Asia-Pacific Journal of Financial Studies

Asia-Pacific Journal of Financial Studies (2022) **51**, 618–650 doi:10.1111/ajfs.12392

The Role of Social Media in Financial Risk Prediction: Evidence from China*

Qi Wang
School of Economics and Finance, Xi'an Jiaotong University, China

Chenghu Zhang
School of Economics and Finance, Xi'an Jiaotong University, China

Zheng Li**
School of Economics and Finance, Xi'an Jiaotong University, China

Abstract

In this paper, we develop an intelligent approach to detect default risk of FinTech lending platforms. Using China's peer-to-peer (P2P) lending market as an empirical application, we assemble a unique dataset of matched default and non-default platforms. We apply state-of-art techniques to extract sentiment and topic features from several stakeholders' social media data, which are used as supportive soft information. Our approach exhibits better predictive abilities than those with hard information only, where the value of dynamic soft information is demonstrated. Our approach serves as a proof of concept to complement traditional methods of financial risk prediction.

Keywords Default risk detection; P2P lending; Sentiment analysis; Social media; Soft information

JEL Classification: G23, G28, G41

1. Introduction

With the rapid development of disruptive technology and financial innovation, P2P lending is a flexible alternative to traditional banking. P2P lending platforms provide online services to establish debt–credit relations without traditional mediators (Agarwal and Zhang, 2020; Liang and Cai, 2020). *Via* qualified third-party internet platforms (Gao *et al.*, 2018), P2P lending offers several merits, such as easy access

*This work was supported by the Major Project of the National Social Science Foundation of China (No. 17ZDA147), the Key Project of the National Social Science Foundation of China (No. 14AZD033).

**Corresponding author: School of Economics and Finance, Xi'an Jiaotong University, 74 Yantaxi Road, Xi'an 710049, China. Tel: +86 17792359360, email: zheng_li@xjtu.edu.cn.

and low marginal costs of connecting investors and borrowers (Michels, 2012). The associated gains in economic efficiency and reduced information asymmetry are beneficial to borrowers (Feng *et al.*, 2015) and lenders (Bachmann *et al.*, 2011).

Despite its merits, P2P lending has received strong criticism for low entry barriers. Numerous defaults may culminate in substantial financial losses, including suspended operations, lost investment and even business close-down (Fu *et al.*, 2020). Therefore, it is crucial to evaluate the default risk associated with P2P lending platforms in an *ex-ante* manner (Luther, 2020). There are two types of P2P lending risk: the default risk of individual borrowers and the default risk of platforms (Suominen and Toivanen, 2016). The majority of existing studies have been designed to assess individual borrowers' default risk (Kaminski *et al.*, 2004; Weiss *et al.*, 2010; Lin *et al.*, 2013; Emekter *et al.*, 2015; Ge *et al.*, 2017; Chen *et al.*, 2018; Duan, 2019; Kim and Kim, 2020; Liu, 2020; Gao *et al.*, 2021); while other studies focused on identifying the role of detailed information in understanding platform risks or performances (Liu and Sun, 2018; Yang *et al.*, 2018, 2019; Gu *et al.*, 2019; Fu *et al.*, 2020).

The Chinese P2P lending market has some distinctive features and special issues (Xia *et al.*, 2017; Luther, 2020), which, together, may offer practical implications for evaluating financial schemes in other economic settings. It is this that has motivated this research. Social media may signal important information on financial market risks (Moody, 2016; Dong *et al.*, 2018). However, such information has not been well recognized in understanding P2P platform risks. To the best of our knowledge, Fu *et al.* (2020) is the only existing study that investigated this topic with the support of investor reviews. Our study considers broader soft information from several stakeholders, including investor reviews, breaking news and company profiles. The framework is grounded in the theory of systemic functional linguistics (SFL) (Matthiessen, 2014), which argues that language is a system of choices or options that writers use to achieve specific goals. Based on the ideational function of SFL, that is, opinions, emotions, and topics, which is one of the most essential elements of SFL (Abbasi and Chen, 2008), this study automatically extracts and embeds sentiment and topic features from multiple stakeholders' social media data (investors, government regulators, and companies) in the proposed risk detection model, along with other supportive hard information. Among all the soft information, the value of dynamic soft information is demonstrated in comparison with static soft information. This dynamic risk assessment approach could provide government regulators valuable inputs to develop proactive interventions. Moreover, it has the flexibility to be extended to other financial fields such as financial technology (FinTech) loans and corporate fraud detection.

The remainder of this paper is organized as follows. After the literature review section, Section 3 introduces the empirical data. The results of the econometric approach are presented in Section 4. Section 5 outlines our deep learning-based framework proposal. The empirical findings are presented in Section 6. The discussion is presented in Section 7. We close our research in Section 8.

2. Literature Review

The literature review section consists of three subsections. First, we briefly summarize recent studies on borrower default risk. Then, a detailed review of platform default is presented, followed by sentiment analysis, topic modeling techniques, and their applications in the finance field.

2.1. Detection of Borrower Default Risk

The research focus of borrower default risk detection is analyzing borrowers' attributes that may influence their default behaviors. Hard information on applicants' characteristics, such as credit ratings, income level, education level, and marital status, is useful (Emekter *et al.*, 2015; Kim and Kim, 2020), or soft information (Ge *et al.*, 2017; Chen *et al.*, 2018; Dorfleitner *et al.*, 2021) play a role on default behaviors. Machine-learning approaches often outperform traditional statistical evaluation methods (Weiss *et al.*, 2010; Lin *et al.*, 2013; Gao *et al.*, 2021). Recently, deep learning methods are gaining more concerns in this field of research (Kaminski *et al.*, 2004; Byanjankar *et al.*, 2015; Li *et al.*, 2016; Duan, 2019). Table 1 summarizes the key features of related studies.

2.2. Detection of Platform Default Risk

Various methods have been developed to evaluate the default risk of P2P lending platforms. Using logistic regression, Gao *et al.* (2021) identified that ownership, popularity, bond yield and loan maturity play a significant role in China's P2P

Table 1 An overview of default risk detection of P2P borrowers

Literature	Technique	Data	Accuracy
Malekipirbazari and Aksakalli (2015)	Random forests	Hard information only	78.00%
Byanjankar *et al.* (2015)	Neural networks	Hard information only	62.70–74.38%
Ge *et al.* (2017)	Logistic regression	Hard information & social media	65.70–66.40%
Xia *et al.* (2017)	Cost-sensitive XGBoost	Hard information only	70.07–74.85%
Li and Chen (2017)	Back-Propagation Neural Network	Hard information only	68.20–81.40%
Liu *et al.* (2018)	Logistic regression	Hard information only	62.99–64.09%
Kim and Cho (2019)	Transductive Support Vector; Machine learning	Hard information only	86.47%
Duan (2019)	Deep Neural Network (DNN)	Hard information only	93.20%
Zhang *et al.* (2020)	Sentiment analysis; Machine learning	Hard information & loan statement	94.18%

platform's failure. Within a survival analysis, Liu *et al.* (2018) found that the lack of high-quality risk management techniques contributes to P2P lending platform risks. Cheng and Guo (2020) revealed that the risk level of platforms rises with investors' risk aversion, while failed platforms are more likely to declare bankruptcy or runoff after significant political events (Pennington *et al.*, 2014). Related studies are also seen in evidence in other countries, such as the USA (Li and Chen, 2017; Liang and Cai, 2020; Babaei and Bamdad, 2021).

However, previous studies suggest that structured data provide limited capacity to detect financial risks (Li and Chen, 2017; Dong *et al.*, 2018; Jiang *et al.*, 2021). Oliveira *et al.* (2017) and Dong *et al.* (2018), among others, suggest that unstructured social media data may add in important information on financial risks; while however, Fu *et al.* (2020) is the only existing application in detecting P2P platform default risk that we are aware of, which only used investor reviews for sentiment analysis. The research overview of default risk detection in P2P lending platforms is presented in Table 2.

2.3. Approaches to Soft Information

Sentiment analysis and its applications.

Human-generated text information may be necessary to business decision making (Klapper and Love, 2004). In the P2P lending market, previous studies find that the narratives (Herzenstein *et al.*, 2011), concrete descriptions (Kim and Cho, 2019), quantitative words and unverifiable disclosures (Peters *et al.*, 2018), emotional keywords (Dorfleitner *et al.*, 2016; Han *et al.*, 2018), text quality (Mou *et al.*, 2019) shown in borrower's loan description have significant associations with the loan success and borrower's default risk (Yoon *et al.*, 2019). Fu *et al.* (2020) employed a BiLSTM-based model to extract keywords from investor reviews to detect default risks of P2P lending platforms.

Unsupervised representation learning has been highly successful in the domain of NLP (Dai and Le, 2015). Embeddings from Language Model (ELMo) (Maskara *et al.*, 2021), Generative Pre-Training (GPT) (Arner *et al.*, 2019) based on the Transformer model (Suominen and Toivanen, 2016) provide more structured memory for handling long-term dependencies in texts. One appealing example is XLNet (Yang and Luo, 2017). This generalized autoregressive method provides a natural way to factorize the joint probability of the predicted tokens. It relaxes the independence assumption in Bidirectional Encoder Representation from Transformers (BERT) (Devlin *et al.*, 2018).

Topic models and its applications.

Topic models can highlight topical patterns and structures in vast corpora by organizing document collections. Blei *et al.* (2003) proposed the Latent Dirichlet Allocation (LDA) approach, a generative probabilistic model for collections of text corpora. Topic models have been studied in various fields. In the realm of electronic commerce research, topic narratives (Bastani *et al.*, 2019), topic distribution (Cao *et al.*, 2020), topic features (Mou *et al.*, 2019; Zhong and Schweidel, 2020) from customer reviews are verified to have a significant relationship with company performance. In the field

Table 2 An overview of default risk detection of P2P lending platforms

Studies	Technique	Data	Observations	Accuracy
Li *et al.* (2016)	Cox proportional hazard model	Platform hard information	311	Not given
Liu *et al.* (2018)	Short-time multi-source regression	Platform hard information	8629	Not given
Xia *et al.* (2019)	Glowworm swarm optimization; Logistic regression	Platform hard information	99 469	83.95%
Yoon *et al.* (2019)	Logistic regression	Platform hard information	1689	87.80–89.90%
Liu *et al.* (2019)	Cox proportional hazard model	Platform hard information	4503	Not given
Fu *et al.* (2020)	Sentiment analysis; EBCC model (deep learning)	Investor review; Platform hard information	6086	77.40–80.34%
Liang and Cai (2020)	LSTM	Platforms' loan data (USA)	Not given	86.40%
Gong *et al.* (2020)	Generalized linear model	Platform hard information	1649	Not given
Babaei and Bamdad (2021)	Artificial Neural Networks	Platforms' loan data (USA)	34 679	Not given
He and Li (2021)	Competing risk model	Platform hard information	6363	Not given
Gao *et al.* (2021)	Logistic regression	Platform hard information	7047	Not given

of corporate management, topic features are used to investigate stock performance (Liu, 2020), stock market efficiency (Xu *et al.*, 2020) and detect corporate fraud (Dong *et al.*, 2018). However, traditional topic models (Bagheri *et al.*, 2014; Pennington *et al.*, 2014; Zirn and Stuckenschmidt, 2014) cannot capture meaningful semantical regularities between tokens based on distributed word representations. However, Moody (2016) proposed an alternative method, namely lda2vec, which can construct word and document representations simultaneously by mixing skip-gram architecture of word2vec with Dirichlet-optimized sparse topic mixtures.

