

Wei, L., G. Li, X. Zhu, and J. Li, 2019, Developing a hierarchical system for energy corporate risk factors based on textual risk disclosures, *Energy Economics*. In press. https://doi.org/10.1016/j.eneco.2019.01.020

Zhu, X., Y. Wang, and J. Li, 2019, Operational risk measurement: a loss distribution approach with segmented dependence, *Journal of Operational Risk* 14, 1–20. https://doi.org/10.21314/JOP.2019.220

Zhu, X., Y. Xie, J. Li, J. Chen, S. Yang, X. Sun, and D. Wu, 2015, Change point detection for subprime crisis in American banking: from the perspective of risk dependence, *International Review of Economics and Finance* 38, 18–28.

Zhu, X., S. Y. Yang, and S. Moazeni, 2016, Firm risk identification through topic analysis of textual financial disclosures, *2016 Symposium Series on Computational Intelligence*, 1–8.

Appendix

Table A1
Confusion matrix of the semi-supervised text mining naive collision approach in the two experiments of clustering quality and predictive power

	AS	MA	AVF	CA	C	D	FE	FII	LO	L	MO	MF	OID	ORF	PE	PS	Reg	Rep	S	TPC	CCR
AS	109	0	0	0	0	0	0	0	0	0	0	6	0	0	0	0	0	0	1	0	0
MA	1	121	0	0	0	0	6	0	0	2	3	0	0	0	0	1	0	0	0	0	0
AVF	0	0	131	2	0	0	6	0	0	14	0	1	1	19	2	3	0	13	4	0	0
CA	0	0	2	127	0	0	6	1	2	2	0	0	11	14	0	10	0	0	0	0	0
C	0	0	0	0	151	0	0	0	7	7	0	1	0	4	2	1	7	0	8	0	0
D	0	0	0	0	0	82	0	0	0	1	0	0	7	0	35	1	0	0	0	0	0
FE	0	0	2	0	0	0	104	0	0	0	4	0	0	17	0	37	0	0	0	3	0
FII	0	0	1	0	0	0	0	63	0	3	0	1	1	0	8	0	0	0	8	0	0
LO	0	0	0	0	0	0	0	0	85	0	0	0	0	1	0	0	15	0	2	0	0
L	0	0	1	0	0	0	1	2	0	89	1	2	2	6	0	8	3	0	23	7	0
MO	0	0	0	8	0	0	0	1	0	0	130	1	0	3	1	3	8	25	0	6	0
MF	0	0	0	0	0	0	0	4	0	1	0	126	7	7	22	0	0	0	11	0	0
OID	0	0	0	0	0	0	0	6	0	17	0	7	104	6	0	5	0	0	1	2	0
ORF	1	0	1	0	0	0	6	0	0	0	7	6	4	16	1	13	0	0	7	1	0
PE	0	0	0	0	0	0	0	0	0	0	0	0	0	0	46	0	0	0	0	0	0
PS	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	76	0	6	0	13	0
Reg	0	1	9	2	0	0	1	4	0	0	0	0	1	1	39	0	117	0	6	1	0
Rep	0	0	0	0	0	0	7	7	0	0	1	0	0	0	0	0	0	93	6	0	0
S	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	65	0	0
TPC	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	94	0
CCR	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	36

Acronyms for the names of bank risk factors stand for the names of bank risk factors.

Table A2
Confusion matrix of the unsupervised Sent-LDA text mining approach in the experiment of clustering quality

