The value of official website information in the credit risk evaluation of SMEs

Cuiqing Jiang, School of Management, Hefei University of Technology, 193 Tunxi Road, Hefei, Anhui 230009, PR China

Chang Yin, School of Management, Hangzhou Dianzi University, Xiasha Higher Education Zone, Hangzhou, Zhejiang 310018, PR China

Qian Tang, School of Computing and Information Systems, Singapore Management University, 80 Stamford Road, 178902, Singapore

Zhao Wang, School of Management, Hefei University of Technology, 193 Tunxi Road, Hefei, Anhui 230009, PR China

Published in Journal of Business Research (2023 December) 169, 114290

DOI: 10.1016/j.jbusres.2023.114290

Abstract: The official websites of small and medium-sized enterprises (SMEs) not only reflect the willingness of an enterprise to disclose information voluntarily, but also can provide information related to the enterprises' historical operations and performance. This research investigates the value of official website information in the credit risk evaluation of SMEs. To study the effect of different kinds of website information on credit risk evaluation, we propose a framework to mine effective features from two kinds of information disclosed on the official website of a SME—design-based information and content-based information—in predicting its credit risk. We select the SMEs in the software and information technology services industry and find that including content-based information in models significantly improves the prediction accuracy. Specifically, the depth and dynamics metrics of the content-based information convey SME performance and mitigate the information asymmetry between SMEs and financial institutions.

Keywords: Credit risk, Information asymmetry, Official website, SMEs

1. Introduction

Small and medium-sized enterprises (SMEs) are the backbone of economies and the drivers of innovation and job creation (Angilella et al., 2019). However, SMEs have higher risks and fewer assets compared with large companies, making it difficult for them to attract capital for further development (Altman & Sabato, 2007). The credit market represents an important source of capital for SMEs, and allows enterprises to obtain financial support from financial institutions according to their credit risk. To evaluate the credit risk, financial information (such as the current ratio, return on operating assets, and debt ratio) has been widely focused on since it directly reflects the financial performance and repayment ability of enterprises (Gordini, 2014). However, for SMEs, due to the lack of mandatory financial reporting requirements, the quality of financial information is limited (Cassar et al., 2015). Thus, to complement the financial information, utilizing non-financial information has a key role in allocating money to "good-credit" enterprises (Ge et al., 2017, Raman et al., 2022).

There is abundant literature on using non-financial information to improve the accuracy of credit risk assessment of SMEs (Donovan, 2021). Information posted on social networks (Facebook, Twitter, Weibo, etc.) has attracted extensive attention from scholars (Zu et al., 2019, Sukumar et al., 2021). However, for unlisted SMEs, few enterprises have this kind of information. With the development of information technology, SMEs start to build official website, where a SME posts positive information and presents its competitiveness to site visitors (Hung et al., 2014) by introducing itself, presenting its products and/or services, releasing news, and so on. We believe that the official website information is related to the enterprises' historical operations and performance. Extending this stream of research, our study explores a new source of non-financial information—SMEs' official website information—to evaluate the credit risk of SMEs.

The enterprise's introduction usually contains the date of establishment, the location, qualification certifications, and the main business of the enterprise. The products/services presented show details of the products/services available, for the convenience of customers who are gathering information or making a purchase. And the news releases record the daily activities of the enterprise, such as meetings, awards, business development, cooperation projects, and innovations. The above information presents an enterprise from different perspectives, including the management level, productivity, social responsibility, and innovation ability. Existing literature has demonstrated that these perspectives are effective in assessing financial performance and repayment ability of enterprises (Bonsall et al., 2017, Chang et al., 2019). Thus, we believe that such website information can complement insufficient financial information and improve the accuracy of the credit risk evaluation of SMEs. In addition, official websites are put on the record by the supervision departments of government, and an enterprise will suffer a loss of reputation and legal punishment if it posts false information deliberately. Therefore, official website information is a kind of open-access, credible, and self-disclosed information. Meanwhile, it contains dynamic information released in a timely manner.

The behavior of building an official website reflects the willingness of an enterprise to disclose information. Firth et al. (2019) has demonstrated that enterprises with good performance like to disclose more information, to reduce information opacity, than the bad ones do. Thus, we assume that SMEs with official websites, who voluntarily disclose information related to their historical operations and performance on official websites, are more likely to be "good-credit" enterprises than SMEs without official websites. We propose the first research question:

RQ1: Are there differences between the credit risk of SMEs with and without official websites?

Although more and more enterprises set up official websites for publicity (Jean & Kim, 2020), these websites vary in the design of the pages, the quality of the website content, and the frequency of updates of dynamic information. An interesting and high-efficiency website should also contain various elements, such as a mission statement, description of service, free information, and ease of navigation. Thus, this paper explores whether these various factors of official website help financial institutions separate the "good" borrowers from loan applicants.

Usually, the information of SMEs' official websites is classified into two kinds of unstructured information: design-based information and content-based information (Lopes & Melão, 2016). The design-based information usually contains information about the website's functioning, service, and page design style. The content-based information is statement or description information, including, but not limited to, the enterprise's introduction of itself, business activities, and relationships with other firms and organizations (Gök et al., 2015). Based on the two kinds of unstructured information, we propose the two research questions:

RQ2: Which kinds of information posted on official websites will be effective to evaluate the credit risk of SMEs?
RQ3: Which features extracted from the effective information will be useful?

This study explores the value of official website information by answering the above questions. We examine the first research question by adding a dummy variable, namely whether an enterprise builds an official website, into prediction models. The results show that the predictive performances of models are significantly improved. We also validate SMEs with official websites have lower credit risk.

To address the second and third research questions, we propose a framework to extract features from official websites. For the design-based information, we design features based on the previous studies. Content-based information is divided into static and dynamic information according to their update frequency. For the static-content-based information, we construct features from two aspects: information breadth and information depth. For the dynamic-content-based information, we not only construct features from the two aspects, but also consider the dynamics metrics to measure the trend of updating of the information. After constructing features, we select the best feature subset by a feature selection method and add the selected features into the prediction models to validate their effects. The results indicate that the content-based information has important predictive value. In particular, the depth of the content-based information and its dynamics metrics can improve the accuracy of credit risk evaluation of SMEs.

This paper makes several contributions. First, to our best knowledge, this paper is the first to explore the value of SMEs' official website information in credit risk evaluation. Second, we show that there are indeed differences between the credit risk of SMEs with and without official websites, and demonstrate that the different kinds of information posted on the website have different effects on the credit risk evaluation. Third, we find that the depth of content-based information and dynamics metrics can mitigate the information asymmetry between SMEs and financial institutions. Finally, we advance the understanding of how official website information affects credit risk assessment by exploring an interpretable model.

Our work has important managerial implications for practice. First, this research helps financial institutions evaluate the credit risk of SMEs more accurately. Given the granting performance of our proposed framework, the improved accuracy can reduce the financial losses caused by defaults. Second, our findings provide a reference for SMEs to use to harness the value of official website information, to help "good-credit" enterprises obtain financial support more easily for further development. Third, the proposed framework provides a solution to people in a variety of disciplines who process unstructured website information.

2. Literature review

2.1. Credit risk evaluation of SMEs

Developing credit risk models is important to minimize lender's losses caused by default loans and allocate financial support to SMEs with good credit. Following the large literature, we conclude that exploring effective predictors and designing excellent prediction methods to build credit risk models are the two main streams. Since this paper focuses on the former only, we summarize the existing literature on exploring effective predictors in credit scenario.

Early studies have utilized accounting ratios, like the current ratio, return on operating assets, and debt ratio, to assess the risk of corporate failure (Yazdanfar & Nilsson, 2008). However, for unlisted SMEs, Altman et al. (2014) argued that the effective accounting ratios are limited availability due to information opacity. Thus, mining predictors from the non-financial information is crucial for financial institutions to complement accounting ratios of SMEs.

Firstly, firm-specific information, such as the firm age, managerial ability, number of employees, and characteristics of the board of directors (Ciampi, 2015; Bonsall et al., 2017), makes a significant contribution to increasing the default prediction power of credit risk model. This paper considered the firm-specific information in the benchmark model. Secondly, network information of enterprises, as an alternative data representing enterprises' external resources (Ravindran et al., 2015), has been widely studied, including distribution and customer networks (Angilella & Mazzù, 2015); the relation networks (Tobback et al., 2017); financial networks (Ahelegbey et al., 2019); and supply chain networks (Zhu et al., 2017). From the network perspective, this paper designs a feature: the number of websites that an enterprise's official website links to.

Thirdly, existing studies extract the enterprises' operations from audit reports (Sánchez et al., 2013), risk events from legal judgments (Yin et al., 2020), and financial performance from financial reports (Tsai & Wang, 2017) using text mining technology. However, there are often insufficient public reports and media coverage for unlisted SMEs, due to the lack of mandatory disclosure requirements.

Extending this stream of research, our study explores a new source of non-financial information—SMEs' official website information. Official websites are put on the record by the supervision departments of government, and an enterprise will suffer a loss of reputation and legal punishment if it posts false information deliberately. Therefore, official website information is a kind of open-access, credible, and self-disclosed information, which can be collected with low labor cost by coding. Meanwhile, it contains dynamic information released in a timely

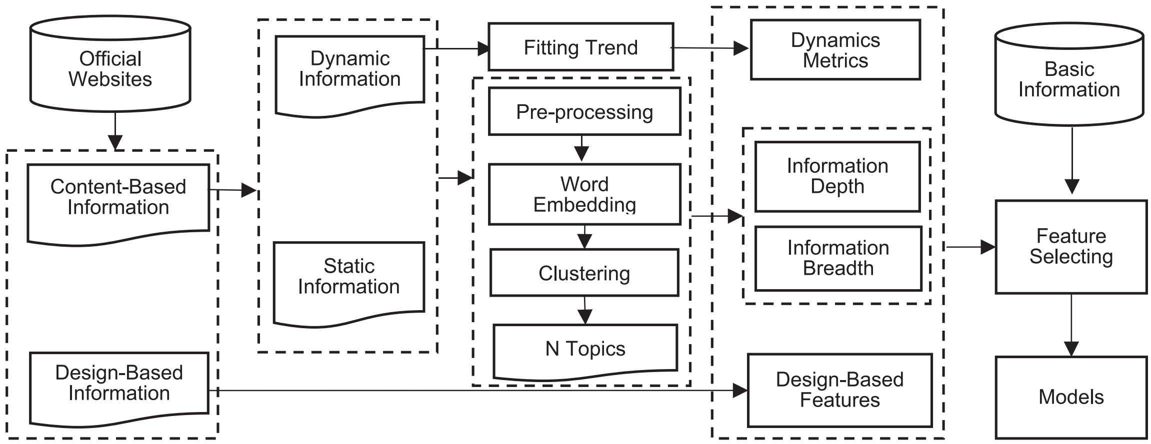

Fig. 1. Framework for evaluating the credit risk of SMEs with official websites.

manner.

2.2. Official website information of SMEs

Enterprise websites, as a digital platform, reflect the digitalization level of SMEs (Salvi et al., 2021). The digitalization can change the relationship between companies and their markets (Chatterjee & Kar, 2020; Caputo et al., 2021). Caputo et al. (2017) highlighted that "the digitalization supports enterprises that are beginning to understand their business environments at a more granular level, are creating new products and services, and are responding more quickly to change as it occurs". Thus, we consider that whether a SME has an official website and the information posted on its official websites, can be associated with its business performance.

Additionally, the enterprises with good repayment ability are willing to disclose more information compared to the enterprises with poor repayment ability (Firth et al., 2019; Kim & An, 2021). Further, studies have proposed that a good quality of disclosure can reduce the information asymmetry between managers, stakeholders, and lenders (Talbi & Omri, 2014). Thus, we consider that whether a SME has an official website (disclosure willingness) and the quality of information posted on the official website (disclosure quality) have effect on the credit risk evaluation of SMEs.

For the SMEs with official websites, their website information may differ. Following García et al. (2017), we classify the website information into design-based and content-based information, and further investigate the value of each category in credit risk evaluation.

2.3. Design-based information

Design-based information of website is critical in engaging users (Cyr, et al., 2009), and plays an important role in website quality assessment. Existing literature has applied functioning and service, web page design, interactivity, and multimedia application, to evaluate website quality (Chiou et al., 2010; Parker et al., 2015), which are also considered in this paper.

However, Vila and Kuster (2011) highlighted that a well-designed website cannot carry higher levels of purchase intention and trust. Thus, website design information has little effect on business performance of SME. Additionally, nowadays SMEs usually hire professional website construction firms to build official websites instead of self-development (Hansen, 2019). These dimensions of design-based information of website provided by the third-party outsourcing are highly similar. Thus, "bad-credit" enterprises can acquire the same quality of design-based information as "good-credit" enterprises, at the same cost. This paper intends to validate this point.

2.4. Content-based information

By visiting official websites of SMEs, visitors can access authentic, accurate and up-to-date information (Rahimnia & Hassanzadeh, 2013). Since content posted on website allows a visitor to make a more informed decision (Hasley & Gregg, 2010), the quality of content-based information will affect the opportunities of business success for SMEs and further influences the credit risk of SMEs.