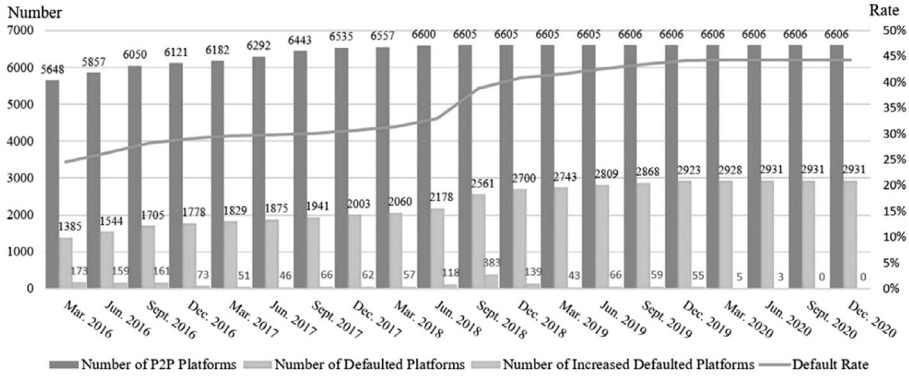
3. Background and Data

3.1. China's P2P Lending Market

China has embraced P2P lending more passionately than other countries over the past decade (He and Li, 2021). By the end of 2020, the number of P2P lending

platforms in China has reached 6607, with total loans exceeding US$1.27 trillion.[1] On December 8, 2017, China's P2P Lending Rectification Office issued the landmark regulation on P2P lending, namely 'Guideline on the Rectification and Acceptance of P2P Lending Risks' (GRAPLR),[2] the P2P lending industry has undergone reshuffling, which has triggered an explosive growth of default platforms since the year of 2018. The quarterly default rate reaches the highest level at the end of 2020, as shown in Figure 1. Moreover, the geographical distribution of default platforms covers more than 90% of China, mainly distributed in the eastern region and expanded to the central region, as shown in Figure 2. In addition, the levels of P2P default risk and 'digital financial inclusion'[3] show a high spatial correlation characteristic. It indicates that potential risks lie in the digital financial business with high

Figure 1 Status of P2P platforms in China

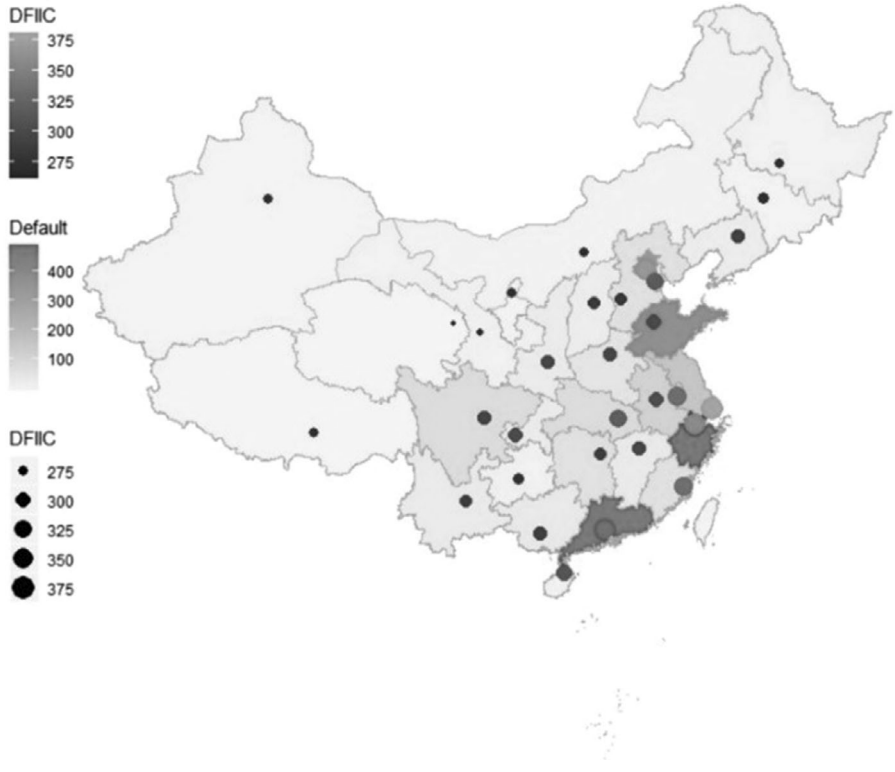


[1]The data comes from the official disclosure of WDZJ. Available from https://www.wdzj.com/dangan/

[2]In December 2017, China's P2P Lending Rectification Office, which is affiliated to the State Council of China and the China Banking and Insurance Regulatory Commission, published its "Guideline on the special rectification and acceptance of P2P lending risks", forcing all regions to complete the archival filing and registration of P2P lending institutions by June 2018. With the introduction of heavy regulatory documents on China's P2P lending in 2017, an industrial regulatory scheme for the depository, filing, and information disclosure has gradually formed. Available from http://www.gzifa.org/gjfg/502

[3]Digital financial inclusion is an emerging financial service that relies on innovative technologies such as big data and artificial intelligence through the continuous improvement of financial infrastructure and further expands the reach of inclusive finance and the depth of services. The Digital Financial Inclusion Index of China (PKU-DFIIC) was compiled by Peking University and covers 31 provinces, 337 cities above the prefecture level, and about 2800 counties in mainland China. PKU-DFIIC generally shows strong regional convergence characteristics and spatial heterogeneity. Available from https://idf.pku.edu.cn/docs/20210421101507614920.pdf

Figure 2 Geographical distribution of P2P lending platform default and DFIIC



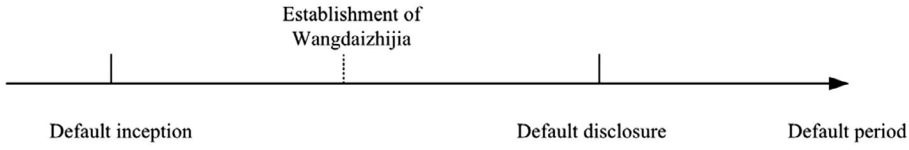
penetration of FinTech, which seriously threatens the sustainability of the emerging financial markets.

3.2. Data and Sample

The empirical data (soft and hard information) come from a few reliable sources. For soft information, WDZJ (https://www.wdzj.com) and WDTY (https://www.p2peye.com) are the best sources for collecting social media data.[4] An improvement over existing research is considering broader social media information: breaking news, investor reviews, and company profiles. *Breaking news* are created by trustworthy editor; *investor reviews* are compiled from all related online discussion forums, including investors' opinions, suggestions, and complaints. C*ompany profiles*

[4]Since its inception in late 2010, WDZJ has become the most widely known third-party P2P lending information platform in China. Like WDZJ, WDTY, founded in early 2012, has become a forum-based comprehensive virtual community and social network platform that provides P2P lending information and platform communication, aiming to supervise the operation of P2P lending platforms in China.

Figure 3 Timeline of default period and establishment of WDZJ



are obtained from the State Administration for Industry and Commerce of the People's Republic of China (SAIC).[5] We use a web crawler to extract all the language-based data in our empirical sample. In addition, from a leading company information search platform QCC (https://www.qcc.com), we extract hard information features of the companies to which the P2P lending platforms belong. Since a P2P lending platform may be disclosed as a default violation of the Ministry of Public Security of China (MPSC) at different times, we consider only the first time and extract social media data prior to that point. As shown in Figure 3, we only consider default platforms disclosed after the establishment of WDZJ.

We use collected information on the Chinese P2P lending market to assess the performance of our baseline approach. All default P2P lending platforms are identified and labeled as follows. We use the Economic Investigation Intervention Announcements (EIIAs) to screen platforms that are disclosed to be the default. Since 2014, under the investigation of the Ministry of Public Security of China (MPSC),[6] the EIIAs provide varying degrees of disclosure of default platforms.

WDZJ has developed a comprehensive database including 397 default platforms following a thorough analysis of EIIAs. Since the P2P Lending Compliance Policy was promulgated by China Banking and Insurance Regulatory Commission (CBIRC)[7] in 2016, we discarded the platforms that have been accused of default violation and involved in the investigation by the MPSC before 2016. We also removed 26 platforms without detailed registration information and 59 without disclosure reports (*e.g.*, the average annualized rate of return and funding guarantee). Finally, we dropped 43 platforms without sufficient social media data (less than 100 pieces of news and reviews) before their first disclosure time. Table 3 presents our sample selection process.

[5]The SAIC is a government agency directly under the State Council of China that conducts market supervision and administrative law enforcement following laws and regulations.

[6]The MPSC is a branch of the State Council of China and the highest leading and commanding organ of the public security system in China.

[7]The CBIRC was established in 2018 and is affiliated with the State Council of China. Its primary responsibilities are to supervise and manage the banking and insurance industries under laws and regulations, guard against and defuse financial risks, protect the legitimate rights and interests of financial consumers, and maintain financial stability.

Table 3 Sample selection process of this study

Distinct platforms	Number
Platforms that have been officially alleged default violation and involved in an investigation by the MPSC	397
Remove those platforms that have been officially alleged default violation and involved in the investigation by the MPSC before 2016	20
Default platforms under the 2016 P2P Online Lending Compliance Policy: 397–20=	377
Remove those platforms with non-detailed registration information	26
Remove those platforms with undisclosed operating reports	59
Remove those platforms that do not have enough social media data in WDZJ	43
Empirical sample size (default platforms): 377–26–59-53=	249

Given the profile of the P2P lending market in China (see Figure 1 in Section 3), random sampling of unbalanced data set in this case would result in an extremely high percentage of non-default platforms in the sample, thus making attempts to investigate significant features for predicting platform default not meaningful (Summers and Sweeney, 1998). To mitigate this concern, we conduct propensity score matching (PSM) (Caliendo and Kopeinig, 2008) to match each default platform with a control platform (non-default) for classification purposes, which is a sampling strategy for handling rare events (Abrahams *et al.*, 2015). Control platforms are matched with the same background (state-owned or non-state-owned), further filtered by PSM on registration capital and firm type. Following existing studies (Dong *et al.*, 2014, 2018; Abrahams *et al.*, 2015), we build a balanced data set with a 1:1 ratio for default and non-default platforms. In summary, our data set includes 249 default platforms and 249 matched non-default platforms, including social media data[8] and platform hard information features. Table 4 describes the summary statistics of our data set. The numbers in parentheses are the average value.

