Human Expertise versus Black Boxes:

Predicting Corporate Defaults via Domain-Informed Textual Analysis

Hui Bu*, Wenjin Tang, Jun Tu, Junbo Wang, Shouyang Wang

ABSTRACT

We develop a transparent textual analysis framework that integrates financial expertise to predict corporate bond defaults in China. Our domain-informed news metric detects latent distress signals up to six months ex-ante, significantly outperforming black-box data-driven models (e.g., ChatGPT, BERTopic) and general dictionaries. Using this metric, we validate a "media governance hypothesis": state-controlled and market-oriented media provide complementary risk information but engage in ownership-dependent selective disclosure for strategically important firms. These findings demonstrate that embedding human domain logic into text analysis enhances predictive interpretability, while illuminating how institutional structures shape information provision in capital markets.

Keywords: textual analysis; knowledge encoding; default prediction; media news

JEL classification: G32, C55, C53

---

*Hui Bu is at School of Economics and Management, Beihang University. Wenjin Tang is at College of Business, Southern University of Science and Technology. Jun Tu is at Antai College of Economics and Management, Shanghai Jiaotong University. Junbo Wang is at College of Business, City University of Hong Kong. Shouyang Wang is at Academy of Mathematics and Systems Science, Chinese Academy of Sciences. Prof. Hui Bu is corresponding author. Email: buhui@buaa.edu.cn. Address: No. 37 Xueyuan Road, Beijing 100191, China. The authors contribute equally to this paper, and the authors are ordered by surname. Dr. Wenjin Tang's Email: tangwj_1995@buaa.edu.cn; Prof. Jun Tu's Email: tujun2022@sjtu.edu.cn ; Prof. Junbo Wang's Email: jwang2@cityu.edu.hk; Prof. Shouyang Wang's Email: sywang@amss.ac.cn.

Acknowledgement: This work was supported by Prof. Bu's grants from the Key Project of the National Natural Science Foundation of China (72531002) and the Key Project of National Key R&D Program of China (2023YFC3304705), this work was partly supported by Prof. Tu's grant from the National Natural Science Foundation of China (72342023), the Prof. Shouyang Wang's grant from the National Natural Science Foundation of China (71988101), and Dr. Tang's grant from the National Natural Science Foundation of China (72503087).

We acknowledge helpful comments from Kewei Hou, Xindan Li, and Jing Xie (Beihang University workshop); Hui Wang and Yifeng Zhu (CUFE seminar); Qiang Ye (University of Science and Technology of China workshop); Yong Zhou and Yun Shi (East China Normal University seminar); Wei Zhang, Xiong Xiong, Xu Feng, and Yongjie Zhang (Tianjin University seminar); Yiming Wang and Xi Wang (Peking University seminar); and Yan Zeng (Sun Yat-sen University seminar). We also thank the participants at these workshops and seminars for their constructive feedback.

This paper received the four Excellent Paper Award at FSERM 2025, BIFE 2023 & ICFT 2023, the 5th Academic Conference on Quantitative Finance and Insurance (2023), 4th Annual Academic Conference of the Risk Management Branch of the Chinese Society of Optimization, Overall Planning and Economic Mathematics (2023). We thank the session chairs, discussants, and participants at these conferences, as well as those at 2025 INFORMS international meeting, 2025 WEHIA Winter Meeting, CFAM 2024, CIRF & CFRI 2024, the 3rd China Financial Frontier Academic Forum (2024), the Conference of Recent Advances on Quantitative Finance (2023), CSCR II, FIRM 2022 for their helpful comments.

I. Introduction

The financial domain distinguishes itself from general contexts by its dual foundation: rigorous theoretical frameworks and extensive, often tacit, practitioner experience. While recent literature has successfully integrated structured economic models into machine learning, for instance, Chen et al. (2023) propose a transfer learning framework that incorporates the Black-Scholes model to enhance option pricing, capturing implicit expertise remains a significant frontier challenge. Unlike closed-form equations, a vast amount of financial wisdom resides in fragmented reasoning, intuition, and discretionary judgment. This tacit knowledge resists straightforward formalization, creating a critical gap in how machines comprehend financial contexts.