The existing research on the content information of websites usually draws on the accuracy, sufficiency, readability, and reliability (Rekik et al., 2018; Sun et al., 2019) as collected by questionnaires or evaluated by experts. However, these indicators cannot convey the detail and rich degree of the content directly or measure the characteristic of dynamic information. Additionally, the existing research ignores the semantic information. Thus, to evaluate content-based information quality, this paper captures topics clustered by the semantic information, and examines the content quality from two aspects: the information breadth and the information depth (Resch & Kock, 2021). In the content quality assessment scenario, the breadth is for topic variety and the depth is in charge of the narrative style of topic concentration (Chen & Ohta, 2010). In addition, for the dynamic information, we design dynamics metrics to measure the trend of how it is updated.

Information breadth. Higher breadth of disclosure may breed liking via signaling to the receiver the discloser's desire to initiate a closer relationship, communication trust, and eliciting a positive affective response from the receiver (Baruh et al., 2018). Consumer research proposes that information breadth leads to more satisfied consumers by increasing available choices, and thus improves the performance of product manufactures (Pentina & Tarafdar, 2014). In this paper, information breadth is measured by the number of topics the content is involved in, reflecting the richness of the content information.

Information depth. Enterprises reduce the uncertainty of themselves and their offering by providing more descriptive and in-depth information (Adjei et al., 2010). For enterprises with poor performance, they generally unwilling to disclose in-depth information because the detailed/concentrated information might reveal their imperfections and unprofessionalism relative to the enterprises with good performance (Dimoka et al., 2012). Existing literature has studied the effect of information depth on firm performance. Metzger and Flanagin (2013) suggested that information depth is an important criteria to evaluate online information quality, and helps firms establish trust with consumers. In this paper, information depth is measured by the degree of detail or the concentration of content in a specific topic.

Dynamics metrics. Dynamic metrics we proposed are used to measure the trend of the news updates during a given observation period. Cui et al. (2018) indicated that the dynamic information gathering and exchange will eventually decrease information asymmetry over time.

Table 1
The features of the three aspects of design-based information.

Aspect	No.	Features
Access Methods	1	Whether it has a WeChat version
	2	Whether it has a foreign language version
Categories of Information Presentation	3	The number of videos presented
	4	The number of honors presented
	5	Whether it has product presentation
Function & Service	6	Whether it has a search engine
	7	The number of external links
	8	The methods offered to interact (telephone number, email address)
	9	Whether the design has a navigation

Additionally, enterprise with good performance has a superior information management capability than "bad" ones (Mithas et al., 2011), so that they update the dynamic information in a more timely manner and more systematically. Therefore, we consider that the trend of updating is a critical indicator when using the dynamic information.

3. Framework

We propose a framework (see Fig. 1) to extract features from official website information of SMEs, and explore their effect on the credit risk evaluation. The steps of the framework are discussed below.

3.1. Constructing design-based features

An effective website design has an important role for organizations that want to maximize their profits by promoting their services or products in a competitive and limited market (Cebi, 2013). According to previous studies, we constructed the design-based information. The specific features are listed in Table 1.

The different access methods of a website, including WeChat and foreign language versions, reflect diversified ways of advertising. The categories of information presentation of a website include v.ideos presentation, honors presentation, and product presentation. V.ideos uploaded on a website can be used to introduce the brand, sell products, and present a viewpoint intuitively and conveniently. The number of honors presented conveys the competitiveness of an enterprise. Product presentation allows visitors to quickly view all the products/services of an enterprise.

In terms of the function and service of a website, we consider whether it has a search engine, the number of external links, the methods offered to interact, whether the design has a navigation. An internal search engine expresses the firm's desire for users to access the information they are looking for quickly and effectively. External links allow visitors to access information posted on other websites efficiently. A website offering various methods for visitors to interact with an enterprise makes communication between the enterprise and the visitors easy. As far as web page design, navigation design can improve the visitors' experience at the website.

3.2. Measuring information breadth and information depth

The content-based information includes an enterprise's introduction of itself and its news titles. The introduction is static and rarely modified. The news of an SME is dynamic and updated periodically. We focus on the news titles since the title is the summarization of the news, which conveys the most important information using the lowest number of words. For an enterprise, we combine all the news title texts during a specific observation period together as the whole content (called the news-title content) for further study. Fig. 2 illustrates the steps of constructing the features of information depth and information breadth.

(1) ***Pre-processing.*** The pre-processing plays an important role in the natural language processing task. We first split the words using Jieba (https://github.com/fxsjy/jieba). We then unify words with the same meaning, such as unifying "Shanghai City" to "Shanghai"; and we filter out stop words and sparse words. This paper sets the sparsity threshold to 0.01 (Stoltz & Taylor, 2019), namely the proportion of the texts containing a certain word should be greater than 0.01.

(2) ***Generating Word Embeddings.*** For language understanding, we generate word embeddings for the words/phrases in corpus. Word embeddings are the mapping of words onto a numerical vector space supposed to preserve the semantic and syntactic similarities between them (Qian et al., 2019). To capture semantic information accurately, we apply a Bidirectional Encoder Representations from Transformers (BERT) model. BERT is bidirectional variant of transformer networks trained to jointly predict a masked word from its context (Devlin et al., 2018). This paper selects a pre-trained BERT model, *BERT-base, Chinese*. Due to the small scale of our corpus, we apply this model without fine-tuning to prevent over-fitting.

(3) ***Clustering Word Embedding via K-means.*** Based on the results of word embeddings, we put the words with similar meanings into one cluster by the advanced algorithm K-means++ (Arthur & Vassilvitskii, 2006). K-means++ solves the initial clustering centers selection problem of K-means algorithm, an efficient and simple method (Li & Wang, 2022), to avoid the problem of local optimization. Before using the K-means++ algorithm, the important model parameter K (cluster number) must be determined. The silhouette coefficient, a well-known measure of clustering quality, has been widely used in the related research to determine the optimal cluster number (Dinh et al., 2019) and considers the intra-cluster and inter-cluster distances. Over different K, the greater the silhouette coefficient value, the better the performance of clustering result. Silhouette coefficient $s(i)$ for a word embedding $i \in I$ is defined as,

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \# \tag{1}$$

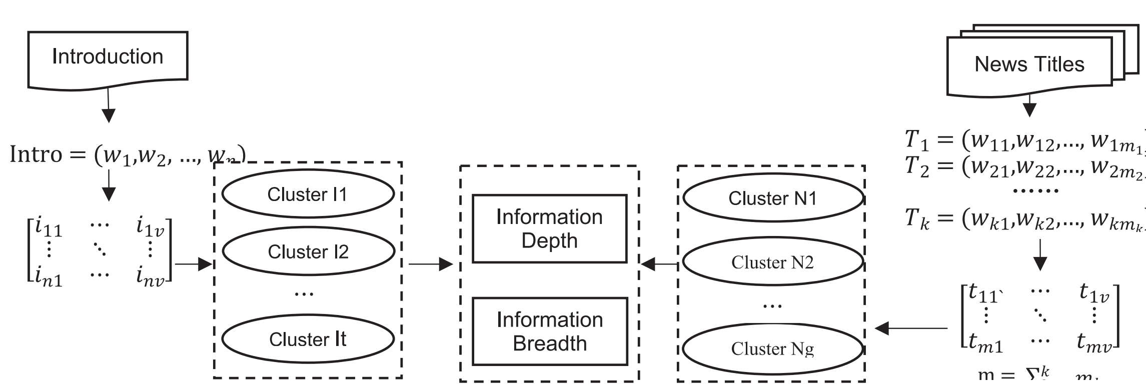

Fig. 2. The steps of constructing information breadth and information depth.

Table 2
Statistics of variables.

	No.	Variable	Summary statisticsMin.	Mean	Max.	S.D.
Basic inform-ation	1	Current ratio (%)	0.120	5.5080	241.400	11.280
	2	Debt asset ratio (%)	0.390	33.080	305.130	22.235
	3	Receivables turnover ratio (years)	0	34.283	18,000	504.244
	4	Inventory turnover ratio (years)	0	171.170	36,000	1,628.724
	5	Total assets turnover (years)	0	0.995	21.320	1.114
	6	Operating profit ratio (%)	–907.37	41.830	99.470	42.753
	7	Rate of return on common stockholders' equity (%)	–32	–0.206	1.714	1.736
	8	Return on total assets ratio (%)	–319.480	0.547	111.760	27.901
	9	Registered capital (ten thousand RMB)	10	4,941	268,572	10,702.210
	10	The number of employees	1	179.500	9,465	438.377
	11	Age (years)	2	10.910	28	4.432
	12	The number of patents	0	94.500	999	106.345
	13	Regions	{The West (78.71%); The Central (10%);The East (8.29%); The Northeast (3%)}			
Website inform-ation	14	Whether an enterprise has built an official website (*BOW*)	{0: the enterprise hasn't built official website (15.36%); 1: the enterprise has built official website (84.64%)}			
	15	Whether it has a WeChat version	{0: not having (40.57%); 1: having (59.43%)}			
	16	Whether it has a foreign language version	{0: not having (67.51%); 1: having (32.49%)}			
	17	The number of videos presented	0	0.926	180	6.688
	18	The number of honors presented	0	14.018	195	19.359
	19	Whether it has product presentation	{0: not having (8.19%); 1: having (91.81%)}			
	20	Whether it has a search engine	{0: not having (67.09%); 1: having (32.91%)}			
	21	The number of external links	0	2.489	62	7.329
	22	The methods offered to interact (telephone number, email address)	0	1.630	2	0.517
	23	Whether the design has a navigation	{0: not having (99.16%); 1: having (0.84%)}			

Notes: Min., minimum; Max., maximum; S.D. refers to standard deviation.

where $a(i)$ is the average Euclidean distance between i and all other word embeddings of the cluster; and $b(i)$ is the minimum of the average Euclidean distances between i and all the word embeddings in other clusters. For a clustering with cluster number $k \in K$, the value of its silhouette coefficient $s(k)$ is the mean of silhouette coefficients of all word embeddings.

(4) *Constructing Features.* After clustering the word embeddings, we calculate the features of information depth and information breadth. Assuming that there are K clusters for the introduction corpus, if all of the words in the introduction of an enterprise appear in M clusters ($M<=K$), the information breadth equals M. The information depth of each cluster of the introduction (formula (2)) is calculated as the summation of the TF-IDF (Term Frequency-Inverse Document Frequency, see formula (3)) value of words that appear in the cluster.

$$\text{InfoDk}_j = \sum \text{TFIDF}_{i,j}\left(w_i \in d_j and w_i \in clusterk\right)\# \quad (2)$$

$$\text{TFIDF}_{i,j} = tf_{i,j} \times idf_i = \frac{n_{i,j}}{\sum_k n_{k,j}} \times log\frac{|D|}{1 + |\{d \in D : w_i \in d\}|}\# \quad (3)$$

where d_j is the introduction document of the enterprise j; w_i is the i-th word or word unit in d_j; $tf_{i,j}$ is the frequency of word w_i in document d_j, which is equal to the number of word w_i in document d_j (namely $n_{i,j}$) divided by the sum of numbers of all words appeared in document d_j; $|D|$ is the number of documents in the collection D; and $|\{d \in D : w_i \in d\}|$ is the number of documents containing the word w_i in D.

TF-IDF is one of the commonly used word weighting schemes in text mining, and the weight of a word increases in proportion to the number of times it appears in the document, but decreases in inverse proportion to the frequency of its appearance in the corpus. We calculate the information breadth and information depth of the news title content in the same manner as the content of the introduction.

3.3. Constructing the dynamics metrics

This paper constructs the dynamics metrics to measure the trend of updating the news released on the official website during a specific observation period. We collect the total number of news stories observed in different years and use a linear regression method to fit these sequences (formula (4)). The trend of updating the news is growth. As a result, for an enterprise, we get the slope k and intercept b for its news update trend. The k and b are the dynamics metrics.

$$Y_t = kX_t + b\# \quad (4)$$

where t is the year in a specific observation period. For an enterprise, Y_t is the total number of news stories observed on the official website in year t. X is the index list of the years by ascending order, and X_t is the index of year t, $X_t \in (0, 1, 2, 3\cdots)$.

3.4. Selecting features

Feature selection focuses on selecting a subset of features from the input data in order to reduce the noise or irrelevant features, avoid overfitting, and improve the predictor performance (Chandrashekar & Sahin, 2014). This paper uses a filter feature selection method because it is independent of classification algorithm.

The correlation-based feature selection (CFS) method is a fully automatic algorithm and assumes that a good feature set contains features that are highly correlated with the class, but uncorrelated with each other (Hall, 1999). The equation for CFS is defined as follows:

$$\text{Merit}_S = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}}\# \quad (5)$$

where Merit_S is the heuristic "merit" of a feature subset S containing k features; $\overline{r_{cf}}$ is the average feature-class correlation; $\overline{r_{ff}}$ is the average feature-feature inter-correlation; the correlation is Pearson's correlation. CFS calculates a matrix of feature-class and feature-feature correlations, and then searches the feature subset space using a best first search strategy based on the value of metric (formula (5)). The feature subset with the highest metric is returned when the search terminates (Karegowda et al., 2010).

Table 3
Predictive performance of models with the feature *BOW*.