4. Econometric Model with Structured Variables

First, we conduct a Probit regression, extending the dimensions of hard platform information from the level of the company, platform, and CEO.

4.1. Dependent Variable

The platform default risk (*PDR*) in this study is defined as the risk that a platform may occur through significant events, including bankruptcy, business termination,

[8]Social media data for non-default platforms are collected until December 28, 2020, when this study started.

Table 4 Overview of the data set

Data set (498 platforms)	Content	No. of pieces	No. of sentences	No. of words	No. of structured financial variables
Unstructured data	Breaking news	33 313 (66.89)	74 049 (148.699)	6 072 089 (12192.94)	–
	Investor reviews	619 829 (1244.63)	4 175 092 (8383.71)	221 279 910 (444337.16)	–
	Company profiles	498 (1)	4215 (8.46)	206 535 (414.72)	–
Structured financial variables	–	–	–	–	28

etc. The disclosure of EIIAs measures *PDR* as follows: the value '1' for default platforms and '0' otherwise.

4.2. Hard Information Features

We consider platform-, company-, and CEO-levels of hard information features that may affect *PDR*. At the platform level, previous studies find that *registered capital (RC)*, *automatic bid (AB)*, *annualized return (AAR)*, the *average term of the loan (ATL)*, *bank depository (BD)*, *platform background (PB)* affect the *PDR* (Yang and Luo, 2017; Liu *et al*., 2019; Yoon *et al*., 2019; Gao *et al*., 2021; He and Li, 2021). Moreover, we examine the effect of *registered address (RA)*, *type of firm (TF)*, *area (AR)* on *PDR*. Detailed variable definitions are provided in Table 5.

At the company level, we consider *abnormal operation (AO)* events to measure the number of abnormal operation events since the company was established. *Tax credit level (TAX)* indicates the number of times the company has been rated as A-level in taxation ratings since its establishment. We compute *information disclosure (IND)* by measuring the number of company annual reports disclosed through official websites. *Social capital (SC)* is created by measuring the number of official accounts operated on two leading social-network platforms in China: Weibo and WeChat. Besides, *ICP registered situation (IR)*, a dummy variable, is used to investigate the compliance of the P2P platform website.

As Chen *et al*. (2006) propose, there may be variations of CEO-level characteristics in corporate fraud. The chairman and CEO are the same people, shares owned by legal entities, *etc*. This is further developed by He and Li (2021) in P2P lending markets. Hence, we consider three variables related to CEOs' behaviors: the *executive (EX)*, *actual controller (AC)* and *legal representative (LR)*. Tables 6 and 7 provide descriptive statistics for the selected qualitative and quantitative variables, respectively.

4.3. Model A: Probit Regression with Heckman Two-Stage Selection Model

As a baseline approach, our empirical analysis starts with the Probit model (see equation 1) and Heckman two-stage selection model (Heckman, 1976, 1979) to assess *PDR*. A variant inflation factors (VIF) test is used to address potential multicollinearity concerns among our key variables, and three variables with VIF > 10 are removed: AAR, IR, and AR.

$$\begin{aligned} Prob(PDR_i = 1|x) = {} & \alpha_0 + \alpha_1 * RC_i + \alpha_2 * RA_i + \alpha_3 * AO_i + \alpha_4 * TAX_i + \alpha_5 \\ & * SC_i + \alpha_6 * IND_i + \alpha_7 * TF_i + \alpha_8 * ASE_i + \alpha_9 * EX_i \\ & + \alpha_{10} * AC_i + \alpha_{11} * LR_i + \alpha_{12} * PB_i + \alpha_{13} * ATL_i \\ & + \alpha_{14} * BD_i + \alpha_{15} * AB_i + \delta_i. \end{aligned} \tag{1}$$

Following Heckman (1979) and Zadrozny* (2004), we construct a Heckman two-stage selection model to address potential endogenous problems in the

Table 5 Variable definitions

Category	Variable	Definition
Dependent variable	*Platform default risk (PDR)*	Dummy variable equals '1' if the platform has been officially alleged default violation by MPSC and '0' otherwise
Company-level variables	*Registered capital (RC)*	The amount of the platform's paid-in capital
	Abnormal operation (AO)	The number of abnormal operations during the company's existence, such as judicial cases, bankruptcy and reorganization, *etc*
	Tax rating (TAX)	The number of times the company has been rated as A-level in taxation ratings since its establishment
	Social capital (SC)	The number of official accounts operated on two leading social platforms: Weibo and WeChat.
	Information disclosure (IND)	The number of company annual reports disclosed through official websites
	DFIIC	The DFIIC of the province where the company is located
CEO-level variables	*Alteration of stock equity (ASE)*	The number of changes in the company's stock equity information made by the CEOs
	Executive (EX)	Dummy variable equals '1' if the company has the executive legally enforced for breach of trust and '0' otherwise
	Actual controller (AC)	Dummy variable equals '1' if the actual controller is the ultimate beneficiary and '0' otherwise
	Legal representative (LR)	Dummy variable equals '1' if the legal representative is the ultimate beneficiary and '0' otherwise
Platform-level variables	*Platform background (PB)*	Dummy variable equals '1' if the controlling shareholder of the platform is the government or a government-related agency and '0' otherwise
	Automatic bid (AB)	Dummy variable equals '1' if the platform has the function of automatically allocating loan targets for investors and '0' otherwise
	ICP registered (IR)	Dummy variable equals '1' if the platform has been registered with Internet Content Provider (ICP) business license and '0' otherwise
	Average annualized return (AAR)	The average annualized return on the platform during the operation period.
	Average term of the loan (ATL)	The average term of the loan (months) of the platform
	Bank depository (BD)	Dummy variable equals '1' if the platform has a custodian bank and '0' otherwise

Table 5 (*Continued*)

Category	Variable	Definition
	Area (AR)	Dummy variable equals '0' if the company is located in the western region of China (AR0), '1' central region (AR1), and '2' eastern region (AR2)
	Registered address (RA)	Dummy variable equals '1' if the company's registered address is specific and '0' otherwise
	Type of firm (TF)	Dummy variable equals '1' if the firm is a limited liability company and '0' joint-stock limited liability company
	Establishment pre-2016 regulations (EP2016)	Dummy variable equals '1' if the platform is founded prior to the 2016 standardized P2P lending industry access regulations and '0' otherwise

Table 6 Descriptive statistics for qualitative variables

Variable	Percentage of '=1'	Variable	Percentage of '=1'
Platform default risk (PDR)	50.00	*ICP registered (IR)*	99.19
Platform background (PB)	19.35	*Bank depository (BD)*	69.35
Registered address (RA)	11.49	*Area (AR0)*	5.44
Executive (EX)	12.90	*Area (AR1)*	8.47
Actual controller (AC)	13.10	*Area (AR2)*	86.09
Legal representative (LR)	64.11	*Type of firm (TF)*	65.32
Automatic bid (AB)	43.75	*Establishment pre-2016 regulations (EP2016)*	92.33

Table 7 Descriptive statistics for quantitative variables

Variable	Mean	Standard Deviation	Min.	Max.
Registered capital (RC)	10 500.33	23 733.71	100.00	300 000.00
Abnormal operation (AO)	0.91	1.02	0	5
Tax rating (TAX)	0.31	0.65	0	4
Social capital (SC)	1.45	1.43	0	12
Information disclosure (IND)	4.64	7.01	0	64
Alteration of stock equity (ASE)	16.39	10.56	0	77
Average annualized return (AAR)	0.11	0.03	0.02	0.48
Average term of loan (ATL)	4.32	5.56	0	38.32
Area's DFIIC (AD)	392.13	32.31	305.50	431.93

disclosure of *PDR* that might be resulted from sample selection bias. In the Heckman first-stage (see equation 2), the regression model is considered to contain an endogenous variable D, where samples $D = 0$ not considered in the model will result in estimation bias.

$$Y = \beta' X + \theta D + \mu. \tag{2}$$

Suppose there is an unobservable interference term, namely the inverse mills ratio (*IMR*), which affects the generation of D. The vector of exogenous variables Z (*i.e.*, instrumental variables related to D) is used to fit the change of D. Then the fitted data generation D^* is obtained (see equation 3).

$$D^* = \alpha_0' Z + \alpha_1' X + \upsilon. \tag{3}$$

If $D^* \geq 0$, then $D = 1$, otherwise $D = 0$. If the random error μ is correlated, then $E(u|D) \neq 0$, resulting in the biased estimator θ, where the deviation can be mitigated by constructing the *IMR* (see equation 4).

$$IMR = E\mu \mid D = \begin{cases} \varphi\left(\hat{\alpha}_0' Z + \hat{\alpha}_1' X\right) / \phi\left(\hat{\alpha}_0' Z + \hat{\alpha}_1' X\right) & if\ D = 1 \\ -\varphi\left(\hat{\alpha}_0' Z + \hat{\alpha}_1' X\right) / \left(1 - \phi\left(\hat{\alpha}_0' Z + \hat{\alpha}_1' X\right)\right) & if\ D = 0 \end{cases} \tag{4}$$

Here $\varphi(\cdot)$ and $\phi(\cdot)$ are the density function and cumulative distribution function of the standard normal distribution, respectively. Then the *IMR* is added to equation (2) for regression (see equation 5).

$$Y = \beta' X + \theta D + \rho\sigma IMR + \varepsilon. \tag{5}$$

Here θ is the unbiased estimator, and the statistical significance of *IMR* can be used to determine whether the sample selection bias exists.

Employing the Heckman model allows us to examine whether there are differences between platforms investigated by MPSC and those overlooked by MPSC with no disclosure of default issues. The exogenous variable must meet the exclusion restriction criteria to ensure non-collinearity between the two-stage models (Sartori, 2003; Bushway *et al.*, 2007; Bärnighausen *et al.*, 2011). Identifying a reasonable exclusion restriction requires significantly associated with whether platforms are involved in MPSC's investigation but does not affect *PDR* in the second-stage model. As a result, the newly created dummy variable under MPSC's investigation (UIM) is used as a proxy for platforms under MPSC's rigorous investigation, with default or non-default outcome disclosed by MPSC when taking the value '1', and '0' otherwise, and becomes the dependent variable in the first-stage model. Since 2016 the State Council of China issued the Notice on the

Special Rectification of Internet Finance Risks (SRIFR).[9] The P2P industry has been subject to comprehensive and rigorous regulation regarding industry access, capital supervision, penalties, *etc.* Given this, P2P platforms *established before the 2016 SRIFR policy* (*EB2016*) may not be compliant enough regarding industry access, operating mode, and capital management, which are more likely to be targeted and investigated by MPSC. Furthermore, given China's digital financial inclusion industry's spatial aggregation characteristic, P2P platforms may be more regulated in areas with a higher DFIIC, where financial risk regulation and industry self-regulation are more prevalent. Conversely, P2P platforms in areas with lower *DFIIC* may be more likely to be investigated by the MPSC because of the lagging regional regulation. Following the criteria on exclusion restriction (Bushway *et al.*, 2007), *EB2016* and *DFIIC* are chosen as the exclusion criteria for the selection equation. We believe that *EB2016* and *DFIIC* will be significantly associated with *UIM*, unrelated to *PDR*.