In the current landscape of Large Language Models (LLMs), the dominant paradigm attempts to bridge this gap through purely data-driven paradigms. Methods such as corpus-based pretraining (e.g., BERT, ChatGPT; Devlin et al., 2019) or domain-specific fine-tuning (e.g., FinBERT; Araci, 2019; Huang et al., 2023) rely on inferring latent structures solely from massive textual exposure. While effective at capturing linguistic probabilities and general semantic associations, these models face inherent identification challenges in specialized financial inference. As noted by Bender et al. (2021), purely data-driven architectures excel at manipulating linguistic form but may lack the grounded understanding required to distinguish valid economic signals from surface-level statistical correlations. Consequently, without explicit theoretical guidance, such approaches remain vulnerable to noise and bias (Fan et al., 2024), risking the prioritization of spurious patterns over causal logic. Furthermore, the reliability of these models is challenged by contamination of pretraining data, which casts doubt on whether performance gains stem from genuine reasoning capabilities or mere memorization (Jiang et al., 2024; Dong et al., 2024)[1]. In high-stakes domains where decision-making demands auditability and robust expert logic, reliance on opaque probabilistic associations presents a significant hurdle. A central question, therefore, remains: how can implicit domain expertise, which is shaped by decades of theoretical development and industry practice, be systematically encoded for

[1] Dong et al. (2024) document that ChatGPT is susceptible to contamination on the HumanEval dataset.

machines to enhance text-based prediction and decision-making?

We examine the integration of domain knowledge in the high-stakes context of predicting first-time corporate bond defaults. This setting offers an ideal laboratory precisely because traditional information channels are often silent prior to initial defaults. Firms frequently exhibit no observable deterioration in financial ratios until distress is imminent, and traditional gatekeepers fail to provide timely warnings. In the Chinese market, for instance, issuers rated AA or higher routinely default without prior downgrades (see Section I of the Internet Appendix). Given this well-documented informational void, where structural models and credit ratings lag, the ability to extract implicit early-warning signals from unstructured text is paramount. While prior literature establishes the value of news for asset pricing (Tetlock, 2007; Chen, Kelly, and Xiu, 2022), systematic evidence regarding its long-horizon predictive power remains limited. We therefore exploit this setting to test whether integrating human expertise can recover predictive signals from news that purely data-driven models miss.

To address this question, we propose a novel framework that transparently operationalizes implicit financial expertise into a machine-readable format. Unlike approaches that rely on latent feature representations inferred indirectly from vast corpora, our method explicitly encodes established credit risk logic, grounded in the "6Cs" of credit evaluation, rating methodologies, and prior academic studies, into the textual analysis process. We implement a two-stage “knowledge deconstruction” procedure to bridge expert logic and textual data. First, we employ expert judgment to decompose credit risk into a set of interpretable themes. Second, we instantiate these themes into specific vocabulary by anchoring expert-identified seed words to the linguistic patterns of the news corpus. This process yields a hierarchical risk-theme-word dictionary tailored for corporate credit risk evaluation.

Our empirical analysis exploits a comprehensive dataset of all corporate and enterprise bonds issued in China from 2014 to 2022, matched with over 2.83 million firm-specific news headlines filtered from the *Choice* database. Using our theme-anchored framework, we construct a credit risk dictionary comprising 21 credit-risk topics and 323 associated words and phrases. We emphasize that this dictionary serves as a proof of concept for our construction framework rather than an exhaustive compendium; however,

the fact that a parsimonious, expert-anchored lexicon achieves strong predictive efficacy implies that our methodology is highly scalable. Building on this dictionary, we construct a novel measure, news-implied theme frequency (*NTITF*), which applies an inverse topic-frequency weighting scheme to capture the intensity of adverse news coverage for each issuer. We then assess the incremental information content of *NTITF* using a conditional logistic regression model following Campbell, Hilscher, and Szilagyi (2008) and evaluate its predictive value through extensive out-of-sample tests across multiple machine-learning models.