	AS	MA	AVF	CA	C	D	FE	FII	LO	L	MO	MF	OID	ORF	PE	PS	Reg	Rep	S	TPC	CCR
AS	26	0	3	0	0	1	1	0	0	0	0	28	0	0	0	0	0	1	0	36	20
MA	1	24	0	2	0	0	12	1	3	1	5	5	3	0	64	4	0	8	0	1	0
AVF	25	0	17	7	9	0	1	20	0	13	11	7	32	0	1	23	1	9	1	13	6
CA	11	8	0	26	7	7	12	2	0	9	0	19	14	5	8	10	6	0	26	5	0
C	0	18	12	3	31	5	0	40	25	15	6	19	0	3	0	0	0	0	0	10	1
D	3	0	0	2	3	56	0	2	6	0	5	0	1	0	0	0	0	0	48	0	0
FE	0	0	0	0	2	10	84	1	0	17	0	1	4	1	1	7	11	28	0	0	0
FII	3	0	0	2	1	0	0	48	0	1	0	0	4	12	7	0	1	0	0	6	0
LO	11	0	13	0	0	0	0	5	24	0	0	1	1	0	0	10	0	0	10	11	17
L	0	0	0	4	0	0	15	13	2	17	4	2	19	12	0	16	2	6	29	4	0
MO	7	15	18	1	10	25	6	1	1	0	22	0	4	18	13	10	8	10	3	9	5
MF	1	5	0	2	4	0	9	31	8	0	2	37	2	20	16	0	3	28	10	0	0
OID	0	1	7	2	1	0	24	26	11	6	11	0	29	16	0	0	1	0	12	1	0
ORF	1	0	0	5	9	0	2	2	0	2	8	1	0	11	9	0	4	0	6	1	2
PE	0	0	0	0	0	7	11	0	9	0	1	0	0	0	5	0	7	0	0	1	5
PS	0	3	0	0	1	11	8	8	7	0	3	14	1	22	1	6	0	8	2	0	2
Reg	4	1	18	32	1	1	1	1	3	9	5	10	0	14	15	1	41	1	9	6	9
Rep	0	13	2	14	1	9	3	3	0	1	18	0	4	1	5	6	15	15	0	4	0
S	11	14	1	9	1	2	0	0	2	3	0	0	0	1	7	2	6	1	6	0	4
TPC	0	13	0	2	11	36	7	7	7	0	0	0	1	8	1	1	1	0	0	2	0
CCR	7	0	0	0	0	0	0	0	9	0	0	0	0	3	0	0	1	10	0	0	7

Acronyms for the names of bank risk factors stand for the names of bank risk factors.

Table A3
The confusion matrix of the unsupervised Sent-LDA text mining approach in the experiment of predictive power

	AS	MA	AVF	CA	C	D	FE	FII	LO	L	MO	MF	OID	ORF	PE	PS	Reg	Rep	S	TPC	CCR
AS	91	2	0	0	1	0	0	1	0	0	0	13	0	0	0	0	0	0	0	0	8
MA	1	23	0	2	0	0	49	2	22	0	8	1	8	7	1	1	1	4	0	0	4
AVF	0	0	38	0	14	0	0	0	4	13	45	7	15	16	0	0	0	1	8	12	23
CA	5	0	4	23	15	16	7	1	2	10	1	7	5	1	4	22	0	2	25	2	23
C	0	0	0	0	55	8	6	44	1	0	1	8	0	0	7	17	32	0	0	7	2
D	2	0	5	0	0	81	4	30	0	0	2	0	1	0	0	1	0	0	0	0	0
FE	0	5	1	2	3	4	92	0	0	20	0	9	2	11	0	0	1	0	0	17	0
FII	10	1	5	1	1	0	0	47	0	3	1	2	0	0	0	0	0	8	0	0	6
LO	0	2	0	13	13	0	0	6	27	0	0	11	5	2	2	2	0	10	10	0	0
L	0	14	6	6	21	0	0	2	3	14	0	1	12	8	6	2	20	3	27	0	0
MO	2	0	1	3	10	0	13	10	4	1	37	9	1	23	5	9	8	23	6	10	11
MF	0	0	6	3	0	5	3	69	25	0	7	23	2	0	0	3	22	0	9	0	1
OID	0	0	8	19	1	4	1	14	0	1	0	10	35	0	0	11	5	1	16	0	22
ORF	1	5	0	0	0	0	1	3	13	5	2	4	9	7	1	0	4	3	1	4	0
PE	0	0	0	0	0	9	0	0	8	0	8	0	11	0	2	0	7	0	1	0	0
PS	0	5	14	17	2	0	2	6	8	1	1	3	5	2	13	12	1	0	0	4	1
Reg	1	35	11	10	1	4	9	0	1	12	1	3	6	2	25	4	38	10	2	6	1
Rep	0	10	0	1	23	1	0	11	14	1	1	2	0	10	0	1	0	28	4	1	6
S	0	11	0	6	0	2	0	0	0	5	5	2	8	5	4	1	3	0	12	0	6
TPC	0	2	0	2	2	1	2	1	1	0	2	0	11	2	19	0	1	5	6	33	7
CCR	4	0	0	0	0	0	0	10	0	0	0	5	0	0	0	9	0	1	0	0	8

Acronyms for the names of bank risk factors stand for the names of bank risk factors.