The Heckman two-stage model in this paper is developed (see equation 6), where i represents an individual platform and σ_i, κ_i are the random errors.

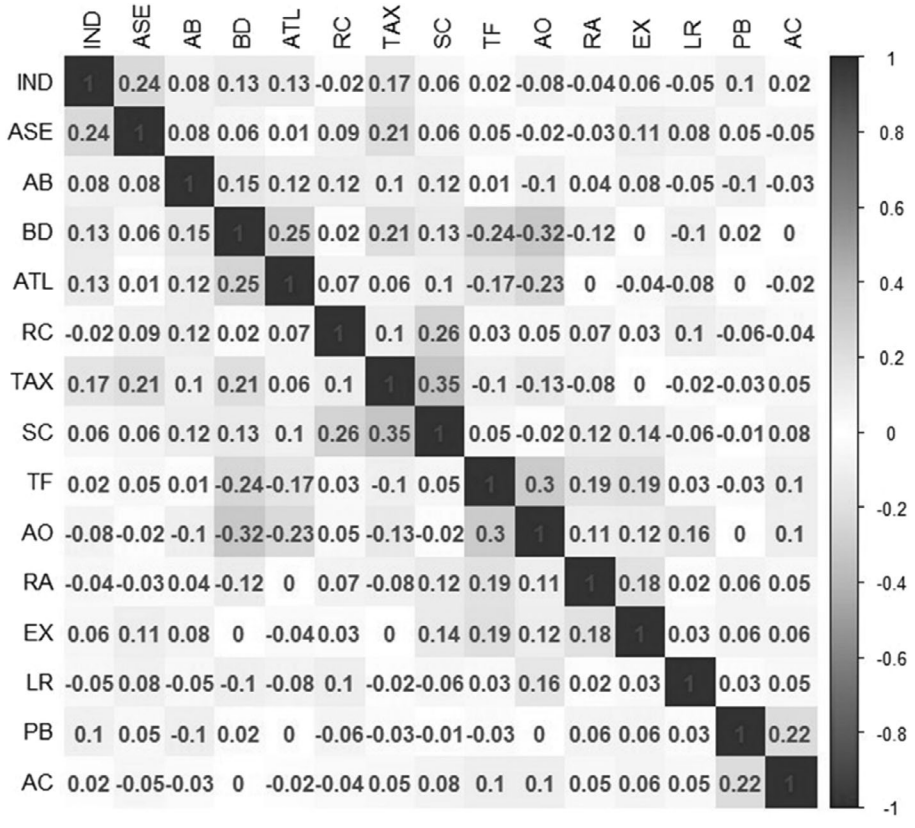
$$\begin{cases} Prob(UIM_i = 1 \mid \omega) = \Phi(\omega_i'\gamma) = \beta_0 + \beta_1 * EB2016_i + \beta_2 * DFIIC_i + \beta_3 * RC_i + \beta_4 * PB_i + \sigma_i \\ Prob(PDR_i = 1 \mid x) = \mu_0 + \rho\sigma\hat{\lambda}(\omega_i'\gamma) + \mu_1 * RC_i + \mu_2 * RA_i + \mu_3 * AO_i + \mu_4 * TAX_i + \mu_5 * SC_i \\ \qquad + \mu_6 * IND_i + \mu_7 * TF_i + \mu_8 * ASE_i + \mu_9 * EX_i + \mu_{10} * AC_i + \mu_{11} * LR_i \\ \qquad + \mu_{12} * PB_i + \mu_{13} * ATL_i + \mu_{14} * BD_i + \mu_{15} * AB_i + \kappa_i. \end{cases} \quad (6)$$

For detailed information on the traditional detection method used in this current study as a baseline to demonstrate the effectiveness of our proposed method, please refer to Yang and Luo (2017) Xia *et al.* (2019) and Yoon *et al.* (2019).

Correlation analysis of critical variables is reported in Figure 4. Table 8 presents the results (Model A), where column (1) presents the Probit model results. The result of the first-stage model is reported in column (2), where the instrumental variables (IV) *EB2016* and *DFIIC* used to identify a valid exclusion restriction are significantly associated with *UIM*. We, therefore, conclude that our IVs are valid and reliable to run the Heckman two-stage model and mitigate the endogeneity problem. The result of the second-stage model is presented in column (3), which reveals that the estimation of *IMR* is significant at the 1% level. Thus, our results

[9]In October 2016, the Special Rectification of Internet Finance Risks (SRIFR) [9] was issued by the State Council of China, requiring the People's Bank of China, the China Banking and Insurance Regulatory Commission to strengthen the rectification of P2P lending, crowdfunding, and third-party payment industries in terms of industry access, capital flow monitoring, penalties, *etc.* This policy is intended to regulate various industries of Internet finance and improve the inclusiveness of financial services in China. Available from http://www.gov.cn/zhengce/content/2016-10/13/content_5118471.htm

Figure 4 Pearson correlations among all variables



suggest that the sample selection bias is identified and does not plague our model (Irfan, 2011; Katmon and Farooque, 2017). Most of the *p*-values in column (3) are statistically significant. The Pseudo R-squared of 0.48 indicates that the model is well-fitting. The identified roles of *RC*, *ATL* and *BD* are consistent with existing studies (Yang and Luo, 2017; Liu *et al.*, 2019; Yoon *et al.*, 2019; Gao *et al.*, 2021; He and Li, 2021). We also find some interesting findings on CEO-level characteristics. For example, there would be a higher risk with the changes in the company's stock equity information made by the CEOs (*ASE*); or if the executive is legally enforced for breach of trust (*EX*); or if the actual controller (*AC*) or legal representative (*LR*) is not the ultimate beneficiary of the company.

5. A Deep Learning-Based Approach with Soft Information

Previous empirical studies concentrate solely on the role of platform hard information in understanding P2P lending risk (see reviewed evidence in Section 2). Grounded in SFL theory, this study extends the strand of default detection research

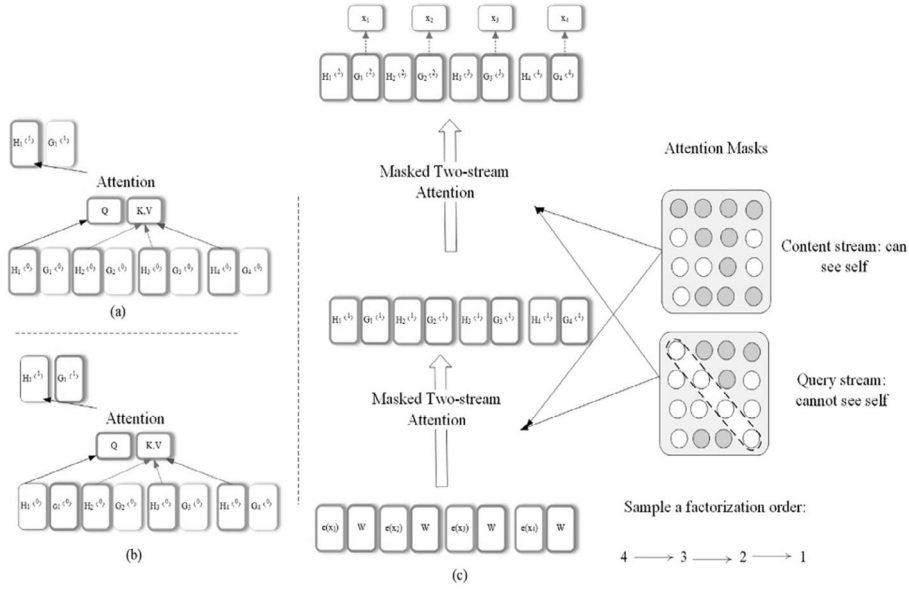
Table 8 Parameter estimation of statistical approach: Probit model and Heckman two-stage model

***, **, * and. Indicate that the variable is significant at the 0.001, 0.01, 0.05, 0.10 levels, respectively. The standard errors are reported in parentheses.

Variables	(1) Probit model DV=*PDR*	(2) Heckman first-stage model DV=*ID*	(3) Heckman second-stage model DV=*PDR*
Instrumental variables			
EB2016		1.378 (0.148)***	
DFIIC		−0.105 (0.064)	
Company-level variables			
RC	−0.255 (0.096)**	0.306 (0.168)	−0.191 (0.094)*
RA	0.953 (0.285)***		0.990 (0.285)***
AO	0.710 (0.086)***		0.685 (0.087)***
TAX	−0.463 (0.154)**		−0.483 (0.151)**
SC	−0.019 (0.068)		−0.012 (0.068)
IND	−0.010 (0.013)		−0.007 (0.012)
TF	0.536 (0.297)		0.594 (0.296)*
CEO-level variables			
ASE	0.019 (0.007)*		0.020 (0.007)**
EX	0.297 (0.098)**		0.300 (0.099)**
AC	1.028 (0.241)***		1.052 (0.241)***
LR	0.330 (0.167)*		0.327. (0.169)
Platform-level variables			
PB	−0.382.(0.205)	0.996 (0.225)***	−0.069 (0.239)
ATL	−0.048 (0.018)**		−0.051 (0.018)**
BD	−0.809 (0.175)***		−0.695 (0.181)***
AB	−0.157 (0.154)		−0.109 (0.156)
IMR			1.348 (0.521)**
Constant	0.329 (0.385)	−0.430 (0.138)**	−0.832 (0.313)**
Wald chi^2	319.65	133.66	327.52
$p < chi^2$	0.000	0.000	0.000
Pseudo R-squared	0.4649	0.1909	0.4763
N	496	646	496

(Cecchini *et al.*, 2010; Abbasi *et al.*, 2012; Dong *et al.*, 2018) by proposing an analytic framework for default detection of P2P lending platforms with the support of social media data. An intelligent approach based on machine learning is established. Language-based features of soft information, including breaking news, investor reviews, and company profiles, are extracted automatically and fed into machine learning classifiers for default risk detection. We compare the predictive performance of models with *vs* without soft information to demonstrate the hard evidence on soft information.