Feature set	Methods	Measures						
		AUC	KS	H	Accuracy	Precision	Recall	F-measure
B	LR	0.845	0.653	0.513	0.829	0.241	0.679	0.352
	SVM	0.828	0.613	0.481	0.831	0.236	0.644	0.341
	XGB	0.876	0.701	0.553	0.886	0.325	0.616	0.420
B+	LR	**0.885(0.000)**	**0.715(0.000)**	**0.576(0.000)**	**0.847(0.000)**	**0.272(0.000)**	**0.717(0.014)**	**0.391(0.000)**
BOW	SVM	**0.883(0.000)**	**0.702(0.000)**	**0.563(0.000)**	**0.843(0.000)**	**0.270(0.000)**	**0.745(0.000)**	**0.392(0.000)**
	XGB	**0.898(0.000)**	**0.737(0.000)**	**0.595(0.000)**	0.880(0.006)	0.321(0.409)	**0.665(0.004)**	0.428(0.550)

Notes: Sample size is 1,400; *B* refers to basic features; p-values of non-parametric paired Wilcoxon test are shown in the parentheses; values of seven measures are represented in bold when they are significantly improved, $p < 0.05$ (similarly hereinafter).

Table 4
Predictive performance of models with the design-based features.

Feature set	Methods	Measures						
		AUC	KS	H	Accuracy	Precision	Recall	F-measure
B	LR	0.871	0.745	0.613	0.862	0.205	0.700	0.308
	SVM	0.870	0.734	0.616	0.892	0.242	0.631	0.338
	XGB	0.888	0.768	0.633	0.868	0.212	0.705	0.317
B +	LR	0.857(0.879)	0.728(0.331)	0.600(0.407)	**0.878(0.000)**	**0.225(0.003)**	0.671(0.160)	**0.326(0.033)**
design-based features	SVM	0.868(0.493)	0.721(0.299)	0.597(0.118)	0.877(0.000)	0.216(0.000)	0.644(0.500)	0.313(0.000)
	XGB	0.865(0.648)	0.743(0.377)	0.617(0.856)	**0.874(0.007)**	0.219(0.247)	0.682(0.131)	0.322(0.597)

Notes: Sample size is 1,185.

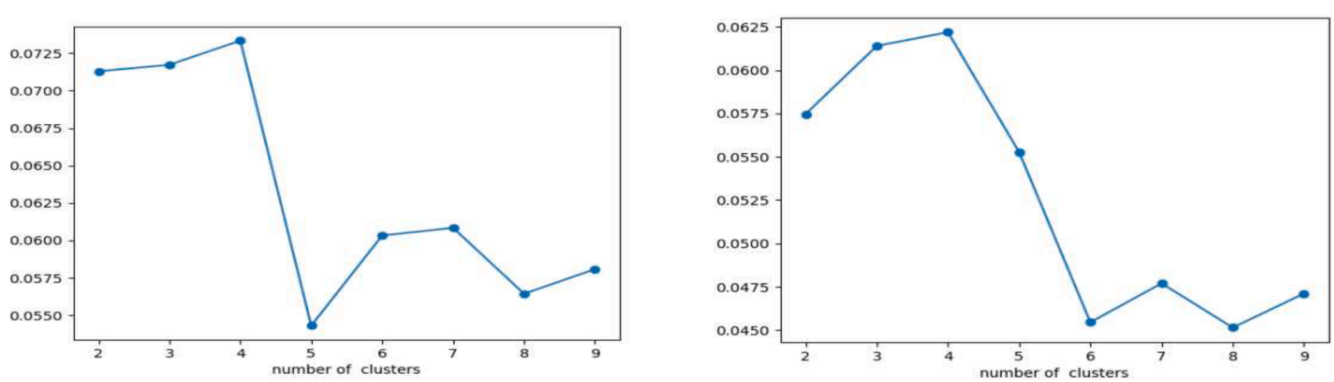

Fig. 3. The value of the silhouette coefficient for K-means++ (introduction corpus).

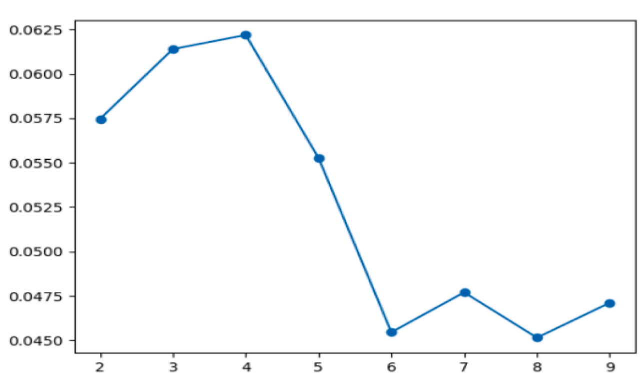

Fig. 4. The value of the silhouette coefficient for K-means++ (news title corpus).

Table 5
Statistics of the selected variables.

No.	Variable	Summary statistics			
		Min.	Mean	Max	S.D.
1	*IntroB*	0	3.879	4	0.608
2	*NewsB*	0	2.441	4	1.920
3	*IntroD4*	0	1.404	3.839	0.706
4	*NewsD2*	0	1.947	9.903	1.941
5	*Dynamic_k*	0	8.091	361.3	19.673
6	*Dynamic_b*	–54.7	3.983	390.3	22.231

3.5. Evaluating predictive performance

We select three classic methods, namely logistic regression (LR), support vector machine (SVM), and eXtreme Gradient Boosting (XGB), to examine the effect of features. And we use seven standard measures to measure the predictive performance of the models. The area under the receiver operating characteristic curve (AUC) of a model is equivalent to the probability that the model will rank a randomly chosen default

Table 6
Predictive performance of models with information breadth and information depth.

Aspect	Feature set	Methods	Measures						
			AUC	KS	H	Accuracy	Precision	Recall	F-measure
	B	LR	0.871	0.745	0.613	0.862	0.205	0.700	0.308
		SVM	0.870	0.734	0.616	0.892	0.242	0.631	0.338
		XGB	0.888	0.768	0.633	0.868	0.212	0.705	0.317
Information breadth	*B* + *IntroB*	LR	0.868	0.739	0.611	**0.870**	**0.212**	0.697	**0.318**
			(0.427)	(0.251)	(0.610)	**(0.038)**	**(0.029)**	(0.842)	**(0.039)**
		SVM	0.867	0.735	0.614	0.891	0.244	0.637	0.341
			(0.338)	(0.913)	(0.256)	(0.475)	(0.804)	(0.504)	(0.931)
		XGB	**0.890**	0.770	0.639	0.871	0.217		0.324
			(0.032)	(0.613)	(0.454)	(0.056)	(0.084)	712(0.425)	(0.102)
	B + *NewsB*	LR	**0.874**	0.750	0.616	**0.869**	**0.214**	0.689	0.316
			(0.041)	(0.241)	(0.567)	**(0.003)**	**(0.020)**	(0.458)	(0.070)
		SVM	**0.877**	0.745	0.618	0.879	0.226	**0.651**	0.324
			(0.025)	(0.124)	(0.735)	(0.000)	(0.010)	**(0.035)**	(0.041)
		XGB	0.889	**0.775**	0.637	**0.872**	0.217	0.700	0.321
			(0.103)	**(0.050)**	(0.309)	**(0.009)**	(0.168)	(0.521)	(0.225)
Information depth	*B* + *IntroD4*	LR	**0.886**	**0.764**	**0.638**	0.868	0.213	0.711	0.317
			(0.007)	**(0.012)**	**(0.004)**	(0.246)	(0.083)	(0.457)	(0.122)
		SVM	**0.884**	**0.760**	**0.645**	0.892	0.245	**0.663**	0.345
			(0.000)	**(0.000)**	**(0.000)**	(0.689)	(0.733)	**(0.004)**	(0.253)
		XGB	**0.899**	**0.778**	**0.657**	**0.913**	**0.286**	**0.625**	**0.378**
			(0.026)	**(0.049)**	**(0.012)**	**(0.000)**	**(0.000)**	**(0.005)**	**(0.000)**
	B + *NewsD2*	LR	**0.885**	**0.773**	0.633	**0.875**	**0.220**	0.692	**0.324**
			(0.007)	**(0.006)**	(0.097)	**(0.000)**	**(0.003)**	(0.794)	**(0.014)**
		SVM	**0.893**	**0.772**	**0.647**	**0.899**	**0.257**	0.644	**0.353**
			(0.000)	**(0.000)**	**(0.001)**	**(0.002)**	**(0.016)**	(0.419)	**(0.017)**
		XGB	**0.905**	**0.791**	**0.673**	**0.915**	**0.297**	**0.618**	**0.384**
			(0.003)	**(0.000)**	**(0.017)**	**(0.000)**	**(0.000)**	**(0.001)**	**(0.000)**
Dynamics metrics	*B* + *Dynamic_k* + *Dynamic_b*	LR	**0.885**	**0.762**	**0.636**	**0.874**	**0.219**	0.703	**0.324**
			(0.005)	**(0.034)**	**(0.033)**	**(0.000)**	**(0.012)**	(0.687)	**(0.041)**
		SVM	**0.894**	**0.765**	**0.647**	**0.908**	**0.275**	0.621	**0.365**
			(0.000)	**(0.000)**	**(0.000)**	**(0.000)**	**(0.000)**	(0.201)	**(0.000)**
		XGB	**0.906**	**0.791**	**0.671**	**0.916**	**0.295**	**0.627**	**0.388**
			(0.001)	**(0.000)**	**(0.018)**	**(0.000)**	**(0.000)**	**(0.003)**	**(0.000)**
ALL	*B* + *IntroD4* +*NewsD2* + *Dynamic_k* + *Dynamic_b*	LR	**0.894**	**0.787**	**0.657**	**0.880**	**0.229**	0.697	**0.336**
			(0.001)	**(0.001)**	**(0.002)**	**(0.000)**	**(0.000)**	(0.931)	**(0.001)**
		SVM	**0.903**	**0.789**	**0.673**	**0.900**	**0.266**	**0.685**	**0.371**
			(0.000)	**(0.000)**	**(0.000)**	**(0.027)**	**(0.004)**	**(0.001)**	**(0.001)**
		XGB	**0.908**	**0.799**	**0.687**	**0.915**	**0.302**	**0.643**	**0.394**
			(0.000)	**(0.000)**	**(0.000)**	**(0.000)**	**(0.000)**	**(0.008)**	**(0.000)**

Notes: Sample size is 1,185.

Table 7
The summarization of the default rate of SMEs with and without official websites.

	Enterprise with official website	Enterprise without official website
Number of SMEs	1,185	215
Number of defaulted SMEs	53	44
Default rate	4.47%	20.47%

observation higher than a randomly chosen non-default observation (Fawcett, 2006). The Kolmogorov-Smirnov statistic (KS) is the maximum difference between the cumulative score distributions of default and non-default observations, and measures the accuracy relative to a single reference point (Lessmann et al., 2015). The H measure (H) avoids the deficiency of AUC that it uses different misclassification cost distributions for different classifier, by fixing a preset beta distribution for classification cost (Hand & Anagnostopoulos, 2014). The accuracy is the proportion of correct predictions (both true positive instances and true negative instances) among all instances. The precision is the percentage of true positive instances among all instances that are predicted to be positive instances, and the recall is the rate of true positive instances among all actually positive instances. Since the precision and recall are contradictory, we use the F-measure, which is the harmonic mean of the two measures (De Weerdt et al., 2011). The higher the AUC, KS, H, accuracy, precision, recall, and F-measure values, the greater the predictive performance of a model.

To estimate the predictive performance of each model, we perform ten independent ten-fold cross validations, resulting in 100 values of performance estimates. Furthermore, we use a non-parametric paired statistic test, Wilcoxon signed-rank test (Wilcoxon, 1992), to examine whether adding features extracted from official website information in the models significantly improved the performance over the baseline model.

4. Empirical evaluation

4.1. Data

We collected a dataset from a commercial bank in the Anhui province of China. Considering that different types of industries may be exposed

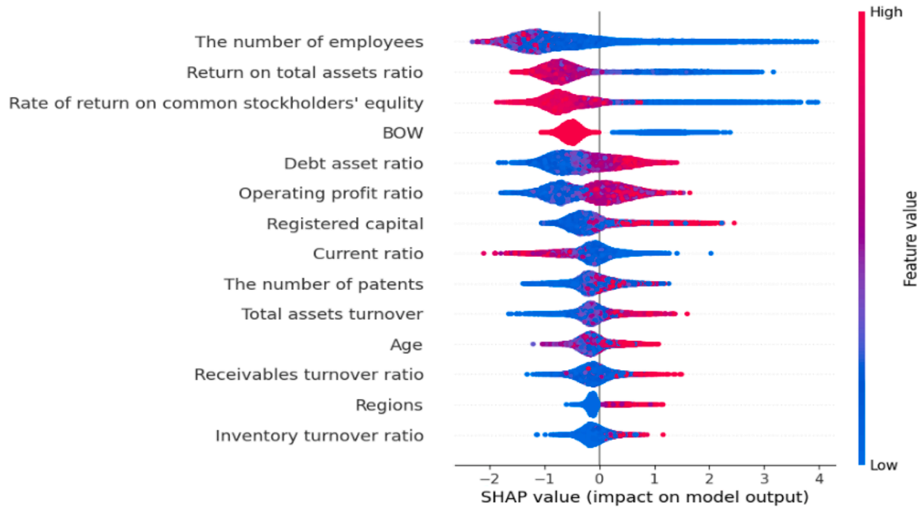

Fig. 5. The SHAP summary plot of the XGB model with feature *BOW*.