Our results document three sets of findings that highlight the value of integrated expertise. First, *NTITF* provides substantial incremental predictive power beyond financial ratios and macroeconomic indicators. Crucially, the measure delivers robust early-warning signals that significantly improve out-of-sample recall, and F_1 scores up to six months prior to default, a horizon where traditional credit ratings typically lag. These findings suggest that expert-guided text analysis effectively captures latent risks, such as corporate misconduct, shareholder misappropriation, and legal risk, well before they materialize in accounting statements.

Second, to validate the necessity of expert logic, we conduct a rigorous "horse race" against leading data-driven alternatives. We benchmark our domain-informed method against both general-purpose large language models (e.g., ChatGPT-4.1) and embedding-based topic models (BERTopic; Grootendorst, 2022) initialized with BERT and FinBERT embeddings (Devlin et al., 2019; Araci, 2019). Empirical evidence shows that our expertise-integrated approach consistently outperforms generic architectures in detecting distress signals. This finding supports the view that for specialized, high-noise tasks, the depth of domain expertise outweighs the breadth of generic data training, particularly when interpretability and the identification of latent risks are prioritized.

Third, we exploit the unique institutional heterogeneity of the Chinese market to unpack the sources of the media's predictive value. We document that state-controlled and market-oriented media play complementary roles, supporting a "*media governance hypothesis"*. State-controlled outlets leverage their privileged access to provide stable, authoritative signals, yet they engage in strategic, selective disclosure by suppressing negative credit news for politically important firms (e.g., CSOEs and SOEs) to preserve

social stability. Conversely, market-oriented outlets act as independent information substitutes, supplying critical incremental risk signals specifically in opaque environments characterized by high information asymmetry. We further trace this complementarity to a distinct specialization in information production: state-controlled media focus on regulation and corporate strategy (particularly for SOEs), whereas market-oriented media deliver incremental value by uncovering idiosyncratic risks such as debt crises and corporate deception.

This study contributes to the literature in three ways. First, methodologically, we contribute a systematic framework for constructing domain-specific dictionaries that operationalizes implicit expert logic. Our theme-anchored approach explicitly maps the "6Cs" principle and established credit analysis frameworks into textual patterns. This provides a replicable protocol for translating abstract financial reasoning into precise indicators, ensuring that the resulting dictionary captures the nuance of domain expertise while retaining the transparency inherent to lexical methods.

Second, we advance the literature on credit risk and information intermediation by demonstrating that unstructured media coverage is an effective information source for assessing corporate credit risk. We show that in opaque markets where credit risk evaluation models based on financial ratios fail and credit ratings lag, expert-filtered news serves as a leading indicator, identifying latent distress signals up to 6 months in advance. This finding validates the media's governing role in resolving information frictions and anticipating tail events when standard information channels are impaired.

Third, we illuminate the political economy of information intermediation by validating a "media governance hypothesis" that reconciles the debate between state-controlled and market-oriented media. We uncover a specialized division of labor rooted in heterogeneous information endowments and incentives: state-controlled media leverage privileged access to official channels to disseminate authoritative coverage of regulation and policy interpretation (particularly for SOEs), thereby providing stable early-warning signals. Conversely, market-oriented media function as independent investigators, contributing distinct incremental information by uncovering idiosyncratic risks, such as debt distress and operational status, especially in low-disclosure environments. These findings demonstrate that ownership structures dictate not only the quality but also the

nature and target of information provision in capital markets.

The remainder of the paper is organized as follows. Section 2 reviews related literature. Section 3 describes the data and the construction of our domain-informed dictionary. Section 4 evaluates the predictive performance and early-warning capability of *NTITF*. Section 5 compares our approach with LLM-based and general-purpose benchmarks. Section 6 analyzes the sources of information and media heterogeneity. Section 7 concludes.