Figure 5 Architecture of XLNet



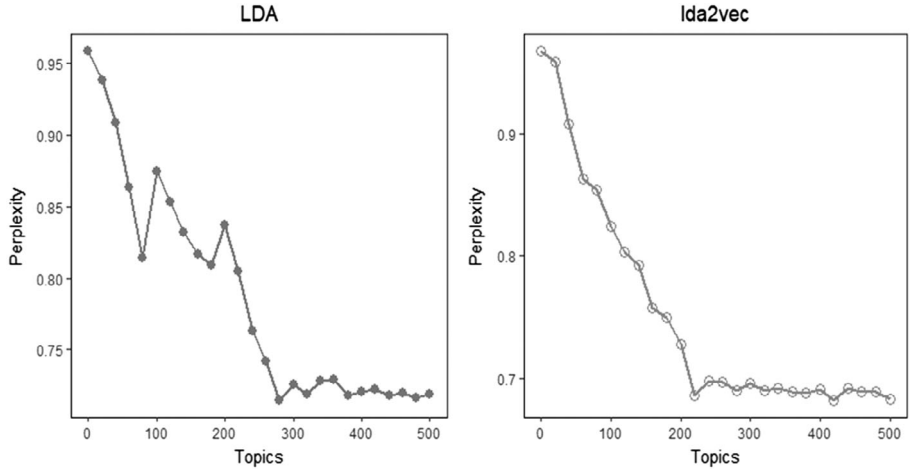
5.1. Sentiment Feature Extraction Using XLNet

Given that sentiment features are implicit in the original forms, such as breaking news and investor reviews, we use XLNet and Word2Vec to extract sentiment features. As a generalized autoregressive pre-training method, XLNet enables learning bidirectional contexts and avoids the shortcomings of the BERT model in training-tuning differences. First, XLNet arranges the tokens of sentences randomly based on the main idea of the Permutation Language Model (PLM) and then uses the autoregressive model to predict the words at the end of the sentence without modifying the order of the sentence tokens so that XLNet have both autoencoding and autoregressive language modeling capabilities. XLNet also adopts Two-Stream Self-Attention to realize that the location information of the token can be seen. In contrast, the content information cannot be seen. Moreover, a component of XLNet, Transformer-XL, is used to transform sentiment features embedded in their original form into measurable factors. The entire framework is shown in Figure 5. We use XLNet pre-training model based on the Chinese corpus, while other languages are also available.

5.2. Topic Feature Extraction Using lda2vec

Thematic topics provide helpful information for predicting intentional financial statement misreporting (Brown *et al.*, 2020). The topics presented in the company profile of a default platform may include some unique features compared to a non-default platform. We employ lda2vec to extract topic features from the company profile, where LDA is used as a baseline method. Following Blei *et al.* (2003), we

Figure 6 Perplexity score of LDA and lda2vec



calculate the perplexity scores of lda2vec and LDA for a different number of topics ranging from 20 to 500, as shown in Figure 6. The model with 280 topic features (a minimum perplexity score of 0.72) and a model with 220 topics (a minimum perplexity score of 0.69) are selected. It has been noted that lda2vec is more efficient relative to LDA.

5.3. Full Framework

The entire framework based on deep learning is shown in Figure 7. The XLNet and Word2Vec method extract sentiment feature from breaking news and investor reviews. In contrast, lda2vec and LDA extract topic features from company profiles. Four machine learning techniques are used and compared: Long Short-Term Memory (LSTM), XGBoost, Random Forest, and Support Vector Machine (SVM). We use accuracy, recall, F1 score and Area Under Curve (AUC) to evaluate the prediction performance of trained classifiers. A tenfold cross-validation technique is employed to assess how the model results generalize to independent test data, which divides the sample into 70% training set and 30% testing set.

6. Results of the Proposed Approach

We conduct a comprehensive analysis to evaluate the performance of our proposed approach systematically. Four classifiers (LSTM, XGBoost, Random Forest, SVM) are used to predict platform default risk. We iteratively include each of the four sets of input features (platform hard information and three sources of soft information: investor reviews, breaking news, company profile) to evaluate their effects on platform default detection.

Figure 7 Proposed framework for P2P platform default detection

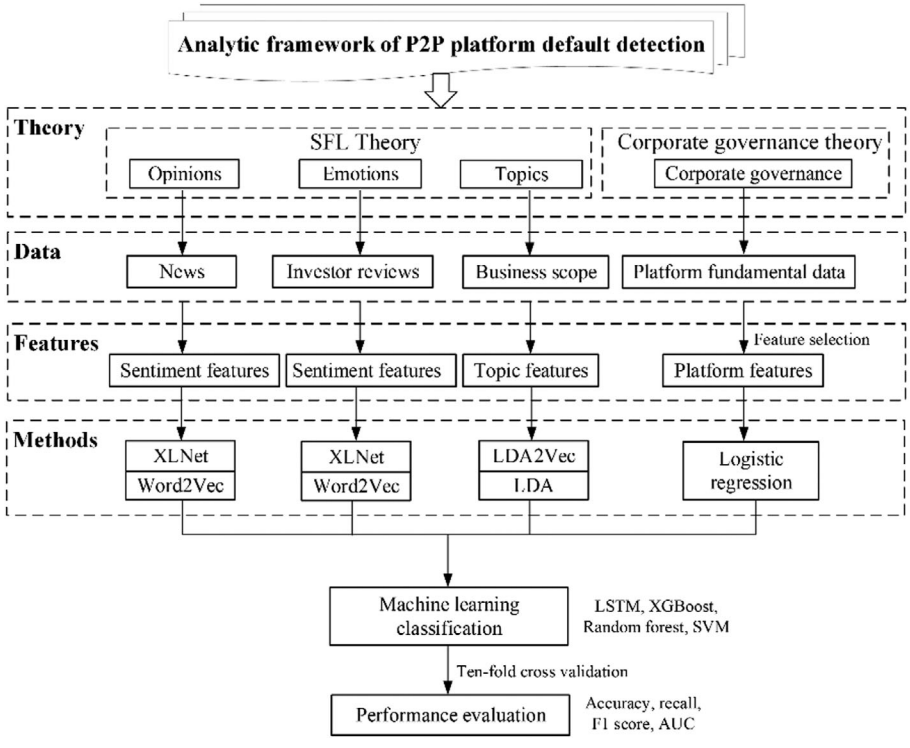


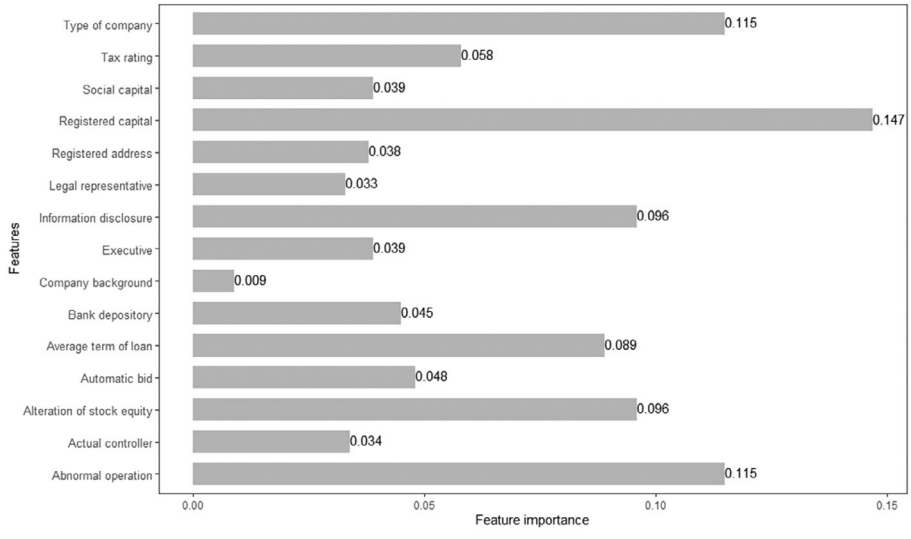
Table 9 Machine learning performance with hard information only

Model B		Average accuracy (%)	Average recall (%)	Average F1 score (%)	Average AUC (%)
LSTM	Training	97.71	98.80	96.71	97.71
	Testing	83.92	85.13	84.84	84.97
XGBoost	Training	96.25	95.93	96.21	96.25
	Testing	82.60	84.75	83.03	83.75
Random Forest	Training	99.71	99.40	99.71	99.72
	Testing	75.84	74.67	75.81	75.82
SVM	Training	99.71	99.92	99.71	99.71
	Testing	52.35	79.73	62.43	52.53

6.1. Model B: Using Hard Information Only

Table 9 documents the average performance of the four classifiers only using platform hard information (Model B). The LSTM model produces the best performance with an average accuracy of 97.71% and 83.92% in training and testing,

Figure 8 Scores of platform hard information features



respectively. Furthermore, we use the XGBoost model to obtain the feature-importance score for each structured variable. The *RC*, *TF*, *AO* are the most important features among all types of selected hard information, as shown in Figure 8.

6.2. Models C1–C3: Using Soft Information Only

In this subsection, we present the predictive performance of three models. Each of them includes one source of soft information: investor review (Model C1), breaking news (Model C2) or company profiles (Model C3). For Models C1 and C2, detailed results are provided in Tables 10 and 11, respectively. The LSTM technique results in the best overall performance among four classifiers. Moreover, XLNeT outperforms word2vec for sentiment features extracted from both investor reviews and breaking news. For Model C3, detailed results are provided in Table 12, and the combination of LSTM+lda2vec delivers the best prediction.