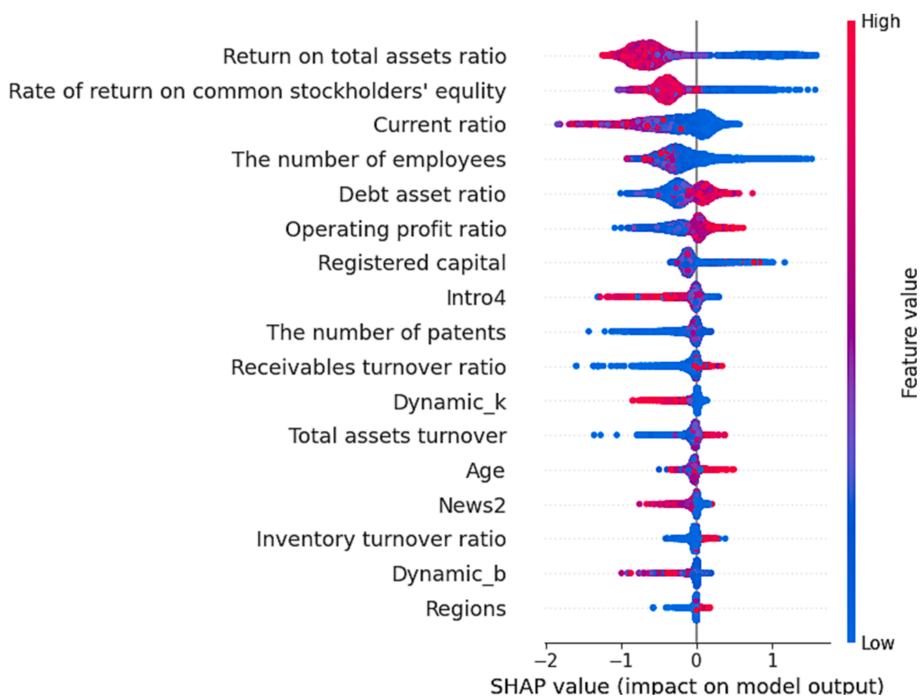

Fig. 6. The SHAP summary plot for the XGB model with effective features.

to different credit risks, we selected the SMEs in the same industry, the software and information technology services industry. The data samples cover 1,400 loan listings that applied for a 12-month loan in December 2017 (ending in December 2018). Our dataset consists of the credit loan records and financial ratios (see Table 2, No. 1 to No. 9) of the SMEs, from one year of financial data before the loan application date. We collected the firm-specific features (see Table 2, No. 10 to No. 13; Mayr et al., 2017) of these SMEs from the QiChacha website (https://www.qcc.com). The financial ratios and firm-specific features are called the "basic features" hereafter. Table 2 shows the statistics of the basic information and official website information, except for the content-based information.

To study the effect of official website information on predicting the credit risk of SMEs, we collected one important dummy variable that indicates whether an enterprise has built an official website, from a government website, https://beian.miit.gov.cn, which is used to organize the ICP/IP addresses registered in China. If an official website was built before the loan application date, the dummy variable of the enterprise is equal to one; otherwise it is 0. We collected the links to the official websites—1,185 in all (there is a one-to-one correspondence between website links and enterprises). By opening these website links, we manually collected the design-based features and clawed the content-based information using codes.

The dependent variable is a binary variable whose value is one if the SME defaulted, and zero if the SME did not default. Following the rules adopted by the bank, we defined a default event as occurring when the payment of a loan is past due over 90 days. There were 1,303 non-default loan observations (positive instances) and 97 default loan observations (negative instances). This is an imbalanced dataset and the imbalanced rate is 13.433 (the number of positive instances/ the number of negative instances). To solve this problem, we employed a popular over-sampling method, the Synthetic Minority Over-sampling

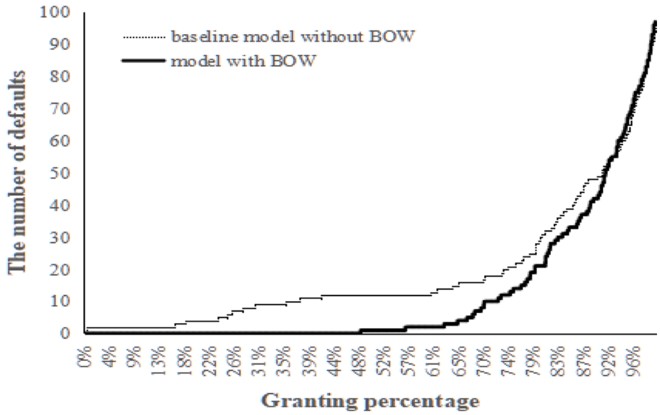

Fig. 7-1. The granting performance of models with and without feature *BOW*.

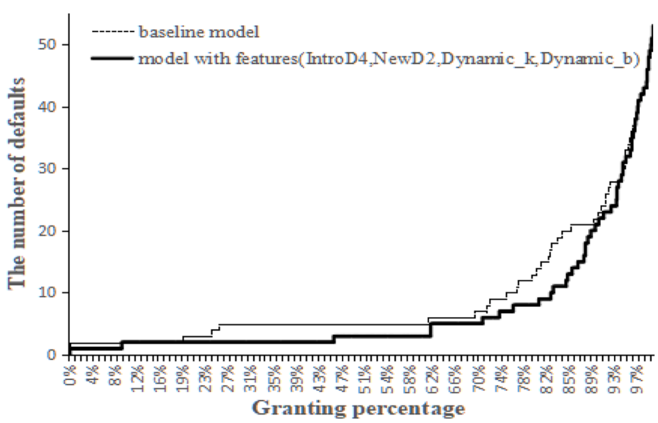

Fig. 7-2. The granting performance of models with and without features (namely *IntroD4, NewsD2, Dynamic_k, and Dynamic_b*).

Technique (SMOTE, Chawla et al., 2002), to resample the training set when training a prediction model, and adjust the imbalance rate of a training set to 1. We also built Generalised Extreme Value (GEV) regression models and models without SMOTE, then compared their predictive performance (see A.ppendix A).

4.2. *The differences between the credit risk of SMEs with or without official websites*

To examine the differences between the credit risk of SMEs with or without official websites, we constructed a feature: whether an enterprise has built an official website (*BOW*) and evaluated its predictive power. The value of feature *BOW* is one if a SME has built an official website; otherwise it is zero. We added the feature *BOW* into prediction models (LR, SVM, and XGB) and compared the results with the baseline models with basic features only (see Table 3).

The predictive performance of the models with feature *BOW* are significantly improved compared with the corresponding baseline models in all seven measures. Thus, *BOW* is a feature with high predictive power. The result demonstrates that there are indeed differences between the credit risk of SMEs with or without official websites (**RQ1**).

In addition, to compare the predictive power of the feature *BOW*, we considered the social network accounts of SMEs and constructed the corresponding social networks features. A.ppendix I shows the comparison of the predictive performance.

4.3. *The effect of design-based information*

From the perspective of prediction, the following evaluations provide a more detailed solution to identify SMEs with high credit risk from loan applicants who have built official websites. We studied which kinds of information posted on the official website will be effective. In our dataset, 1,185 SMEs built official websites before the loan application date.

To examine the effectiveness of integrating design-based information into the credit risk evaluation of SMEs, we added the design-based features into prediction models, and compared them with corresponding baseline models adding basic features only. Table 4 presents the predictive performances of different models, and reports that adding the design-based features into models cannot improve the predictive performance significantly (**RQ2**).

We considered that the design-based information was "frozen" once the website was built and therefore could not reflect the status of the enterprise at the moment of loan application. Additionally, SMEs usually hire professional website construction firms to build their official websites. "Bad-credit" enterprises can acquire the same quality of design-based information as "good-credit" enterprises, at the same cost. Therefore, this kind of information cannot mitigate the information asymmetry between SMEs and financial institutions.

4.4. *The effect of content-based information*

Content-based information includes the introduction to the SME and the news titles it has shared; the observation period of the news was from 2014 to 2017 (A.ppendix B shows more analysis details). The raw content was pre-processed by splitting words and removing stop words and sparse words. Then we trained each word into a vector using BERT model. To examine the vectors' quality, we used other two Word2vec algorithms. See A.ppendix C for results comparison.

For the corpus of introductions and news titles, we clustered the vectors respectively, using the K-means++ algorithm. To determine the optimal cluster number for each corpus, we used the silhouette coefficient to validate the performance of the clustering results. Fig. 3 and Fig. 4 present the values of the silhouette coefficient under different clustering numbers. As a result, the optimal cluster numbers of the collection of introduction texts and of news titles are both four. A.ppendix D presents the word lists of the four clusters. To validate the robust of optimal cluster number, we employed other two clustering quality metrics: Calinski-Harabasz Index and Davies-Bouldin Index (see A.ppendix H).

Based on the clusters, for an enterprise, we calculated the features about information depth and information breadth of the introduction content and news-title content. The features *IntroB* and *NewsB* refer to the information breadth of the introduction content and that of the news-title content, respectively. The features about the information depth of the introduction content are *IntroD1* to *IntroD4* and those of the news-title content are *NewsD1* to *NewsD4*. To remove the irrelevant and redundant features, we used a feature selection method, correlation-based feature selection (CFS), to select the best feature subset (see A.ppendix E). We found that *IntroD4* and *NewsD2* are the effective features. In addition, for each enterprise, we also fit its trend of updating the news content during our observation period using linear regression. The slope and intercept are considered as the dynamics metrics, namely *Dynamic_k* and *Dynamic_b*. Table 5 presents the statistics of information breadth, information depth and dynamics metrics.

To examine the effectiveness of information breadth, we added the two features *IntroB* and *NewsB* into prediction models, respectively. The results report that the information breadth of content-based information is not effective in predicting the credit risk of SMEs (see Table 6).

We examined the effectiveness of information depth by adding the features, *IntroD4* and *NewsD2*, into the prediction models. In Table 6, compared with the baseline models, the predictive performance of

Table A.1
Predictive performance of models without SMOTE method.

Aspect	Feature set	Methods	Measures						
			AUC	KS	H	Accuracy	Precision	Recall	F-measure
	B	LR	0.877	0.741	0.614	**0.954**	**0.446**	0.234	0.280
			(0.152)	(0.125)	(0.301)	**(0.000)**	**(0.000)**	(0.000)	(0.210)
		SVM	0.845	0.687	0.582	0.383	0.068	**0.944**	0.125
			(0.000)	(0.000)	(0.005)	(0.000)	(0.000)	**(0.000)**	(0.000)
		XGB	0.884	0.752	0.630	**0.955**	**0.463**	0.236	0.287
			(0.026)	(0.015)	(0.331)	**(0.000)**	**(0.000)**	(0.000)	(0.227)
Information breadth	*B* + *IntroB*	LR	0.878	0.745	0.619	**0.953**	**0.415**	0.233	0.274
			(0.766)	(0.588)	(0.842)	**(0.000)**	**(0.000)**	(0.000)	(0.061)
		SVM	0.853	0.702	0.593	0.412	0.070	**0.935**	0.129
			(0.005)	(0.007)	(0.065)	(0.000)	(0.000)	**(0.000)**	(0.000)
		XGB	0.886	0.758	0.636	**0.954**	**0.445**	0.237	0.284
			(0.020)	(0.098)	(0.431)	**(0.000)**	**(0.000)**	(0.000)	(0.090)
	B + *NewsB*	LR	0.881	0.750	0.620	**0.952**	**0.428**	0.217	0.258
			(0.182)	(0.382)	(0.487)	**(0.000)**	**(0.000)**	(0.000)	(0.018)
		SVM	0.841	0.680	0.577	0.374	0.066	**0.934**	0.122
			(0.000)	(0.000)	(0.005)	(0.000)	(0.000)	**(0.000)**	(0.000)
		XGB	0.888	0.760	0.633	**0.954**	**0.432**	0.219	0.264
			(0.048)	(0.013)	(0.233)	**(0.000)**	**(0.000)**	(0.000)	(0.018)
Information depth	*B* + *IntroD4*	LR	0.885	0.761	0.635	**0.954**	**0.435**	0.235	0.282
			(0.099)	(0.544)	(0.474)	**(0.000)**	**(0.000)**	(0.000)	(0.156)
		SVM	0.857	0.698	0.595	0.283	0.059	**0.953**	0.111
			(0.000)	(0.000)	(0.001)	(0.000)	(0.000)	**(0.000)**	(0.000)
		XGB	0.894	0.769	0.666	**0.951**	**0.452**	0.267	0.303
			(0.625)	(0.684)	(0.262)	**(0.000)**	**(0.000)**	(0.000)	(0.000)
	B + *NewsD2*	LR	0.884	0.760	0.631	**0.951**	**0.363**	0.180	0.215
			(0.031)	(0.025)	(0.178)	**(0.000)**	**(0.000)**	(0.000)	(0.000)
		SVM	0.859	0.704	0.598	0.290	0.059	**0.955**	0.112
			(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	**(0.000)**	(0.000)
		XGB	0.892	0.769	0.670	**0.952**	**0.440**	0.274	0.309
			(0.033)	(0.141)	(0.087)	**(0.000)**	**(0.000)**	(0.000)	(0.000)
Dynamics metrics	*B* + *Dynamic_k* + *Dynamic_b*	LR	0.890	0.765	0.638	**0.951**	**0.391**	0.200	0.243
			(0.845)	(0.857)	(0.831)	**(0.000)**	**(0.000)**	(0.000)	(0.001)
		SVM	0.869	0.717	0.608	0.331	0.063	**0.957**	0.119
			(0.000)	(0.000)	(0.002)	(0.000)	(0.000)	**(0.000)**	(0.000)
		XGB	0.894	0.769	0.662	**0.951**	**0.433**	0.272	0.303
			(0.023)	(0.048)	(0.620)	**(0.000)**	**(0.000)**	(0.000)	(0.000)
ALL	*B* + *IntroD4* +*NewsD2* + *Dynamic_k* + *Dynamic_b*	LR	0.895	0.778	0.649	**0.950**	**0.358**	0.193	0.227
			(0.574)	(0.276)	(0.220)	**(0.000)**	**(0.001)**	(0.000)	(0.000)
		SVM	0.871	0.722	0.613	0.264	0.098	**0.910**	0.141
			(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	**(0.000)**	(0.000)
		XGB	0.904	0.789	0.678	**0.951**	**0.426**	0.285	0.317
			(0.124)	(0.028)	(0.196)	**(0.000)**	**(0.000)**	(0.000)	(0.000)

Notes: Sample size is 1,185; values of seven measures are represented in bold when they are significantly improved, comparing with the corresponding values in Table 6 ($p < 0.05$).