II. Institutional Background

A. Stylized Facts of Corporate Bond Defaults in China

The inaugural corporate bond default in China occurred in March 2014 with the "11 Chaori Bond," issued by the publicly listed Shanghai Chaori Solar Energy. This landmark event is widely cited as a watershed moment that shattered the long-standing assumption of implicit guarantees in the Chinese financial market. Using data sourced from the WIND database, we analyze the issuer structure of bond defaults from 2014 through 2022. Our sample covers five major types of instruments traded in the inter-bank market and the Shanghai (SSE) and Shenzhen (SZSE) stock exchanges: listed company bonds, enterprise bonds, short-term financing bills, medium-term notes, and private placements.

Table IA.1 in the Internet Appendix reports the detailed statistics. We identify 264 issuers that defaulted for the first time, corresponding to 317 initial bond defaults and 965 subsequent defaults. Among the 264 first-time defaulters, 67 are listed companies, and 42 are state-owned enterprises (SOEs), of which 12 are central state-owned enterprises (CSOEs), according to the issuers' nature at the time a corporate bond is issued. By the end of 2022, the market also included 38,551 bonds that matured without default, issued by 3,932 firms[2]. The non-defaulting issuers sample with outstanding debt comprises 3,748 firms, consisting of 1,113 listed companies, 2,205 SOEs, and 705 CSOEs. The distinct ownership structures of issuers and the variation in issuance frameworks make the Chinese credit bond market an ideal laboratory for research.

[2] The ratio of the defaulted issuers for the first time and non-default issuers in the Chinese corporate bond market during the period from January 2014 to December 2022 is close to 1:15.

Based on the statistics in Tables IA.II and IA.III of the Internet Appendix, we document six stylized facts regarding the dynamics of the Chinese corporate bond market: First, default severity has escalated significantly. The annual frequency and total value of defaults have surged to four and six times their historical averages, respectively. Second, the implicit guarantee for SOEs has eroded. The resilience previously observed in state-owned enterprises has weakened since 2015, with further deterioration after 2020. Third, first-time defaults have become the norm. Since 2018, the number of issuers defaulting for the first time has remained elevated, now accounting for over 90% of all defaulting issuers. Fourth, high ratings fail to preclude default. Since 2015, even issuers rated AA and AAA have frequently defaulted without prior downgrades. Fifth, pre-default ratings are deceptively high. Predicting distress is challenging because nearly 70% of issuers hold investment-grade ratings immediately prior to default, with over 41% rated AA or higher. Finally, rating agencies exhibit significant lag. External agencies tend to be reactive, typically downgrading distressed firms only after a crisis has materialized. The diminishing effectiveness of external credit ratings underscores the difficulty of forecasting first-time defaults, thereby providing the practical motivation for this study.

B. The Institutional Background of Chinese Media

A distinct feature of China's media landscape is the prevalence of state-controlled entities, creating an ownership structure different from those of other major markets. Based on the "List of Internet News and Information Source Units" released by the Cyberspace Administration of China (CAC) on October 20, 2021, we identify 38 state-controlled news units operating 241 specific media outlets, based on our news data collection period from May 1, 2015, to December 31, 2022, sourced from the Choice database (*Eastmoney*)[3]. These outlets span diverse channels, including newspapers, official websites, mobile applications, and social media accounts (e.g., Weibo and WeChat). In this study, we classify articles from these 241 outlets as "state-controlled media" and those from all others as "market-oriented media." This distinction is critical for isolating the divergent functions

[3] The data of the news headline contains title, publishing source media, showtime, URL link, and other labels. We can collect the main text of each news item from its URL.

and incentives of media based on their institutional affiliation.

In our news data sample, we identify 13,253 media outlets, of which 3,928 provide firm-specific content. The news sample comprises 9,043,976 news articles. Tables IA.IV and IA.V in the Internet Appendix report the top 20 media outlets ranked by total news volume and by the volume of credit risk-related news published prior to first-time defaults.