6.3. Comparing the Role of Each Source of Soft Information and Hard Information

Based on the LSTM technique, which delivers the best performance under each source of information, we compare the role of breaking news, investor review, company profile, and hard information separately (see Figure 9). It has been noted that sentiment features extracted from breaking news contribute the most to prediction accuracy, followed by investor review. This evidence demonstrates the significance of soft information (particularly breaking news and investor review revealed in this current study) in predicting default risks. At the same time, existing research

Table 10 Performance with investor reviews only

Model C1			Average accuracy (%)	Average recall (%)	Average F1 score (%)	Average AUC (%)
XLNeT	LSTM	Training	98.72	97.90	97.64	99.71
		Testing	88.94	87.74	88.16	88.20
	XGBoost	Training	98.77	99.67	98.64	98.64
		Testing	85.17	86.17	85.22	85.14
	Random Forrest	Training	96.83	99.44	96.99	96.76
		Testing	84.42	85.61	84.77	85.63
	SVM	Training	98.85	99.44	98.88	98.83
		Testing	84.96	83.27	83.68	84.57
Word2vec	LSTM	Training	97.71	98.24	97.39	97.65
		Testing	85.18	83.37	84.44	85.76
	XGBoost	Training	96.83	94.77	96.74	96.81
		Testing	79.87	81.58	80.52	79.83
	Random Forrest	Training	97.41	95.38	97.35	97.40
		Testing	81.21	77.22	80.33	81.46
	SVM	Training	88.76	82.08	87.93	88.74
		Testing	83.22	77.33	82.27	83.26

Table 11 Performance with breaking news only

Model C2			Average accuracy (%)	Average recall (%)	Average F1 score (%)	Average AUC (%)
XLNeT	LSTM	Training	99.71	97.86	98.75	99.87
		Testing	89.26	86.11	88.57	89.16
	XGBoost	Training	84.73	84.10	84.81	84.73
		Testing	85.14	87.53	85.50	85.36
	Random Forrest	Training	98.56	98.30	98.58	98.56
		Testing	85.23	80.56	84.06	85.09
	SVM	Training	99.71	99.43	99.72	99.71
		Testing	84.26	86.11	85.57	84.15
Word2vec	LSTM	Training	98.65	97.38	97.60	98.75
		Testing	86.69	85.38	86.40	86.22
	XGBoost	Training	96.83	94.77	96.74	96.81
		Testing	79.87	81.58	80.52	79.83
	Random Forrest	Training	97.12	97.04	97.04	97.12
		Testing	81.33	77.22	81.21	81.46
	SVM	Training	87.03	79.43	86.07	87.10
		Testing	84.56	75.34	82.71	84.38

Table 12 Performance with company profiles only

Model C3			Average accuracy (%)	Average recall (%)	Average F1 score (%)	Average AUC (%)
lda2vec	LSTM	Training	97.39	96.10	97.26	98.25
		Testing	73.24	74.64	74.22	73.15
	XGBoost	Training	99.14	99.98	98.14	99.14
		Testing	72.15	70.68	72.68	73.14
	Random Forrest	Training	99.71	100.00	99.71	99.71
		Testing	65.10	69.74	67.09	65.01
	SVM	Training	95.68	95.35	95.63	95.67
		Testing	57.72	43.42	51.16	58.01
LDA	LSTM	Training	98.48	97.47	98.15	98.15
		Testing	70.63	69.90	70.15	69.74
	XGBoost	Training	98.75	97.69	97.66	97.11
		Testing	68.38	69.80	68.70	69.15
	Random Forrest	Training	98.79	99.65	99.77	98.26
		Testing	65.49	66.16	66.15	65.18
	SVM	Training	95.45	96.11	94.86	95.66
		Testing	52.79	51.14	51.57	49.68

concentrates solely on hard information, except Fu *et al.* (2020), which only used investor review as an additional source of information.

6.4. Model D: Incorporation of Hard Information and Soft Information

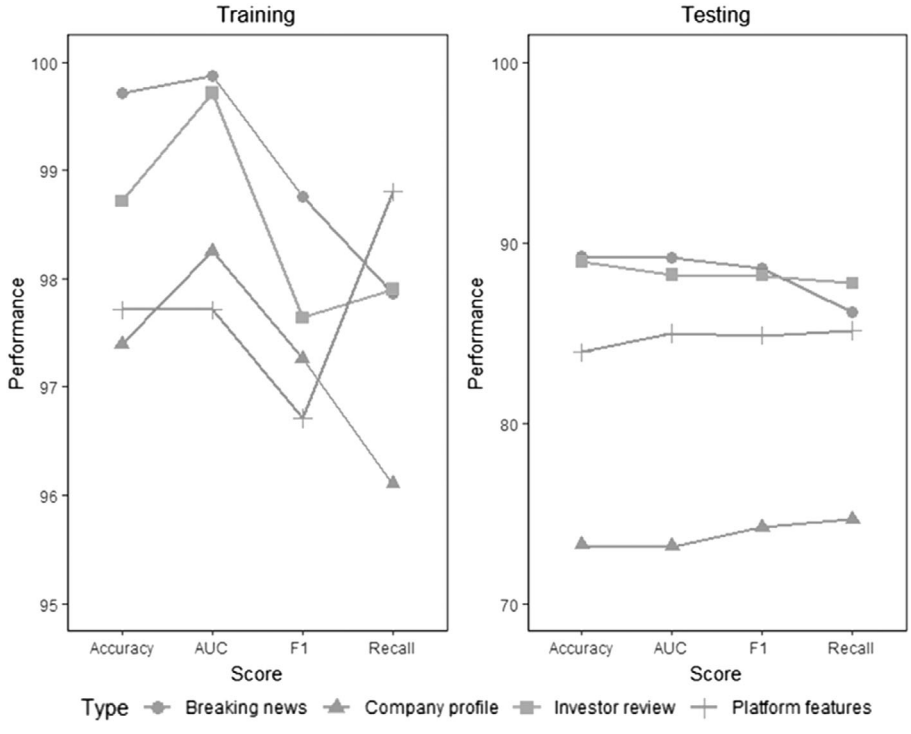
Model D with the full features (hard+soft) is shown in Table 13. Figure 10, which produces the best performance, is the ultimate model for our empirical analysis. Using the LSTM technique, the model delivers the best out-of-sample (testing) performance. The average accuracy, recall, F1 measure, and AUC on the testing data set are 94.63%, 92.11%, 94.59% and 93.68%, respectively. Overall, our findings suggest incremental value with the support of soft information (sentiment features and topic features) and hard information for risk prediction. For this empirical data set, the combination of XLNet+lda2vec outperforms Word2Vec + lda2vec.

6.5. Robustness Check

We use different proportions of unbalanced data sets to test the robustness of our proposed framework. The LSTM models with full features (hard and soft information), trained by three data sets (Default: Non-default = 1:2, 1:5 & 1:10) and the empirical data set (1:1) deliver the average accuracy rate being 93.75%, 88.89%, 89.46% and 94.63% respectively. The results show that our proposed risk-detection approach is robust to alternative data sets.

To verify the validity, we assess the model performance before and after the inflection point of a significant Chinese internet finance regulation-GRAPLR, fully

Figure 9 The individual role of three sources of soft information and hard information



effective in June 2018. Following Farooq and Qamar (2019) and Fernández *et al.* (2018), an over-sampling strategy is used to obtain the balanced data set before and after June 2018, as shown in Table 14 Figure 11. Notably, its out-of-sample performance after the inflection point triggered by 'GRAPLR' is even better than its performance before the inflection point. This risk-detection approach delivers better forecasts after the Chinese P2P lending industry was subjected to stringent regulation under 'GRAPLR'. The results, to some extent, suggest that it can anticipate the influence of major events and radical changes in the market, which, in turn, highlight the important role of soft information in risk prediction.

7. Discussion

Table 15 summarizes a between-study and within-study comparison to demonstrate the benefits of using soft information and the contributions of this study. Among our reviewed studies listed in Section 2, Xia *et al.* (2019), Yoon *et al.* (2019), and Fu *et al.* (2020) report their forecasts for Chinese P2P lending platforms. From an out-of-sample prediction aspect, our proposed approach with a systematic treatment of soft and hard information outperforms Xia *et al.* (2019) (hard information

Table 13 Performance with full features (hard and soft information)

Model D			Average accuracy (%)	Average recall (%)	Average F1 score (%)	Average AUC (%)
Word2Vec + lda2vec	LSTM	Training	95.68	97.18	95.97	96.18
		Testing	88.36	87.74	88.18	86.57
	XGBoost	Training	94.23	91.86	94.05	94.22
		Testing	85.91	83.71	84.79	85.97
	Random Forrest	Training	99.96	99.87	99.97	99.93
		Testing	85.23	81.58	84.93	85.31
	SVM	Training	96.74	97.12	96.78	96.89
		Testing	84.64	85.28	85.15	85.29
XLNet+lda2vec	LSTM	Training	98.74	97.10	97.54	98.07
		Testing	94.63	92.11	94.59	93.68
	XGBoost	Training	97.98	98.26	97.97	97.99
		Testing	93.21	94.16	93.56	93.79
	Random Forrest	Training	99.86	99.79	99.94	99.98
		Testing	88.27	84.21	88.59	88.68
	SVM	Training	97.88	97.65	98.77	97.95
		Testing	98.74	97.10	97.54	98.07

Figure 10 Performance under different methods

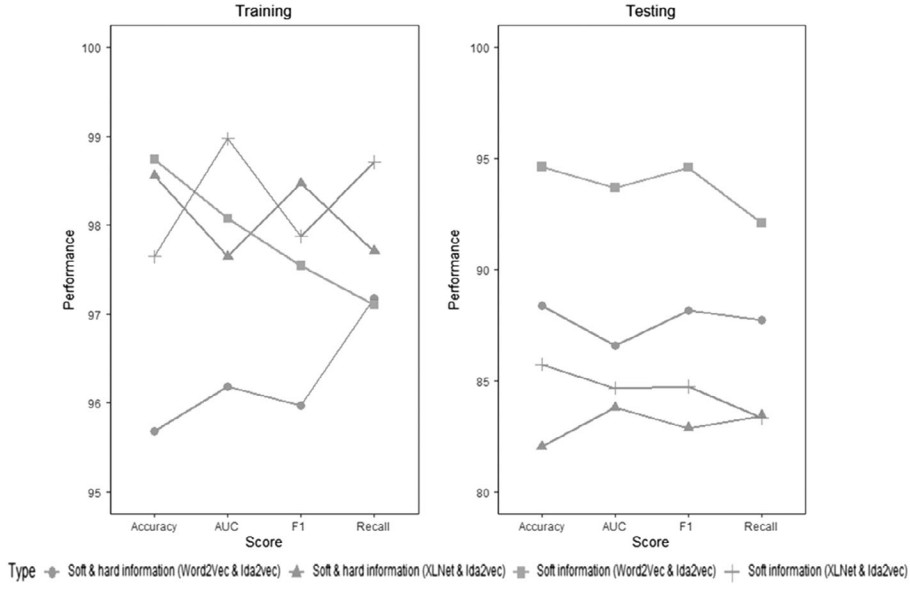


Table 14 Pre versus post the landmark regulation

Model E			Pre the landmark regulation				Post the landmark regulation			
			Average accuracy (%)	Average recall (%)	Average F1 score (%)	Average AUC (%)	Average accuracy (%)	Average recall (%)	Average F1 score (%)	Average AUC (%)
Full features (hard+soft)	LSTM	Training	98.70	96.13	98.26	98.32	98.73	98.89	97.41	97.66
		Testing	91.27	92.40	91.47	92.61	94.63	95.89	94.59	94.66
	XGBoost	Training	99.14	98.87	99.15	99.15	95.11	92.57	95.01	95.13
		Testing	90.28	91.75	90.78	91.38	94.46	93.51	94.68	93.38
	Random Forrest	Training	96.54	97.71	96.61	96.53	99.71	99.40	99.71	99.70
		Testing	89.26	84.93	88.57	89.18	93.66	92.68	92.77	92.56
	SVM	Training	99.71	99.42	99.72	99.70	98.28	99.18	98.31	98.25
		Testing	83.22	83.68	81.75	90.60	91.40	91.46	90.37	90.60
Platform hard information features	LSTM	Training	98.77	96.71	98.37	98.72	98.85	98.31	98.86	98.86
		Testing	82.93	83.14	81.52	82.94	84.64	85.18	83.50	81.67
	XGBoost	Training	88.77	93.71	89.37	88.72	98.85	98.31	98.86	98.86
		Testing	79.93	80.14	79.52	79.94	82.64	81.18	81.50	81.67
	Random Forrest	Training	97.12	97.21	97.21	97.12	93.29	90.42	92.96	93.23
		Testing	75.80	75.88	75.71	75.90	79.93	80.42	79.80	79.94
	SVM	Training	93.95	92.35	93.73	93.92	98.28	98.83	98.26	98.28
		Testing	61.88	66.92	61.63	62.12	62.60	79.40	67.46	62.37