Table A.2
Predictive performance of GEV regression models.

Feature set	Measures						
	AUC	KS	H	Accuracy	Precision	Recall	F-measure
B	0.830	0.666	0.544	0.953	0.444	0.244	0.286
B + IntroB	0.834	0.676	0.549	0.952	0.408	0.230	0.270
B + NewsB	0.828	0.668	0.543	0.953	0.451	0.244	0.288
B + IntroD4	0.834	0.684	0.558	0.952	0.416	0.243	0.279
B + NewsD2	0.839	0.688	0.559	0.952	0.436	0.253	0.292
B + *Dynamic_k* + *Dynamic_b*	0.834	0.686	0.558	0.952	0.429	0.233	0.275
ALL	0.846	0.707	0.569	0.952	0.442	0.240	0.282

corresponding models integrating feature *IntroD4* (or feature *NewsD2*) are significantly improved respectively. The results indicate that the feature *IntroD4* and the feature *NewsD2* are strong predictors at predicting the credit risk of SMEs. Similarity, Table 6 reports the predictive performance of models with the dynamics metrics are significantly improved ($p < 0.05$).

Finally, we integrated all the effective features extracted from the content-based information into the prediction models, resulting in the best predictive performance (see Table 6). All the features added into the models are moderately correlated to each other (see A.ppendix F). We

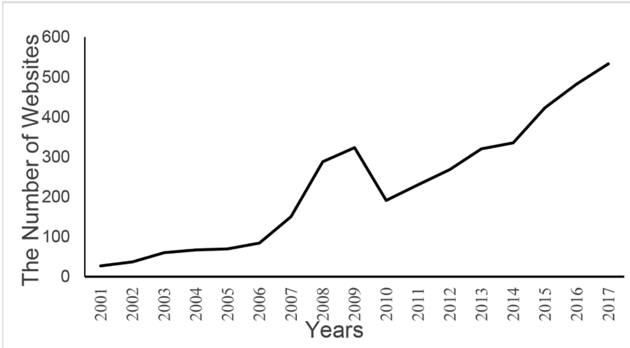

Fig. B1. The number of registered websites.

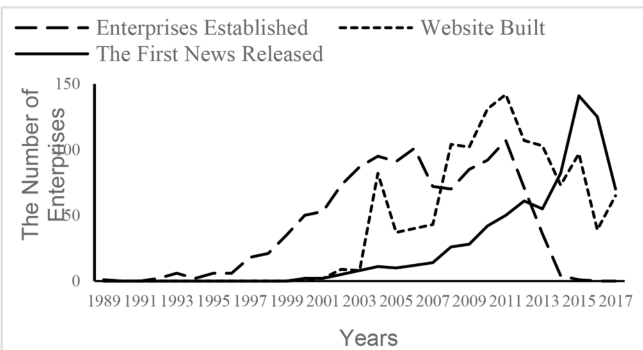

Fig. B2. Statistics of enterprises and websites.

Table C.1
The results of the CFS method for features constructed by Skip-Gram method.

No.	Feature sets	CFS_merits	Feature sets	CFS_merits
1	*SkipIntroD1*	0.035	*SkipNewsD1*	0.049
2	*SkipIntroD2*	0.081	*SkipNewsD2*	0.029
3	*SkipIntroD3*	0.011	*SkipNewsD3*	0.047
4	*SkipIntroD4*	0.071	*SkipNewsD4*	0.025
5	*SkipIntroD2, SkipIntroD1*	0.070	*SkipNewsD1, SkipNewsD2*	0.041
6	*SkipIntroD2, SkipIntroD3*	0.057	***SkipNewsD1, SkipNewsD3***	**0.050**
7	***SkipIntroD2, SkipIntroD4***	**0.091**	*SkipNewsD1, SkipNewsD4*	0.039
8	*SkipIntroD2, SkipIntroD4, SkipIntroD1*	0.081	*SkipNewsD1, SkipNewsD3, SkipNewsD2*	0.045
9	*SkipIntroD2, SkipIntroD4, SkipIntroD3*	0.073	*SkipNewsD1, SkipNewsD3, SkipNewsD4*	0.043

Table C.2
The results of the CFS method for features constructed by CBoW method.

No.	Feature sets	CFS_merits	Feature sets	CFS_merits
1	*CBoWIntroD1*	0.089	*CBoWNewsD1*	0.034
2	*CBoWIntroD2*	0.046	*CBoWNewsD2*	0.034
3	*CBoWIntroD3*	0.015	*CBoWNewsD3*	0.052
4	*CBoWIntroD4*	0.026	*CBoWNewsD4*	0.018
5	***CBoWIntroD1, CBoWIntroD2***	**0.093**	***CBoWNewsD3, CBoWNewsD1***	**0.053**
6	*CBoWIntroD1, CBoWIntroD3*	0.073	*CBoWNewsD3, CBoWNewsD2*	0.048
7	*CBoWIntroD1, CBoWIntroD4*	0.077	*CBoWNewsD3, CBoWNewsD3*	0.040
8	*CBoWIntroD1, CBoWIntroD2, CBoWIntroD3*	0.081	*CBoWNewsD3, CBoWNewsD1, CBoWNewsD2*	0.047
9	*CBoWIntroD1, CBoWIntroD2, CBoWIntroD4*	0.085	*CBoWNewsD3, CBoWNewsD1, CBoWNewsD4*	0.044

concluded that content-based information has an effect on the credit risk prediction of SMEs, and can significantly improve the performance of prediction model (**RQ2**). The features *IntroD4*, *NewsD2*, *Dynamic_k*, and *Dynamic_b* can be used as signals to mitigate the information asymmetry when financial institutions evaluate the credit risk of SMEs (**RQ3**).

5. Discussion

5.1. The predictive mechanism of effective features

This paper conducted the above experiments to solve research questions **RQ1**, **RQ2**, and **RQ3**. We identified the effective features, *BOW*, *IntroD4*, *NewsD2*, *Dynamic_k*, and *Dynamic_b*. This section further explores the predictive mechanism of these effective features. In this paper, models with XGB algorithm provide the best predictive performance in different tasks (see Table 3 and Table 6). However, these models, as a black box prediction, lack the interpretability to gather clues for making loan decisions (Bussmann et al., 2021).

To overcome this problem, we employed the SHAP (Shapley Additive exPlanations; Lundberg & Lee, 2017) to interpret the predictive models and analyze the predictive mechanism of each effective feature. SHAP is based on game theory and local explanations, and it offers a means to estimate the contribution of each feature (Parsa et al., 2020). For each predicted sample, the SHAP value is the value assigned to each feature in the sample (Yang et al., 2021), which can be considered as the predictive ability of the feature on the sample.

5.1.1. SMEs with official websites have lower credit risk

To explore the relationship between the feature *BOW* and the credit risk of SMEs, we first calculated the default rates of the two groups. Table 7 shows that the default rate of the SMEs with official websites is much lower than that of SMEs without official websites. Further, we used the SHAP method to interpret the XGB model with feature *BOW* and observe how the feature *BOW* play a role in the predictive model. Fig. 5 displays the SHAP summary plot that orders the features based on their importance to affect default prediction. The horizontal position of the dot is the impact of the feature on the prediction, and the color of the dot represents the value of that feature for the prediction. In Fig. 5, as the feature *BOW* decreased (the dot color transition from red to blue), the probability of default increased (SHAP values change from negative to positive). Thus, we consider that the SMEs with official websites have a lower credit risk.

5.1.2. The analysis of content-based information

Fig. 6 displays the SHAP summary plot for the XGB model with the effective features constructed by content-based information. The results show that the features, *IntroD4* and *NewsD2*, are negatively correlated with the probability of default. We conclude that the SMEs are less likely to default if they describe the technology they use in running their business in detail in their introduction; and if they frequently release news on their website about their business.

For the dynamic metrics, *Dynamic_k* and *Dynamic_b*, they are negatively correlated with the probability of default. We conclude that SMEs who frequently publish news on their official websites and/or have a large base of news, in a specific observation period, have a lower default probability.

5.2. The economic benefit of official website information

After analyzing the value of official website information from the perspective of predictive performance, we further discuss the economic

Table C.3
Predictive performance of models with features constructed by Skip-Gram method.

Aspect	Feature set	Methods	Measures						
			AUC	KS	H	Accuracy	Precision	Recall	F-measure
Information breadth	*B+ SkipIntroB*	LR	0.863	0.737	0.600	0.860	0.196	0.666	0.293
		SVM	0.861	0.720	0.589	0.884	0.222	0.622	0.318
		XGB	0.880	0.757	0.636	0.861	0.208	0.729	0.315
	B+ SkipNewsB	LR	0.867	0.747	0.609	0.865	0.203	0.681	0.304
		SVM	0.870	0.736	0.607	0.875	0.212	0.637	0.307
		XGB	0.868	0.739	0.615	0.864	0.207	0.699	0.310
Information depth	*B + SkipIntroD2 + SkipIntroD4*	LR	0.868	0.752	0.612	0.866	0.206	0.700	0.309
		SVM	0.870	0.730	0.607	0.879	0.220	0.655	0.319
		XGB	0.877	0.756	0.633	0.856	0.205	0.739	0.312
	B + SkipNewsD1 + SkipNewsD3	LR	0.865	0.743	0.605	0.863	0.202	0.694	0.304
		SVM	0.867	0.735	0.604	0.873	0.213	0.653	0.310
		XGB	0.868	0.738	0.616	0.859	0.201	0.709	0.304

Table C.4
Predictive performance of models with features constructed by CBoW method.

Aspect	Feature set	Methods	Measures						
			AUC	KS	H	Accuracy	Precision	Recall	F-measure
Information breadth	*B+ CBoWIntroB*	LR	0.859	0.728	0.586	0.863	0.195	0.648	0.291
		SVM	0.857	0.715	0.586	0.885	0.222	0.617	0.316
		XGB	0.873	0.742	0.620	0.855	0.201	0.722	0.305
	B+ CBoWNewsB	LR	0.866	0.745	0.608	0.865	0.205	0.681	0.305
		SVM	0.870	0.733	0.605	0.875	0.212	0.638	0.308
		XGB	0.867	0.736	0.610	0.862	0.204	0.687	0.305
Information depth	*B + CBoWIntroD1 + CBoWIntroD2*	LR	0.863	0.735	0.595	0.870	0.206	0.652	0.304
		SVM	0.868	0.731	0.614	0.894	0.241	0.613	0.336
		XGB	0.871	0.744	0.622	0.855	0.205	0.737	0.312
	B+ CBoWNewsD1+ CBoWNewsD3	LR	0.865	0.746	0.615	0.871	0.214	0.707	0.319
		SVM	0.881	0.751	0.627	0.884	0.231	0.648	0.328
		XGB	0.869	0.746	0.627	0.862	0.204	0.708	0.308

Table D.1
The word list of each cluster of the introduction-text corpus.

No.	Management (cluster1)	word freq.	Development (clsuter2)	word freq.	Product (clsuter3)	word freq.	Technology (clsuter4)	word freq.
1	领先 Leading	205	发展 Development	585	服务 Service	902	高新技术 High and new technology	566
2	产业 Industry	200	创新 Innovation	401	技术 Technology	732	解决方案 Solution	537
3	销售 Selling	200	国内 Home	385	产品 Product	703	大数据 Big data	323
4	国际 International	187	市场 Market	341	行业 Business	693	软件著作权 Software copyright	240
5	集团 Group	164	信息化 Informatization	322	领域 Domain	602	物联网 Internet of things	221
6	先进 Advanced	152	安全 Safety	250	研发 R&D	552	云计算 Cloud computing	192
7	城市 City	150	政府 Government	250	客户 Customer	548	自主研发 R&D independently	188
8	涵盖 Covering	149	中心 Center	220	应用 Application	547	系统集成 System integration	174
9	子公司 Subsidiary	147	积累 Accumulate	205	平台 Platform	526	合作伙伴 Cooperative partner	171
10	国家级 National	144	提升 Promoting	204	专业 Profession	489	软件产品 Software product	163

effect of our findings in practice. To translate the improved predictive performance into financial loss that could be prevented, we introduce the granting performance which refers to the number of defaults under different granting ratios (Wang et al., 2020).