Table IA.V details the contribution of different media types. State-controlled media produced 91,513 credit risk-related articles, accounting for 17.52% of the total volume, while market-oriented outlets contributed 367,754 articles. Despite their smaller aggregate volume, state-controlled media hold a dominant position among top-tier influencers. Five state-controlled outlets appear in the top 20 by credit risk news volume, with three ranking in the top 10: *Securities Times, China Securities Journal, and Securities Daily*. These statistics underscore the pivotal role of state-controlled media in disseminating and interpreting material corporate information in the Chinese capital market.

On the other side, within the top 20 outlets, market-oriented media exhibit the highest concentration of pre-default warning signals. Specifically, the three outlets with the highest ratio of pre-default credit risk news to total issuer coverage are all market-oriented: *Chinese Securities Journal* (16.96%), *The Paper* (16.38%), and *The Beijing News* (13.68%). These findings suggest that while state-controlled and market-oriented media both contribute to information discovery, market-oriented outlets play a relatively more active role in disseminating early warning signals regarding credit risk.

Furthermore, traditional newspapers dominate the landscape of early warning dissemination, as shown in Table IA.V. Of the top 20 outlets providing such information, 16 are traditional newspapers, occupying the top four positions. In contrast, digital aggregators such as *Eastmoney.com* and *Jiemian.com* (despite ranking first and fifth in total news volume, respectively), contribute a disproportionately low volume of unique early warning content. After controlling for total coverage, these platforms exhibit the lowest density of risk-relevant information among the top 20. This pattern likely reflects their operational model, which relies heavily on reposting content from primary sources. These results underscore the enduring value of traditional investigative journalism in uncovering latent credit risks.

III. Related Literature

Compared with the rapid development of alternative-data variables for personal credit scoring (e.g., social media data, call association relationships, electronic footprint data, e-commerce platform data), the development of new firm-level variables for corporate credit evaluation and bond default prediction has advanced more slowly. The cornerstone of credit risk evaluation remains firm-disclosed financial ratios, such as Altman's z-score model for manufacturer bankruptcies proposed in 1968, the ZETA model proposed by Altman et al., proposed in 1977, Ohlson's O-score logit models for non-manufacturers proposed in 1980, and conditional logit models for financial distress and bond defaults, for example, Campbell et al. (2008). Non-financial variables typically fall into three categories: macroeconomic, industry, sovereign credit, and regulatory indicators (Agarwal et al., 2016); securities market characteristics such as market capitalization, stock price dynamics, volatility, and valuation ratios (Campbell et al., 2008; Olson et al., 2012; Tian et al., 2015); and firm-level managerial and governance features—dividend policies, share pledging, ownership structures, internal control quality, and qualitative records such as litigation, filing histories, and audit opinions (Altman et al., 2010). Together, these variables improve default-risk measurement but still struggle to incorporate unstructured or forward-looking signals, motivating approaches that exploit text and other big-data sources.

Over the past decade, the growing availability of textual data—from corporate disclosures (Loughran and McDonald, 2014), analyst reports, credit rating reports (Agrawal et al., 2016), earnings calls, bond prospectuses, and loan documents to financial news (Tetlock, 2007; Tetlock et al., 2008)—has introduced qualitatively rich information into empirical finance beyond traditional quantitative indicators. Common approaches include linguistic tone analysis (Tetlock, 2007; Agrawal et al., 2016), sentiment analysis (Li et al., 2020), topic modeling, document similarity, and readability (Loughran and McDonald, 2014). Research on news-based measures for credit risk evaluation remains nascent.