Figure 11 Pre versus post the landmark regulation (LR) performance

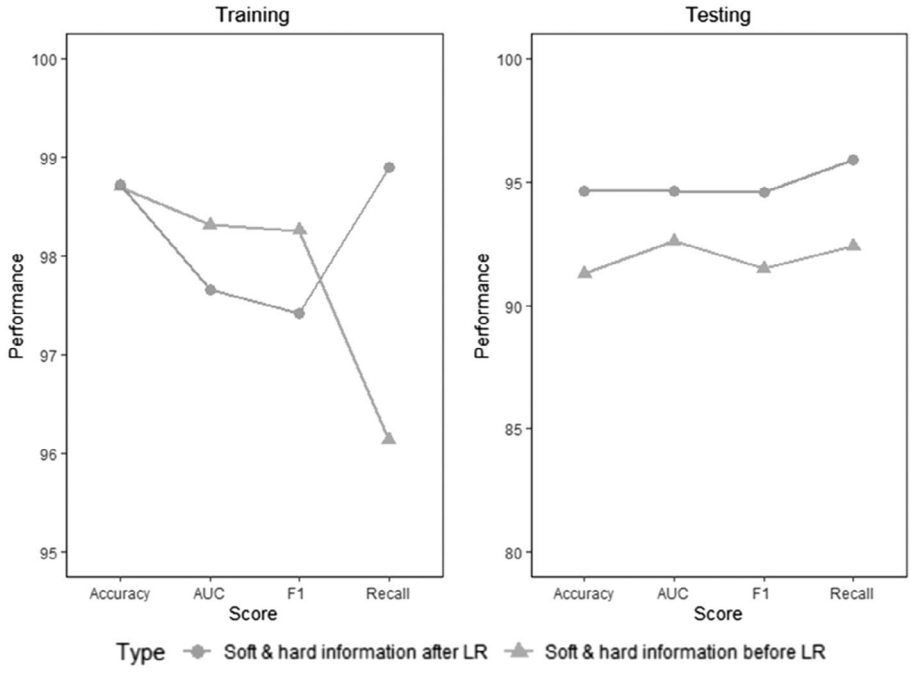


Table 15 Comparison between studies and within this study

Studies	Methods	Feature sets	Accuracy (%)
Xia *et al.* (2019)	Glowworm swarm optimization; Logistic regression	Hard information	83.95
Yoon *et al.* (2019)	Logistic regression	Hard information	87.80–89.90
Fu *et al.* (2020)	Sentiment analysis; EBCC model (deep learning)	Investor reviews & Hard information	77.40–80.34
This study: Model A	Logistic regression	Hard information only	79.51
Model B	LSTM	Hard information only	83.92
Model C1	LSTM+XLNet	Investor reviews only (dynamic)	88.94
Model C2	LSTM+XLNet	Breaking news only (dynamic)	89.26
Model C3	LSTM+lda2vec	Company profiles only (static)	73.24
Model D (Proposed method)	LSTM + XLNet + lda2vec	Full features (hard + soft)	94.63

only) and Fu *et al.* (2020) (investor reviews as additional information). Another factor contributing to our superior performance is that we test different techniques (four machine learning methods, two methods for sentiment features and two methods for topic features) and select the combination with the best performance (*i.e.*, LSTM+XLNet+lda2vec). In contrast, the other three studies only used one method.

Compared to the LSTM model with hard information only, the LSTM+XLNet model with breaking news only and the LSTM+XLNet model with investor reviews only deliver better performance than company profiles. These results imply that China's P2P lending platforms tend to be sensitive to breaking news and investor reviews. At the same time, however, company profiles represent static information with limited disclosure (*e.g.*, during the start-up period). With hard information only (see Table 9), the LSTM and XGBoost models outperform Probit models. However, relative to the statistical approach, all four machine learning techniques deliver better forecasts with the support of sentiment features extracted from breaking news (see Table 11) or from investor reviews (see Table 10). These findings further highlight the critical role of these two sources of soft information in detecting P2P default risk.

8. Conclusions

Under event-driven uncertainty, traditional statistical models established on identified features or relationships, or machine learning approaches isolated from other information sources, cannot predict dramatic changes and the impact of significant events in financial markets. Previous studies have mainly used hard information (Xia *et al.*, 2019; Yoon *et al.*, 2019) and investors reviews (Fu *et al.*, 2020) to assess P2P platform risk. However, little effort has been made to empirically distinguish between different sources of stakeholder powers (investors, government regulators, and companies) through soft information from social media platforms. To the best of our knowledge, this study is the first to quantify the contributions of diversified soft information posted by multiple stakeholders (investors, government regulators, and companies) to risk prediction accuracy. This paper empirically validates the incremental effectiveness of incorporating soft and hard information in P2P default risk assessment and verifies the significant improvement of forecasting performance.

Our findings reveal that real-time soft information from several stakeholders on social media platforms has the most significant performance for default risk evaluation. The findings show that breaking news from government regulators disseminated by public media is more time-sensitive and authentic than investors' self-perceptions and sentimental behaviors toward financial risks, potentially mitigating information asymmetry and allowing financial markets to withstand uncertainty in a short period. Moreover, dynamic soft information (breaking news and investor reviews) contributes to P2P default detection than static soft information (company profiles). The latter is the historical business scope created by the companies during

their start-up period. Government regulators and investors will be able to make more rational decisions when assessing and reacting to unpredictable market risks by tracking dynamic soft data, which may even help determine the company's creditworthiness. By distinguishing the role of different sources of soft information in improving the accuracy of risk prediction, this paper presents an insightful study on risk prediction in the context of China's P2P lending market. At the same time, there are some concerns regarding its scientific novelty.

References

Abbasi, A., C. Albrecht, A. Vance, and J. Hansen, 2012, Metafraud: A meta-learning framework for detecting financial fraud, *MIS Quarterly* 36, pp. 1293–1327.

Abbasi, A., and H. Chen, 2008, CyberGate: A design framework and system for text analysis of computer-mediated communication, *MIS Quarterly* 32, pp. 811–837.

Abrahams, A. S., W. Fan, G. A. Wang, Z. Zhang, and J. Jiao, 2015, An integrated text analytic framework for product defect discovery, *Production and Operations Management* 24, pp. 975–990.

Agarwal, S., and J. Zhang, 2020, FinTech, lending and payment innovation: A review, *Asia-Pacific Journal of Financial Studies* 49, pp. 353–367.

Arner, D., R. Buckley, and D. Zetzsche, 2019, Fintech, regtech and systemic risk, In The rise of global technology risk, Systemic Risk in the Financial Sector, pp. 69–82.

Babaei, G., and S. Bamdad, 2021, A new hybrid instance-based learning model for decision-making in the P2P lending market, *Computational Economics* 57, pp. 419–432.

Bachmann, A., A. Becker, D. Buerckner, M. Hilker, F. Kock, M. Lehmann et al., 2011, Online peer-to-peer lending – a literature review, *Journal of Internet Banking and Commerce* 16, 2.

Bagheri, A., M. Saraee, and F. de Jong, 2014, ADM-LDA: An aspect detection model based on topic modelling using the structure of review sentences, *Journal of Information Science* 40, pp. 621–636.

Bärnighausen, T., J. Bor, S. Wandira-Kazibwe, and D. Canning, 2011, Correcting HIV prevalence estimates for survey nonparticipation using Heckman-type selection models, *Epidemiology* 22, pp. 27–35.

Bastani, K., H. Namavari, and J. Shaffer, 2019, Latent Dirichlet allocation (LDA) for topic modeling of the CFPB consumer complaints, *Expert Systems with Applications* 127, pp. 256–271.

Blei, D. M., A. Y. Ng, and M. I. Jordan, 2003, Latent Dirichlet allocation, *Journal of Machine Learning Research* 3, pp. 993–1022.

Brown, N. C., R. M. Crowley, and W. B. Elliott, 2020, What are you saying? Using topic to detect financial misreporting, *Journal of Accounting Research* 58, pp. 237–291.

Bushway, S., B. D. Johnson, and L. A. Slocum, 2007, Is the magic still there? The use of the Heckman two-step correction for selection bias in criminology, *Journal of Quantitative Criminology* 23, pp. 151–178.

Byanjankar A., M. Heikkila, and J. Mezei, 2015.*Predicting credit risk in peer-to-peer lending: A neural network approach*.

Caliendo, M., and S. Kopeinig, 2008, Some practical guidance for the implementation of propensity score matching, *Journal of Economic Surveys* 22, pp. 31–72.

Cao, N., S. Ji, D. K. W. Chiu, M. He, and X. Sun, 2020, A deceptive review detection framework: Combination of coarse and fine-grained features, *Expert Systems with Applications* 156, p. 113465.

Cecchini, M., H. Aytug, G. J. Koehler, and P. Pathak, 2010, Detecting management fraud in public companies, *Management Science* 56, pp. 1146–1160.

Chen, G., M. Firth, D. N. Gao, and O. M. Rui, 2006, Ownership structure, corporate governance, and fraud: Evidence from China, *Journal of Corporate Finance* 12, 424–448.

Chen, X., B. Huang, and D. Ye, 2018, The role of punctuation in P2P lending: Evidence from China, *Economic Modelling* 68, pp. 634–643.

Cheng, H., and R. Guo, 2020, Risk preference of the investors and the risk of peer-to-peer lending platform, *Emerging Markets Finance and Trade* 56, pp. 1520–1531.

Dai, A. M., and Q. V. Le, 2015, Semi-supervised sequence learning, *Advances in Neural Information Processing Systems* 28.

Devlin J., M.-W. Chang, K. Lee, and K. Toutanova, 2018.*BERT: Pre-training of deep bidirectional transformers for language understanding*.

Dong W., S. Liao, B. Fang, X. Cheng, C. Zhu, and W. Fan, 2014, The detection of fraudulent financial statements: An integrated language model approach, Proceedings - Pacific Asia Conference on Information Systems, PACIS.

Dong, W., S. Liao, and Z. Zhang, 2018, Leveraging financial social media data for corporate fraud detection, *Journal of Management Information Systems* 35, pp. 461–487.