First, we rank the loan applications based on their default probabilities estimated by the prediction model with official website information and the baseline model, respectively. Through a ten-fold cross-validation, each enterprise has an opportunity as test sample and has an estimated default probability. We build the two prediction models based on the XGB algorithm, because the values of evaluation metrics of XGB are the highest under different feature sets. Second, we calculate the number of defaults under different cut-off values of the percentage of applications approved in our dataset (i.e., the granting performance).

In Fig. 7-1., the granting performance of model with feature *BOW* is

Table D.2
The word list of each cluster of the news-title corpus.

No.	Communication (cluster1)	word freq.	Business (cluster2)	word freq.	Development (cluster3)	word freq.	Management (cluster4)	word freq.
1	科技 Technology	300	荣获 Having honor to obtain	396	发展 development	363	成功 success	409
2	大数据 Big data	206	项目 Project	378	创新 Innovation	312	年度 annual	392
3	互联网+ Internet plus	199	平台 Platform	327	顺利 Without a hitch	260	管理 Management	282
4	互联网 Internet	197	技术 Technology	325	未来 Future	241	产业 Industry	205
5	研讨会 Seminar	187	活动 Activity	323	市场 Market	234	董事长 Chairman	203
6	战略合作 Strategic cooperation	184	系统 System	315	安全 Safety	228	集团 Group	192
7	解决方案 Solution	149	服务 Service	309	信息化 Informatization	220	论坛 Forum	192
8	圆满结束 A successful close	134	助力 Assisting	307	圆满 Completeness	218	员工 Staff	186
9	博览会 Exposition	133	行业 Business	292	领导 Leading	217	落幕 Ending	165
10	高峰论坛 Summit	132	产品 Product	290	中心 Center	201	通知 informing	161

Table E.1
The results of the CFS method.

No.	Feature sets	CFS_merits	Feature sets	CFS_merits
1	*IntroD1*	0.051	*NewsD1*	0.047
2	*IntroD2*	0.041	***NewsD2***	**0.057**
3	*IntroD3*	0.023	*NewsD3*	0.033
4	***IntroD4***	**0.096**	*NewsD4*	0.047
5	*IntroD4*, *IntroD1*	0.093	*NewsD2*, *NewsD1*	0.055
6	*IntroD4*, *IntroD2*	0.087	*NewsD2*, *NewsD3*	0.046
7	*IntroD4*, *IntroD3*	0.072	*NewsD2*, *NewsD4*	0.054

Table F.1
The results of multi-collinearity test among all features.

	No.	Variable	Tolerance	VIF
Basic inform-ation	1	Current ratio (%)	0.800	1.251
	2	Debt asset ratio (%)	0.721	1.387
	3	Receivables turnover ratio (years)	0.984	1.017
	4	Inventory turnover ratio (years)	0.858	1.166
	5	Total assets turnover (years)	0.812	1.231
	6	Operating profit ratio (%)	0.864	1.158
	7	Rate of return on common stockholders' equity (%)	0.557	1.795
	8	Return on total assets ratio (%)	0.545	1.836
	9	Registered capital (ten thousand RMB)	0.713	1.403
	10	The number of employees	0.672	1.488
	11	Age (years)	0.896	1.116
	12	The number of patents	0.901	1.110
	13	Regions	0.965	1.036
Website inform-ation	14	IntroB	0.857	1.167
	15	NewsB	0.339	2.952
	16	IntroD4	0.743	1.345
	17	NewsD2	0.327	3.060
	18	Dynamic_k	0.513	1.949
	19	Dynamic_b	0.625	1.601

better than the baseline model when the granting percentage between 0% and 93.2%. It indicates that if a lender approves top N% ($0 < N < 93.2$) applicants ranked by default probabilities estimated by prediction model, the number of defaults of our model will be lower than that of baseline model. And the maximum difference of granting performance of the two models is 12 (the granting ratio is 41.43%), which means that integrating the feature *BOW* into the baseline prediction model can

Table G.1
The results of the CFS method for features constructed by DBSCAN method.

No.	Feature sets	merits	Feature sets	merits
1	*DBSCANIntroD1*	0.069	*DBSCANNewsD1*	0.044
2	*DBSCANIntroD2*	0.027	*DBSCANNewsD2*	0.015
3	*DBSCANIntroD3*	0.059	*DBSCANNewsD3*	0.041
4	*DBSCANIntroD4*	0.024	*DBSCANNewsD4*	0.018
5	*DBSCANIntroD1, DBSCANIntroD2*	0.066	*DBSCANNewsD1, DBSCANNewsD2*	0.041
6	***DBSCANIntroD1, DBSCANIntroD3***	**0.082**	***DBSCANNewsD1, DBSCANNewsD3***	**0.057**
7	*DBSCANIntroD1, DBSCANIntroD4*	0.058	*DBSCANNewsD1, DBSCANNewsD4*	0.037
8	*DBSCANIntroD1, DBSCANIntroD3, DBSCANIntroD2*	0.080	*DBSCANNewsD1, DBSCANNewsD3, DBSCANNewsD2*	0.055
9	*DBSCANIntroD1, DBSCANIntroD3, DBSCANIntroD4*	0.075	*DBSCANNewsD1, DBSCANNewsD3, DBSCANNewsD4*	0.051

Table G.2
The results of the CFS method for features constructed by Mean Shift method.

No.	Feature sets	merits	Feature sets	merits
1	***MeanShiftIntroD1***	**0.069**	***MeanShiftNewsD1***	**0.053**
2	*MeanShiftIntroD2*	0.032	*MeanShiftNewsD2*	0.017
3	*MeanShiftIntroD3*	0.005	*MeanShiftNewsD3*	0.007
4	*MeanShiftIntroD4*	0.014	*MeanShiftNewsD4*	0.010
5	*MeanShiftIntroD5*	0.009	*MeanShiftNewsD5*	0.014
6	*MeanShiftIntroD6*	0.029	*MeanShiftNewsD6*	0.013
7	*MeanShiftIntroD1, MeanShiftIntroD2*	0.065	*MeanShiftNewsD7*	0.010
8	*MeanShiftIntroD1, MeanShiftIntroD3*	0.052	*MeanShiftNewsD1, MeanShiftNewsD2*	0.047
9	*MeanShiftIntroD1, MeanShiftIntroD4*	0.055	*MeanShiftNewsD1, MeanShiftNewsD3*	0.041
10	*MeanShiftIntroD1, MeanShiftIntroD5*	0.053	*MeanShiftNewsD1, MeanShiftNewsD4*	0.042
11	*MeanShiftIntroD1, MeanShiftIntroD6*	0.066	*MeanShiftNewsD1, MeanShiftNewsD5*	0.045
12			*MeanShiftNewsD1, MeanShiftNewsD6*	0.045
13			*MeanShiftNewsD1, MeanShiftNewsD7*	0.043

Table G.3
Predictive performance of models with features constructed by DBSCAN method.

Aspect	Feature set	Methods	Measures						
			AUC	KS	H	Accuracy	Precision	Recall	F-measure
Information breadth	*B+ DBSCANIntroB*	LR	0.874	0.732	0.605	0.954	0.433	0.234	0.278
		SVM	0.862	0.716	0.592	0.879	0.219	0.643	0.317
		XGB	0.877	0.750	0.624	0.870	0.216	0.691	0.320
	B+ DBSCANNewsB	LR	0.874	0.739	0.611	0.953	0.423	0.220	0.264
		SVM	0.860	0.726	0.593	0.890	0.235	0.622	0.329
		XGB	0.873	0.747	0.621	0.870	0.213	0.677	0.315
Information depth	*B + DBSCANIntroD1 + DBSCANIntroD3*	LR	0.875	0.751	0.624	0.954	0.435	0.225	0.272
		SVM	0.866	0.736	0.613	0.879	0.222	0.658	0.322
		XGB	0.876	0.760	0.638	0.870	0.219	0.724	0.329
	B+ DBSCANNewsD1 + DBSCANNewsD3	LR	0.879	0.746	0.620	0.951	0.376	0.205	0.242
		SVM	0.871	0.738	0.604	0.871	0.211	0.664	0.311
		XGB	0.876	0.756	0.625	0.867	0.209	0.697	0.314

Table G.4
Predictive performance of models with features constructed by Mean Shift method.

Aspect	Feature set	Methods	Measures						
			AUC	KS	H	Accuracy	Precision	Recall	F-measure
Information breadth	*B+ MeanShiftIntroB*	LR	0.870	0.729	0.603	0.954	0.432	0.228	0.273
		SVM	0.868	0.724	0.602	0.895	0.247	0.631	0.343
		XGB	0.876	0.742	0.617	0.869	0.213	0.696	0.315
	B+ MeanShiftNewsB	LR	0.874	0.737	0.608	0.953	0.419	0.226	0.268
		SVM	0.860	0.721	0.592	0.871	0.210	0.656	0.307
		XGB	0.879	0.753	0.625	0.871	0.213	0.692	0.317
Information depth	*B + MeanShiftIntroD1*	LR	0.879	0.752	0.621	0.953	0.418	0.223	0.265
		SVM	0.864	0.727	0.601	0.878	0.218	0.644	0.316
		XGB	0.890	0.778	0.651	0.868	0.221	0.734	0.330
	B+ MeanShiftNewsD1	LR	0.880	0.751	0.621	0.952	0.377	0.201	0.238
		SVM	0.866	0.731	0.601	0.872	0.214	0.665	0.313
		XGB	0.883	0.762	0.633	0.869	0.214	0.706	0.319

Table H.1
The evaluation of clustering quality under different cluster numbers.

Corpus	Metric	Cluster number							
		2	3	4	5	6	7	8	9
Introduction corpus	CHI	83.229	107.227	**128.516**	69.499	60.552	53.884	49.205	45.224
	DBI	3.899	3.435	**2.275**	4.220	4.151	4.026	4.100	4.007
News title corpus	CHI	189.318	241.486	**326.147**	158.280	138.499	122.842	110.789	101.351
	DBI	4.558	3.916	**2.391**	4.393	4.393	4.312	4.414	4.345

Note: CHI refers the Calinski-Harabasz Index; and DBI refers to the Davies-Bouldin Index.

identify 12.37% (12/97) more default loans. Further, assuming the average granting (loan) amount is RMB 1,000,000, and the total number of loans approved is 580 (1400 * 41.43%), the total amount of credit loans is RMB 580,000,000. In practice, when a SME defaults on a loan, 30% of the loan amount would be lost on average[1]. Therefore, considering the feature *BOW* can save the bank RMB 3,600,000 (12 * 1,000,000 *30%) on average due to reduced loan defaults.

After using the model with feature *BOW*, lenders can further estimate the credit risk for SMEs with official websites. In Fig. 7-2, the granting performance of model with features *IntroD4*, *NewsD2*, *Dynamic_k*, and *Dynamic_b*, are better than the baseline model when the granting percentage between 0% and 99.16%. Specifically, the maximum difference of granting performance of the two models is 9 when the granting ratio is 84.30%, which indicates that for SMEs with official websites, our model can further identify 16.98% (9/53) more default loans than the baseline model. Thus, considering the features *IntroD4*, *NewsD2*, *Dynamic_k*, and *Dynamic_b*, can save the bank RMB 2,700,000 (9 * 1,000,000 *30%) on average due to reduced loan defaults.

We demonstrate that, from an economic perspective, combining the SMEs' official website information into loan decisions can help financial institutions reduce their financial losses due to loan defaults.

6. Conclusion

This paper examines the effect of official website information as a complement to insufficient financial information on the credit risk evaluation of SMEs by solving three research questions. We further analyze the predictive mechanism of each effective feature by SHAP method. Finally, we evaluate the economic benefit of official website information in the credit risk context by granting performance.

[1] The 30% loss rate is provided by the bank we work with for this paper based on their business experience.

Table I.1
Predictive performance of models with the social network features.

Feature set	Methods	Measures						
		AUC	KS	H	Accuracy	Precision	Recall	F-measure
B	LR	0.845	0.653	0.513	0.829	0.241	0.679	0.352
	SVM	0.828	0.613	0.481	0.831	0.236	0.644	0.341
	XGB	0.876	0.701	0.553	0.886	0.325	0.616	0.420
B+ BOW	LR	0.885	0.715	0.576	0.847	0.272	0.717	0.391
	SVM	0.883	0.702	0.563	0.843	0.270	0.745	0.392
	XGB	0.898	0.737	0.595	0.880	0.321	0.665	0.428
B + Reg_Weibo	LR	0.857	0.665	0.529	0.826	0.243	0.719	0.359
	SVM	0.832	0.619	0.490	0.837	0.248	0.653	0.355
	XGB	0.877	0.703	0.549	0.891	0.329	0.551	0.405
B + Reg_WeChat	LR	0.864	0.670	0.526	0.828	0.248	0.718	0.365
	SVM	0.844	0.639	0.505	0.818	0.234	0.716	0.349
	XGB	0.877	0.709	0.560	0.892	0.337	0.576	0.428

Notes: Sample size is 1,400.