Early text analysis in finance studies primarily relied on general-purpose dictionaries, yet the evidence on whether news contains incremental information beyond financial ratios and market variables remains mixed under this approach (e.g. Tetlock, 2007; Tetlock et al.,

2008). The use of general-purpose dictionaries in financial applications faces well-documented limitations, including inaccurate tone classification (Loughran & McDonald, 2011), limited domain applicability (Loughran & McDonald, 2016), and a lack of contextual sensitivity (Bochkay et al., 2022). Even domain-specific dictionaries developed for one type of financial text may not transfer well to other contexts. Loughran and McDonald (2016), for example, cautioned that the Loughran and McDonald (2011)'s dictionaries derived from 10-K filings may require substantial adaptation for settings such as conference calls or social media, where tone and language usage differ markedly. Some work shows that carefully designed, context-specific lexicons markedly improve the accuracy and informativeness of tone and content measurement in financial text analysis (Price et al., 2012; Davis et al., 2015). To accomplish specialized tasks, task-specific dictionaries must be developed. Yet a central question remains: how can domain knowledge be effectively integrated into textual analysis in a systematic and replicable manner?

Over the past decade, computational linguistics has witnessed rapid advancements in deep-learning-based natural language processing (NLP) algorithms. Early contextual embedding models, such as ELMo, laid the groundwork for capturing word meaning in context, while generative pretraining models like OpenAI GPT and bidirectional encoder representations from transformers (BERT) introduced transformer-based architectures capable of learning contextualized embeddings from massive corpora (Devlin et al. 2019; Huang et al., 2023). These large language models (LLMs) have demonstrated substantial performance improvements over traditional NLP algorithms in tasks such as language translation, named entity recognition, and sentiment classification of general text (Devlin et al., 2019). Their growing capabilities have also begun to reshape empirical finance. Recent studies demonstrate that LLMs such as ChatGPT can generate economically meaningful forecasts, including stock return prediction (Chen, Kelly, and Xiu, 2022) and corporate bankruptcy assessment (Bie et al., 2024), among other applications.

A key feature of LLMs is their ability to learn general semantic and syntactic patterns from large volumes of unlabeled text through unsupervised pretraining, as exemplified by BERT, which was pretrained on Wikipedia and BookCorpus (~3.3 billion tokens) to capture nuanced language structures (Devlin et al., 2019). LLMs can then be fine-tuned on smaller, task-specific datasets, combining unsupervised pretraining with supervised

learning, a paradigm that underpins domain-specific adaptations such as FinBERT (Araci, 2019; Huang et al., 2023). Despite their predictive performance, LLMs share the limitations of deep learning models: they are often described as "black boxes", prone to internalizing irrelevant or misleading latent knowledge, and sensitive to noise or spurious correlations in the training corpus. While they capture general linguistic and financial language patterns, LLMs may struggle to internalize task-specific or domain-specific expertise, particularly for highly specialized financial prediction tasks.

Building on these insights, this study proposes a complementary approach: directly embedding domain-specific expertise into textual analysis through a theme-anchored, knowledge-informed dictionary. Rather than relying solely on large-scale unsupervised pretraining, our method maps expert-identified credit-risk topics to their linguistic representations in the corpus, enabling machine learning models to leverage structured domain knowledge alongside textual patterns. This approach enhances interpretability, reduces reliance on purely statistical associations, and allows more stable extraction of predictive signals for specialized tasks, such as first-time corporate bond defaults, where general language patterns may be insufficient to capture nuanced risk information.

IV. Converting Domain-specific Expertise to Textual Analysis

A. Constructing a Domain-informed Credit Risk Dictionary

The dictionary approach serves as a cornerstone in accounting and finance literature for quantifying textual information, prized for its straightforward implementation and, most importantly, its interpretability. However, the efficacy of this method depends heavily on the quality of the underlying lexicon, and general-purpose dictionaries often introduce excessive noise when applied to specific contexts. Prioritizing interpretability over complexity, we adhere to the dictionary paradigm but enhance its precision by integrating domain knowledge. Specifically, we implement a two-stage knowledge deconstruction procedure to align expert logic with textual data. Our objective is to construct a specialized dictionary that mitigates noise and captures the true information value of news, providing a transparent and replicable framework for domain-specific text analysis.