Dorfleitner, G., E.-M. Oswald, and R. Zhang, 2021, From credit risk to social impact: On the funding determinants in interest-free peer-to-peer lending, *Journal of Business Ethics* 170, pp. 375–400.

Dorfleitner, G., C. Priberny, S. Schuster, J. Stoiber, M. Weber, I. de Castro et al., 2016, Description-text related soft information in peer-to-peer lending – Evidence from two leading European platforms, *Journal of Banking & Finance* 64, pp. 169–187.

Duan, J., 2019, Financial system modeling using deep neural networks (DNNs) for effective risk assessment and prediction, *Journal of the Franklin Institute-Engineering and Applied Mathematics* 356, pp. 4716–4731.

Emekter, R., Y. Tu, B. Jirasakuldech, and M. Lu, 2015, Evaluating credit risk and loan performance in online Peer-to-Peer (P2P) lending, *Applied Economics* 47, pp. 54–70.

Farooq, U., and M. A. J. Qamar, 2019, Predicting multistage financial distress: Reflections on sampling, feature and model selection criteria, *Journal of Forecasting* 38, pp. 632–648.

Feng, Y., X. Fan, and Y. Yoon, 2015, Lenders and borrowers' strategies in online peer-to-peer lending market: An empirical analysis of ppdai.com, *Journal of Electronic Commerce Research* 16, pp. 242–260.

Fernández, A. D., M. López-Martín, T. Montero-Romero, F. Martínez-Estudillo, and F. Fernández-Navarro, 2018, Financial soundness prediction using a multi-classification model: Evidence from current financial crisis in OECD banks, *Computational Economics* 52, pp. 275–297.

Fu, X., T. Ouyang, J. Chen, and X. Luo, 2020, Listening to the investors: A novel framework for online lending default prediction using deep learning neural networks, *Information Processing & Management* 57, pp. 102–236.

Gao, M., J. Yen, and M. Liu, 2021, Determinants of defaults on P2P lending platforms in China, *International Review of Economics & Finance* 72, pp. 334–348.

Gao, Y., S. H. Yu, and Y. C. Shiue, 2018, The performance of the P2P finance industry in China, *Electronic Commerce Research and Applications* 30, pp. 138–148.

Ge, R., J. Feng, B. Gu, and P. Zhang, 2017, Predicting and deterring default with social media information in peer-to-peer lending, *Journal of Management Information Systems* 34, pp. 401–424.

Gong, Q., C. Liu, Q. Peng, and L. Wang, 2020, Will CEOs with banking experience lower default risks? Evidence from P2P lending platforms in China, *Finance Research Letters* 36.

Gu, D., T. Lu, P. Luo, and C. Zhang, 2019, The impact of venture capital investment on the performance of peer-to-peer lending platforms: Evidence from China, *Asia-Pacific Journal of Financial Studies* 48, pp. 640–665.

Han, J.-T., Q. Chen, J.-G. Liu, X.-L. Luo, and W. Fan, 2018, The persuasion of borrowers' voluntary information in peer to peer lending: An empirical study based on elaboration likelihood model, *Computers in Human Behavior* 78, pp. 200–214.

He, Q., and X. Li, 2021, The failure of Chinese peer-to-peer lending platforms: Finance and politics, *Journal of Corporate Finance* 66, 101852.

Heckman, J. J., 1976, The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models, In Annals of Economic and social measurement, Vol. 5, number 4 (NBER).

Heckman, J. J., 1979, Sample selection bias as a specification error, *Econometrica: Journal of the Econometric Society* 47, pp. 153–161.

Herzenstein, M., S. Sonenshein, and U. M. Dholakia, 2011, Tell me a good story and i may lend you money: The role of narratives in peer-to-peer lending decisions, *Journal of Marketing Research* 48, pp. 138–149.

Irfan, S., 2011, Modeling wages of females in the UK, *International Journal of Business and Social Science* 2, 11.

Jiang, J., L. Liao, Z. Wang, and X. Zhang, 2021, Government affiliation and peer-to-peer lending platforms in China, *Journal of Empirical Finance* 62, pp. 87–106.

Kaminski, K. A., W. T. Sterling, and L. Guan, 2004, Can financial ratios detect fraudulent financial reporting? *Managerial Auditing Journal* 19, pp. 15–28.

Katmon, N., and O. Farooque, 2017, Exploring the impact of internal corporate governance on the relation between disclosure quality and earnings management in the UK listed companies, *Journal of Business Ethics* 142, pp. 345–367.

Kim, A., and S.-B. Cho, 2019, An ensemble semi-supervised learning method for predicting defaults in social lending, *Engineering Applications of Artificial Intelligence* 81, pp. 193–199.

Kim, D., and J. Kim, 2020, Distinctive features of student borrowers and suboptimal investor decision-making: Evidence from the P2P lending market, *Asia-Pacific Journal of Financial Studies* 49, pp. 860–883.

Klapper, L. F., and I. Love, 2004, Corporate governance, investor protection, and performance in emerging markets, *Journal of Corporate Finance* 10, pp. 703–728.

Li, J., S. Hsu, Z. Chen, and Y. Chen, 2016, Risks of P2P lending platforms in China: Modeling failure using a cox hazard model, *Chinese Economy* 49, pp. 161–172.

Li, Lin C.-T. and Chen S.-F., 2017, Micro-lending default awareness using artificial neural network (Taichung, Taiwan, Association for Computing Machinery).

Liang, L., and X. Cai, 2020, Forecasting peer-to-peer platform default rate with LSTM neural network, *Electronic Commerce Research and Applications* 43, p. 100997.

Lin, M., N. R. Prabhala, and S. Viswanathan, 2013, Judging borrowers by the company they keep: Friendship networks and information asymmetry in online peer-to-Peer Lending, *Management Science* 59, pp. 17–35.

Liu, P., H. Li, and G. Sun, 2018, P2P lending platform risk observing method based on short-time multi-source regression algorithm, *2018 Ieee International Conference on Communications*, pp. 1–6.

Liu, Q., L. Zou, X. Yang, and J. Tang, 2019, Survival or die: a survival analysis on peer-to-peer lending platforms in China, *Accounting and Finance* 59, pp. 2105–2131.

Liu, X., 2020, Analyzing the impact of user-generated content on B2B Firms' stock performance: Big data analysis with machine learning methods, *Industrial Marketing Management* 86, pp. 30–39.

Luther, J., 2020, Twenty-first century financial regulation: P2P Lending, fintech, and the argument for a special purpose fintech charter approach, *University of Pennsylvania Law Review* 168, pp. 1013–1059.

Malekipirbazari, M., and V. Aksakalli, 2015, Risk assessment in social lending via random forests, *Expert Systems with Applications* 42, pp. 4621–4631.

Maskara, P. K., E. Kuvvet, and G. Chen, 2021, The role of P2P platforms in enhancing financial inclusion in the United States: An analysis of peer-to-peer lending across the rural-urban divide, *Financial Management.* 50, pp. 747–774.

Matthiessen C., 2014, An introduction to functional grammar, introduction.

Michels, J., 2012, Do unverifiable disclosures matter? Evidence from peer-to-peer lending, accounting review, *The Accounting Review* 87, pp. 1385–1413.

Moody C. E., 2016, Mixing Dirichlet Topic Models and Word Embeddings to Make lda2vec arXiv, arXiv, pp.7 pp.-7 pp.

Mou, J., G. Ren, C. Qin, and K. Kurcz, 2019, Understanding the topics of export cross-border e-commerce consumers feedback: An LDA approach, *Electronic Commerce Research* 19, pp. 749–777.

Oliveira, N., P. Cortez, and N. Areal, 2017, The impact of microblogging data for stock market prediction: Using Twitter to predict returns, volatility, trading volume and survey sentiment indices, *Expert Systems with Applications* 73, pp. 125–144.

Pennington J., R. Socher, and C. Manning, 2014, Glove: Global vectors for word representation, In *Proceedings of the 2014 conference on empirical methods in natural language processing*, pp. 1532–1543.

Peters M., M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, 2018, Deep contextualized word representations.

Sartori, A. E., 2003, An estimator for some binary-outcome selection models without exclusion restrictions, *Political Analysis* 11, pp. 111–138.

Summers, S. L., and J. T. Sweeney, 1998, Fraudulently misstated financial statements and insider trading: An empirical analysis, *Accounting Review* 73, pp. 131–146.

Suominen, A., and H. Toivanen, 2016, Map of science with topic modeling: Comparison of unsupervised learning and human-assigned subject classification, *Journal of the Association for Information Science and Technology* 67, pp. 2464–2476.

Weiss G., K. Pelger, and A. Horsch, 2010, Mitigating adverse selection in P2P Lending – Empirical Evidence from Prosper.com, SSRN Electronic Journal.

Xia, N. Z., X. Zhu, and L. Ni, 2019, A novel key influencing factors selection approach of P2P lending investment risk, *Mathematical Problems in Engineering* 2019, pp. 1–12.

Xia, Y., C. Liu, and N. Liu, 2017, Cost-sensitive boosted tree for loan evaluation in peer-to-peer lending, *Electronic Commerce Research and Applications* 24, pp. 30–49.

Xu, Q., Q. Li, C. Jiang, J. Wu, and X. Zhang, 2020, Social media, interaction information and stock market efficiency: Evidence from the Shenzhen stock exchange easy interaction platform in China, *Asia-Pacific Journal of Accounting & Economics*, pp. 1–28.

Yang, C. P., F. Shi, and C. Wen, 2018, Internet finance: Its uncertain legal foundations and the role of big data in its development, *Emerging Markets Finance and Trade* 54, pp. 721–732.

Yang, D. Z., Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le, 2019, XLNet: Generalized autoregressive pretraining for language understanding, *Advances in Neural Information Processing Systems* 32.

Yang, J., and D. Luo, 2017, The P2P risk assessment model based on the improved AdaBoost-SVM algorithm, *Journal of Financial Risk Management* 6, pp. 201–209.

Yoon, Y., Y. Li, and Y. Feng, 2019, Factors affecting platform default risk in online peer-to-peer (P2P) lending business: An empirical study using Chinese online P2P platform data, *Electronic Commerce Research* 19, pp. 131–158.

Zadrozny, B., 2004, Learning and evaluating classifiers under sample selection bias, In Proceedings of the twenty-first international conference on Machine learning, pp. 114.

Zhang, W., C. Wang, Y. Zhang, and J. Wang, 2020, Credit risk evaluation model with textual features from loan descriptions for P2P lending, *Electronic Commerce Research and Applications* 42.

Zhong, N., and D. A. Schweidel, 2020, Capturing changes in social media content: A multiple latent changepoint topic model, *Marketing Science* 39, pp. 827–846.

Zirn, C., and H. Stuckenschmidt, 2014, Multidimensional topic analysis in political texts, *Data & Knowledge Engineering* 90, pp. 38–53.