The key findings are summarized as follows. First, whether an enterprise builds an official website is a credible non-financial information to convey the SME's credit risk, and the SMEs with official websites are less likely to default. Second, the design-based information is useless in credit risk evaluation because it is a "freeze-frame" taken when the official website is built. Third, regarding the content-based information, information depth and dynamics metrics have been demonstrated to effectively improve the predictive performance of models. Finally, combining the SMEs' official website information into loan decisions can help financial institutions reduce their financial losses due to loan defaults.

This paper makes significant contributions to academic research. First, this paper contributes to the literature on credit risk evaluation of SMEs by extending to predicting credit risk using non-financial information. To the best of our knowledge, we are the first to investigate the value of information from SMEs' official websites as a complement to financial information in evaluating the credit risk. Second, this paper contributes to the literature on SMEs' official website assessment. We classify the website information into two categories: design-based and content-based information. Furthermore, content-based information is divided into static information and dynamic information. We also develop different information dimensions, including information breadth, information depth, and dynamic metrics, to measure the content-based information quality.

Third, our paper makes a methodology contribution by proposing a text mining framework to identify the effective textual features, especially from official website information. In particular, we demonstrate how to identify the key website information that distinguishes the "bad" borrowers from the "good" ones. The framework, however, is generally applicable to differentiating good-performing SMEs from bad-performing ones for other business decisions. Fourth, we interpret our conclusions after examining the predictive performance. We analyze the predictive mechanism of the effective features based on SHAP method. These predictive mechanisms help us understand the value of official website information in SMEs' credit risk evaluation, and reveal that the information disclosure willingness and quality of an enterprise affect credit risk evaluation.

The implications to practice are threefold. First, our study assists financial institutions in improving the accuracy of credit risk evaluation, avoiding adverse selections, and reducing the financial loss caused by default loans. Although our framework brings a manual burden on collecting information, we can decrease this burden by developing data clawing codes, and collect information automatically. The cost of code development is far below than the economic benefit we improved. Second, our findings provide a reference for SMEs to follow to harness the benefit of having good official website information, helping enterprises highlight their strengths so as to more easily obtain financial support for further development. With the development of information technology, an SME's official website has become an important channel for investors or other interest groups to use to obtain information. Thus, it is vital for an enterprise to disclose meaningful information on its website. Third, the proposed framework provides a technical solution of processing unstructured website information, to guide performance evaluation, product marketing, and other business strategies, in addition to credit evaluation.

As in prior studies, there are many limitations in this paper. First, due to storage and technical limitations, we ignored the visual information—like v.ideos and pictures—posted on the websites. Visual information conveys its meaning vividly and directly. In future work, we intend to explore the effect of visual information of website on the credit risk evaluation of SMEs. Second, we do not consider the issue of fake or inaccurate information, as the SMEs' official websites are under close monitoring and supervision of the relevant government agencies. Future research work can look into this issue as another research direction. Third, more comprehensive experiences using multiple datasets in various industries should be conducted to further validate the generalizability of the proposed framework.

CRediT authorship contribution statement

Cuiqing Jiang: Supervision, Funding acquisition, Conceptualization, Project administration, Resources, Validation. **Chang Yin:** Visualization, Validation, Methodology, Formal analysis, Conceptualization, Data curation, Writing - original draft, Writing - review & editing. **Qian Tang:** Writing – review & editing, Supervision, Conceptualization. **Zhao Wang:** Visualization, Software, Methodology, Funding acquisition.

Declaration of Competing Interest

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

Acknowledgments

This work was funded by the National Natural Science Foundation, China (Grant Nos. 71731005, 72101073) and the Anhui Provincial Natural Science Foundation, China (Grant No. 2108085MG234).

Appendix A:. Results comparison for data imbalanced method

This paper used a non-parametric test (full pairwise), Wilcoxon test, to test the statistical significance of predictive performance of models with and without SMOTE method. Table A.1 shows the predictive performance of models without SMOTE method. And the results correspond to that in Table 6. The p-values of Wilcoxon test are shown in the parentheses, which indicate that the predictive performance of models with SMOTE is better than that of models without SMOTE.

In addition, we built credit risk models based on extreme values to address the data imbalanced problem. Calabrese et al. (2015) estimated bank default using Generalised Extreme Value (GEV) regression model, which concentrates estimation efforts on the tail of the distribution, adopting a link function that lets the predicted default probability to approach one slower than it approaches zero. Based on the R package BGEVA, we obtained the predictive performance of GEV regression models, shown in Table A.2. Compared with Table 6, the results indicate that the predictive performance of models with SMOTE are better than that of GEV regression models.

*Reference for Appendix A:

Calabrese, R. and Giudici, P. (2015) Estimating bank default with generalised extreme value regression models. Journal of the Operational Research society 66(11), 1783–1792.

Appendix B:. Analysis of news to determine observation period

Fig. B.1 shows the number of websites registered in China in different years and indicates that the number of registered websites increased rapidly after 2006. Fig. B.2 shows how the date when a SME releases its first news on the official website relates to when it is established and when it builds its official website.

We analyze the dates when the first news was released on the official website of each SME. From Fig. B.2, we summarize that before 2014, most SMEs did not release news, although their official websites had been built. To ensure that most SMEs would have news information posted, we chose the observation period from 2014 to 2017 for news release information.

Appendix C:. Results comparison for word embedding algorithms

This paper used a pre-trained BERT model to generate word embeddings. To examine the quality of these word embeddings, we employed other two Word2vec algorithms, Skip-Gram (Mikolov et al., 2013) and CBoW (Bansai et al., 2018), to generate word embeddings. Further, we constructed features based on these word embeddings, following our proposed framework.

For Skip-Gram, the features *SkipIntroB* and *SkipNewsB* refer to the information breadth of the introduction content and news-title content, respectively. The features about the information depth of the introduction content are *SkipIntroD1* to *SkipIntroD4*, and that of the news-title content are *SkipNewsD1* to *SkipNewsD4*. After feature selection (see Table C.1), we selected the effective feature sets "*SkipIntroD2*, *SkipIntroD4*" and "*SkipNewsD1*, *SkipNewsD3*" and added them into models. The predictive performance of models presents in Table C.3.

Similarly, for CBoW, we constructed the features *CBoWIntroB, CBoWNewsB, CBoWIntroD1* to *CBoWIntroD4*, and *CBoWNewsD1* to *CBoWNewsD4*. We selected the effective feature sets "*CBoWIntroD1*, *CBoWIntroD2*" and "*CBoWNewsD3*, *CBoWNewsD1*" (see Table C.2) and added them into models. The predictive performance of models present in Table C.4.

In Table C.3 and Table C.4, the predictive performance of models are lower than that of models with features constructed by BERT (see Table 6). These results indicate that features constructed by BERT are stronger predictors to evaluate the credit risk of SMEs than the features constructed by Skip-Gram and CBoW.

*Reference for A.ppendix C:

Mikolov, T., Sutskever, I., Chen, K., et al. (2013). Distributed representations of words and phrases and their compositionality. Advances in neural information processing systems, 26.

Bansal, B., & Srivastava, S. (2018). Sentiment classification of online consumer reviews using word vector representations. Procedia computer science, 132, 1147–1153.

Appendix D:. Word lists of clusters

Based on the K-means algorithm, we clustered the word embeddings trained by the corpus of website introduction texts and the corpus of posted news titles into four clusters each and presented the top 10 words in each cluster in Table D.1 and Table D.2.

Appendix E:. Results of correlation test

Feature selection results of CFS method for information depth of content-based information are shown in Table E.1. The results indicate that the information depth of cluster 4 of the SMEs' introductions (*IntroD4*) is the best subset of information depth of the introduction content, and indicate that the information depth of cluster 2 of the news-title content (*NewsD2*) is the best feature subset of the information depth of news titles.

Appendix F:. Results of multi-collinearity test

Before modeling, we tested the multi-collinearity among the features, extracted from basic information and content-based information, using

Tolerance and Variance Inflation Factor (VIF). The tolerance is simply the inverse of the VIF. And the higher the VIF, the more likely is the multicollinearity among the features. If the value of VIF is $1 < VIF < 5$, it specifies that the features are moderately correlated to each other (Shrestha et al., 2020). Table F.1 shows the results of Tolerance and VIF, which indicate that our features are moderately correlated to each other.

*Reference for A.ppendix G:

Shrestha, N. (2020). Detecting multicollinearity in regression analysis. American Journal of Applied Mathematics and Statistics, 8(2), 39–42.

Appendix G:. Results comparison for clutersing methods

To compare K-Means++ algorithm, this paper employed other two clustering algorithms, DBSCAN and Mean Shift. For DBSCAN (Hahsler et al., 2019), we used silhouette coefficient to determine parameters, and clustered the word embeddings trained by the introduction corpus and the news titles corpus into four clusters each. Then, we constructed features following the proposed framework. Specifically, the features *DBSCANIntroB* and *DBSCANNewsB* refer to the information breadth of the introduction content and news-title content, respectively. The features about the information depth of the introduction content are *DBSCANIntroD1* to *DBSCANIntroD4*, and that of the news-title content are *DBSCANNewsD1* to *DBSCANNewsD4*.

Similarly, Mean Shift algorithm (Beck et al., 2019) clustered the word embeddings of the introduction corpus and the news title corpus into six and seven clusters, respectively. And we constructed the corresponding features *MeanShiftIntroB*, *MeanShiftNewsB*, *MeanShiftIntroD1* to *MeanShiftIntroD6, and MeanShiftNewsD1* to *MeanShiftNewsD7*.

We used the CFS method to select effective feature subsets for prediction. Table G.1 and Table G.2 present the results of feature selection. The effective feature subsets are "*DBSCANIntroD1*, *DBSCANIntroD3*", "*DBSCANNewsD1*, *DBSCANNewsD3*", "*MeanShiftIntroD1*", and "*MeanShiftNewsD1*". We combined these feature subsets with basic features and added them into models. The predictive performance of models are shown in the Table G.3, and G.4, which are lower than that of models with features constructed by K-Means++ algorithm (see Table 6).

*Reference for A.ppendix G:

Hahsler, M., Piekenbrock, M., & Doran, D. (2019). dbscan: Fast density-based clustering with R. Journal of Statistical Software, 91, 1–30.

Beck, G., Duong, T., Lebbah, M., Azzag, H., & Cérin, C. (2019). A distributed approximate nearest neighbors algorithm for efficient large scale mean shift clustering. Journal of Parallel and Distributed Computing, 134, 128–139.

Appendix H:. The rubust of clustering results evaluation

To validate the robustness of the optimal cluster number, in addition to the silhouette coefficient, we employed two widely-used metrics, Calinski-Harabasz Index (CHI) and Davies-Bouldin Index (DBI). CHI evaluates the quality of clustering based on the average sum of squares of between and within clusters (Caliński & Harabasz, 1974). DBI is based on the average similarity between each cluster and its most similar one (Davies & Bouldin, 1979). Both metrics results in the same optimal cluster number as determined by silhouette coefficient (see Table H.1).

*Reference for A.ppendix H:

Caliński, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, *3*(1), 1–27.

Davies, D. L., & Bouldin, D. W. (1979). A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*, (2), 224–227.

Appendix I:. The results of social network features

We have searched the Weibo and WeChat accounts for the SMEs in our sample, and found that there are 438 (31.32%) and 832 (59.43%) SMEs registered Weibo accounts and WeChat accounts before lending, respectively. We constructed the social network features, *Reg_Weibo* and *Reg_WeChat*, namely whether an enterprise registered a Weibo account and whether an enterprise registered a WeChat account, respectively. The value of feature *Reg_Weibo* (*Reg_WeChat*) is one if a SME has registered a Weibo (WeChat) account; otherwise, it is zero.

We added the two social network features into prediction models (LR, SVM, and XGB) and the results are shown in Table I.1. Compared with the models with *BOW*, the predictive performance of models with *Reg_Weibo* and *Reg_WeChat* are lower.

References

Adjei, M. T., Noble, S. M., & Noble, C. H. (2010). The influence of C2C communications in online brand communities on customer purchase behavior. *Journal of the Academy of Marketing Science, 38*(5), 634–653.

Ahelegbey, D. F., Giudici, P., & Hadji-Misheva, B. (2019). Latent factor models for credit scoring in P2P systems. *Physica A: Statistical Mechanics and its Applications, 522*, 112–121.

Altman, E. I., & Sabato, G. (2007). Modelling credit risk for SMEs: Evidence from the US market. *Abacus, 43*(3), 332–357.

Altman, E. I., Sabato, G., & Wilson, N. (2014). *The value of non-financial information in SME risk management*. SSRN.

Angilella, S., & Mazzù, S. (2015). The financing of innovative SMEs: A multicriteria credit rating model. *European Journal of Operational Research, 244*(2), 540–554.