The two-stage knowledge deconstruction procedure for a hierarchical risk-theme

word-dictionary tailored for corporate credit risk evaluation is illustrated in Figure 1. First, we adopt expert judgment to decompose latent credit risk into a set of interpretable themes. The domain knowledge synthesized in this study derives from three authoritative sources: the structural frameworks of credit rating agencies, the “6Cs” of credit principle, and key risk indicators documented by prior studies. Specifically, we consolidate the topics and perspectives used by eight external credit risk rating agencies in the Chinese bond market, including S&P, Fitch, Moody's, Dongfang Jincheng, Zhongchengxin, Dagong International, United Credit, Shanghai New Century (Table IA.VI of Internet Appendix), integrating them with the “6Cs” of Credit, a fundamental framework used by lenders to evaluate a borrower's creditworthiness. Additionally, we collect the variables used in prior studies of our reference (Table IA.VII of the Internet Appendix). As reported in literature review section, the non-financial variables are mainly three categories: variables of macro-economy, industry, sovereign credit, and regulatory environment (Agarwal et al., 2016; among others); variables of the securities market, including the market value, stock price, volatility, P/E ratio (Campbell et al., 2008; Olson et al., 2012; Tian et al., 2015); and variables of enterprises management, including dividends, pledge of shares, ownership structure, internal governance, qualitative data relating to such variables as legal action by creditors to recover unpaid debts, company filing histories, comprehensive audit report/opinion data and firm-specific characteristics (Altman et al., 2010).

We apply a systematic set of filtering rules to synthesize the selected domain knowledge into coherent risk themes. Specifically: (1) we incorporate all topics and evaluative perspectives utilized by major credit rating agencies; (2) we retain the dimensions of the "6Cs" framework that are observable to external parties; and (3) we include key predictors validated in prior credit risk literature. This process yields 21 focal topics essential for credit risk evaluation, including asset quality and financial leverage, profitability, cash flow adequacy, financial reports and audit opinions, market risk, credit risk rating, debt crisis, production and operation status, asset freeze, asset disposal and financing capability, legal risk, development strategy, strategic investors, shareholding pledging and shareholder reduction, shareholder misappropriation, internal control, personnel management, violation of laws and rules, regulatory inquiry, deception, and financial fraud.

In the second step, we instantiate these themes into specific vocabulary by anchoring expert-identified seed words from domain materials to the linguistic patterns of the news headline corpus. We first extract the expert-identified seed words directly from domain materials to serve as the conceptual core of each theme. To ensure coverage of actual media usage, we expand this seed set by analyzing collocation patterns in the news headline corpus[4]. Specifically, we adopt words or phrases if they are: (1) explicitly identified as seed words by experts, or (2) high-frequency tokens that exhibit strong semantic links to the seed words in our training sample. Through this procedure, we ultimately construct a specialized credit risk dictionary comprising 323 distinct words/phrases across the 21 topics, including 156 bigrams (two-word combinations) that capture nuanced semantic meanings. We emphasize that this dictionary serves as a proof of concept for our construction framework rather than an exhaustive compendium; however, the fact that a parsimonious, expert-anchored lexicon achieves strong predictive efficacy implies that our methodology is highly scalable.

[Figure 1. Framework for constructing a domain-specific credit risk dictionary.]

A. Domain-informed Topic-driven News Metric

To explore the usefulness of our proposed domain-informed credit risk dictionary, we construct a topic-driven credit risk news metric based on the number of credit risk topics, with topic-related weighting. Taking topic information into account can address the limitation of the most commonly used measure in the previous literature, which counts the number of negative words, assuming that all words in a dictionary carry the same weight, as noted by Bochkay et al. (2022). We adopt inverse topic frequency (*ITF*) as a topic-related weighting scheme, which follows the idea of inverse document frequency (*IDF*) weights used by Loughran and McDonald (2011). The news measure of firm *i* at day *t* using firm-level credit risk-related news released in the period $D(t)$ before *t*, as shown in Eq.(1).