Angilella, S., et al. (2019). A credit risk model with an automatic override for innovative small and medium-sized enterprises. *Journal of the Operational Research Society, 70* (10), 1784–1800.

Arthur, D., & Vassilvitskii, S. (2006). *k-means++: The advantages of careful seeding.* Stanford.

Baruh, L., et al. (2018). When more is more? The impact of breadth and depth of information disclosure on attributional confidence about and interpersonal attraction to a social network site profile owner. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace, 12*(1).

Bonsall, S. B., IV, Holzman, E. R., & Miller, B. P. (2017). Managerial ability and credit risk assessment. *Management Science, 63*(5), 1425–1449.
Bussmann, N., Giudici, P., Marinelli, D., & Papenbrock, J. (2021). Explainable machine learning in credit risk management. *Computational Economics, 57*, 203–216.
Caliński, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods, 3*(1), 1–27.
Caputo, F., Evangelista, F., Perko, I., & Russo, G. (2017). The role of big data in value co-creation for the knowledge economy. *In 10th Annual Conference of the EuroMed Academy of Business*.
Caputo, F., Magni, D., et al. (2021). Knowledge hiding in socioeconomic settings: Matching organizational and environmental antecedents. *Journal of Business Research, 135*, 19–27.
Cassar, G., Ittner, C. D., et al. (2015). Alternative information sources and information asymmetry reduction: Evidence from small business debt. *Journal of Accounting and Economics, 59*(2–3), 242–263.
Cebi, S. (2013). Determining importance degrees of website design parameters based on interactions and types of websites. *Decision Support Systems, 54*(2), 1030–1043.
Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. *Computers & Electrical Engineering, 40*(1), 16–28.
Chang, X., Chen, Y., Wang, S. Q., et al. (2019). Credit default swaps and corporate innovation. *Journal of Financial Economics, 134*(2), 474–500.
Chatterjee, S., & Kar, A. K. (2020). Why do small and medium enterprises use social media marketing and what is the impact: Empirical insights from India. *International Journal of Information Management, 53*, Article 102103.
Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research, 16*, 321–357.
Chen, M., & Ohta, T. (2010). Using blog content depth and breadth to access and classify blogs. *International Journal of Business and Information, 5*(1), 26.
Chiou, W. C., Lin, C. C., & Perng, C. (2010). A strategic framework for website evaluation based on a review of the literature from 1995–2006. *Information & Management, 47*(5–6), 282–290.
Ciampi, F. (2015). Corporate governance characteristics and default prediction modeling for small enterprises. An empirical analysis of Italian firms. *Journal of Business Research, 68*(5), 1012–1025.
Cui, J., Jo, H., & Na, H. (2018). Does corporate social responsibility affect information asymmetry? *Journal of Business Ethics, 148*(3), 549–572.
Cyr, D., Head, M., Larios, H., & Pan, B. (2009). Exploring human images in website design: A multi-method approach. *MIS Quarterly*, 539–566.
De Weerdt, J., De Backer, M., Vanthienen, J., & Baesens, B. (2011). A robust F-measure for evaluating discovered process models. *In 2011 IEEE Symposium on Computational Intelligence and Data Mining* (CIDM) (pp. 148–155).
Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). *Bert: Pre-training of deep bidirectional transformers for language understanding.* arXiv preprint arXiv:1810.04805.
Dimoka, A., Hong, Y., & Pavlou, P. A. (2012). On product uncertainty in online markets: Theory and evidence. *MIS Quarterly*, 395–426.
Dinh, D. T., Fujinami, T., & Huynh, V. N. (2019). Estimating the optimal number of clusters in categorical data clustering by silhouette coefficient. In *International Symposium on Knowledge and Systems Sciences* (pp. 1–17). Singapore: Springer.
Donovan, J. (2021). Financial reporting and entrepreneurial finance: Evidence from equity crowdfunding. *Management Science, 67*(11), 6629–7289.
Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters, 27*(8), 861–874.
Firth, M., Lin, C., et al. (2019). Hello, is anybody there? Corporate accessibility for outside shareholders as a signal of agency problems. *Review of Accounting Studies, 24*(4), 1317–1358.
García, M. G., Carrillo-Durán, M. V., & Jimenez, J. L. T. (2017). Online corporate communications: Website usability and content. *Journal of Communication Management, 21*(2), 140–154.
Ge, R., Feng, J., et al. (2017). Predicting and deterring default with social media information in peer-to-peer lending. *Journal of Management Information Systems, 34*(2), 401–424.
Gök, A., et al. (2015). Use of web mining in studying innovation. *Scientometrics, 102*(1), 653–671.
Gordini, N. (2014). A genetic algorithm approach for SMEs bankruptcy prediction: Empirical evidence from Italy. *Expert Systems with Applications, 41*(14), 6433–6445.
Hall, M. A. (1999). *Correlation-based feature selection for machine learning.* https://www.cs.waikato.ac.nz/~ml/publications/1999/99MH-Thesis.pdf.
Hand, D. J., & Anagnostopoulos, C. (2014). A better Beta for the H measure of classification performance. *Pattern Recognition Letters, 40*, 41–46.
Hansen, B. (2019). The digital revolution–digital entrepreneurship and transformation in Beijing. *Small Enterprise Research, 26*(1), 36–54.
Hasley, J. P., & Gregg, D. G. (2010). An exploratory study of website information content. *Journal of Theoretical and Applied Electronic Commerce Research, 5*(3), 27–38.
Hung, W. H., Chang, L. M., Lin, C. P., & Hsiao, C. H. (2014). E-readiness of website acceptance and implementation in SMEs. *Computers in Human Behavior, 40*, 44–55.
Jean, R.-J., & Kim, D. (2020). Internet and SMEs' internationalization: The role of platform and website. *Journal of International Management, 26*(1), Article 100690.
Karegowda, A. G., Manjunath, A. S., & Jayaram, M. A. (2010). Comparative study of attribute selection using gain ratio and correlation based feature selection. *International Journal of Information Technology and Knowledge Management, 2*(2), 271–277.
Kim, Y., & An, J. (2021). Voluntary disclosure and rating disagreement among credit rating agencies: Evidence from Korea. *Asia-Pacific Journal of Financial Studies, 50*(3), 288–307.
Lessmann, S., et al. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research, 247*(1), 124–136.
Lopes, L. A., & Melão, N. F. (2016). Website content and design in SME: Insights from Portugal. *International Journal of Electronic Business, 13*(1), 70–97.
Li, H., & Wang, J. (2022). Collaborative annealing power k-means++ clustering. *Knowledge-Based Systems, 255*, Article 109593.
Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In *Advances in neural information processing systems* (p. 30).
Mayr, S., Mitter, C., et al. (2017). Corporate crisis and sustainable reorganization: Evidence from bankrupt Austrian SMEs. *Journal of Small Business Management, 55*(1), 108–127.
Metzger, M. J., & Flanagin, A. J. (2013). Credibility and trust of information in online environments: The use of cognitive heuristics. *Journal of Pragmatics, 59*, 210–220.
Mithas, S., Ramasubbu, N., & Sambamurthy, V. (2011). How information management capability influences firm performance. *MIS Quarterly*, 237–256.
Parker, C. M., et al. (2015). How website design options affect content prominence: A literature-derived framework applied to SME websites. *Journal of Internet Commerce, 14*(2), 139–176.
Parsa, A. B., Movahedi, A., Taghipour, H., Derrible, S., & Mohammadian, A. K. (2020). Toward safer highways, application of XGBoost and SHAP for real-time accident detection and feature analysis. *Accident Analysis & Prevention, 136*, Article 105405.
Pentina, I., & Tarafdar, M. (2014). From "information" to "knowing": Exploring the role of social media in contemporary news consumption. *Computers in Human Behavior, 35*, 211–223.
Qian, Y., Du, Y., Deng, X., et al. (2019). Detecting new Chinese words from massive domain texts with word embedding. *Journal of Information Science, 45*(2), 196–211.
Rahimnia, F., & Hassanzadeh, J. F. (2013). The impact of website content dimension and e-trust on e-marketing effectiveness: The case of Iranian commercial saffron corporations. *Information & Management, 50*(5), 240–247.
Raman, R., Aljafari, R., Venkatesh, V., & Richardson, V. (2022). Mixed-methods research in the age of analytics, an exemplar leveraging sentiments from news articles to predict firm performance. *International Journal of Information Management, 64*, Article 102451.
Ravindran, K., et al. (2015). Social capital and contract duration in buyer-supplier networks for information technology outsourcing. *Information Systems Research, 26*(2), 379–397.
Rekik, R., Kallel, I., Casillas, J., & Alimi, A. M. (2018). Assessing web sites quality: A systematic literature review by text and association rules mining. *International Journal of Information Management, 38*(1), 201–216.
Resch, C., & Kock, A. (2021). The influence of information depth and information breadth on brokers' idea newness in online maker communities. *Research Policy, 50*(8), Article 104142.
Sánchez, C. P., de Llano Monelos, P., et al. (2013). A parsimonious model to forecast financial distress, based on audit evidence. *Contaduría y administración, 58*(4), 151–173.
Salvi, A., Vitolla, F., Rubino, M., et al. (2021). Online information on digitalisation processes and its impact on firm value. *Journal of Business Research, 124*, 437–444.
Stoltz, D. S., & Taylor, M. A. (2019). Concept Mover's Distance: Measuring concept engagement via word embeddings in texts. *Journal of Computational Social Science, 2*(2), 293–313.
Sun, Y., Zhang, Y., et al. (2019). Consumer evaluation of the quality of online health information: Systematic literature review of relevant criteria and indicators. *Journal of Medical Internet Research, 21*(5), e12522.
Sukumar, A., Jafari-Sadeghi, V., & Xu, Z. (2021). The influences of social media on Chinese start-up stage entrepreneurship. World review of entrepreneurship. *Management and Sustainable Development, 17*(5), 559–578.
Talbi, D., & Omri, M. A. (2014). Voluntary disclosure frequency and cost of debt: An analysis in the Tunisian context. *International Journal of Managerial and Financial Accounting, 6*(2), 167–174.
Tobback, E., Bellotti, T., Moeyersoms, J., et al. (2017). Bankruptcy prediction for SMEs using relational data. *Decision Support Systems, 102*, 69–81.
Tsai, M. F., & Wang, C. J. (2017). On the risk prediction and analysis of soft information in finance reports. *European Journal of Operational Research, 257*(1), 243–250.
Vila, N., & Kuster, I. (2011). Consumer feelings and behaviours towards well designed websites. *Information & Management, 48*(4–5), 166–177.
Wang, Z., Jiang, C., Zhao, H., et al. (2020). Mining semantic soft factors for credit risk evaluation in peer-to-peer lending. *Journal of Management Information Systems, 37*(1), 282–308.
Wilcoxon, F. (1992). *Individual comparisons by ranking methods. In Breakthroughs in statistics* (pp. 196–202). New York, NY: Springer.
Yazdanfar, D., & Nilsson, M. (2008). The bankruptcy determinants of Swedish SMEs. *ISBE International Entrepreneurship Conference*.
Yang, C., Chen, M., & Yuan, Q. (2021). The application of XGBoost and SHAP to examining the factors in freight truck-related crashes: An exploratory analysis. *Accident Analysis & Prevention, 158*, Article 106153.
Yin, C., Jiang, C., Jain, H. K., & Wang, Z. (2020). Evaluating the credit risk of SMEs using legal judgments. *Decision Support Systems, 136*, Article 113364.
Zhu, Y., Xie, C., Wang, G. J., et al. (2017). Comparison of individual, ensemble and integrated ensemble machine learning methods to predict China's SME credit risk in supply chain finance. *Neural Computing and Applications, 28*(1), 41–50.
Zu, X., Diao, X., & Meng, Z. (2019). The impact of social media input intensity on firm performance: Evidence from Sina Weibo. *Physica A: Statistical Mechanics and its Applications, 536*, Article 122556.

Cuiqing Jiang is a Professor at School of Management, Hefei University of Technology, P. R. China. He received his PhD degree in 2007 from Hefei University of Technology. His research interests include big data analytics, business intelligence, financial technology (Fintech), and inclusive finance. He has published in such journals as *MIS Quarterly*, *Information Systems Research*, *Journal of the Association for Information Systems*, *Journal of Management Information Systems*, *European Journal of Operational Research*, *Information Sciences*, *Decision Support Systems* and many others.

Chang Yin received her PhD degree in 2022 from Hefei University of Technology. Her research interests include credit risk evaluation of SMEs and data mining. She has published in such journals as *Journal of Product Innovation Management* and *Decision Support Systems*.

Qian Tang is an Assistant Professor of Information Systems, Singapore Management University. She received her PhD degree in Management Information Systems, University of Texas at Austin. Her research interests include business analytics, social media and social networks, online word of mouth, and information security. She has published in such journals as *Production and Operations Management*, *Information Systems Research*, *Journal of Management Information Systems, Journal of the Association for Information Systems*, and many others.

Zhao Wang is an Assistant Professor at the School of Management, Hefei University of Technology. He received his PhD degree in management science and engineering from that university. His research interests include data mining and credit scoring. He has published in such journals as *Information Systems Research, Journal of Management Information Systems*, *European Journal of Operational Research*, *Annals of Operations Research*, *Electronic Commerce Research and Applications*, and many others.