$$NTITF_{it} = \sum_k \sum_{j \in D(t)} TMentioned_{ijk} \cdot \ln\left(\frac{N_m}{N_{k,m}}\right), \quad (1)$$

where $TMentioned_{ijk}$ is a dummy variable that equals 1 when news headline *j* contains

[4] By using the Chinese text segmentation toolkit for natural language processing (NLP) *Jieba*, we can find the words and phrases that usually used in the news headlines of defaulted issuers in our training sample.

the credit risk words of topic *k* for news of firm *i* released in the period $D(t)$ before *t*, for example, setting the calculation period 3-month in this study (i.e., $D(t) = [t-90, t]$). The term $\ln\left(\frac{N_m}{N_{k,m}}\right)$ represents inverse topic frequency updated monthly, which assigns more weight to less frequently mentioned topics and gives less weight to more frequently mentioned topics N_m represents aggregate count of credit risk-related news from the beginning of news collection period until the calendar month immediately preceding the calendar month of news published date *t*, i.e., $m = \text{month}(t) - 1$. $N_{k,m}$ represents the aggregate count of credit risk related news headlines containing the words in topic *k* until preceding month *m*.

We also provide alternative news metrics using a different weighting scheme based on word count. More detailed information can be found in the Table A. I of the Appendix. The summary statistics of the news measures used in this study are reported in Table IA.VIII of Internet Appendix.

B. Data and Sample

C.1. Raw Data

To investigate the information content of news, we focus on the ex-ante prediction of first-time corporate bond defaults. We select the Chinese market as our laboratory not only because the limited early-warning capacity of credit rating agencies presents a significant challenge, but also due to the complexity of issuer ownership structures. As discussed in Section II, state entities occupy a pivotal position in China's economy, creating a unique nature of corporate and media ownership. This institutional setting enriches our analysis by allowing us to examine the interplay between issuer ownership and media incentives, specifically, how this interaction shapes the flow and credibility of financial information. We collect all news headlines from May. 1, 2015 to Dec. 31, 2022 from the *Choice* database (i.e., *Eastmonty.com*), and matched with over 2.83 million firm-specific news headlines filtered from 9.04 million news headlines.

Our sample construction is driven by news data availability and the requirement for a minimum six-month forecasting window to evaluate predictive power. The final sample covers corporate bond issuers in the primary market from November 1, 2015, to December

31, 2022 (detailed in Panel B of Table IA.I in the Internet Appendix). We identify 235 first-time defaulting issuers, comprising 63 listed companies, 39 state-owned enterprises (SOEs), and 9 central SOEs, and match them with a control group of 3,294 non-defaulting issuers with outstanding bonds.

To validate our model, we split the full sample into training and test sets using December 31, 2019, as the cut-off. The training sample (from November 1, 2015, to December 31, 2019) includes 131 defaulting and 3,178 non-defaulting issuers. The testing sample (from January 1, 2020, to December 31, 2022) comprises 104 defaulting and 2,087 non-defaulting issuers.

C.2. Descriptive Statistics of Credit Risk-related News

We link news to specific enterprises by matching firm names, abbreviations, bond names, and the names of issuers' listed subsidiaries, resulting in a total of 2,834,237 firm-level news items in our news collection period. Our dataset demonstrates comprehensive coverage of Chinese corporate bond issuers, with an overall news coverage rate of 87.37%. We observe that media attention correlates with information availability; specifically, non-listed companies exhibit a significantly lower news coverage rate of 82.57%, compared to listed companies at 99.42%. Detailed descriptive statistics comparing news volume across different firm characteristics (e.g., ownership structure and listing status) are reported in Panel A of Table 1. Moreover, defaulting issuers attract extensive media attention, evidenced by an overall coverage rate of 97.02%. Crucially, this attention is not merely reactive; 93.19% of these issuers receive news coverage prior to their initial default. This high level of pre-event reporting indicates that media outlets serve as active information intermediaries rather than merely reporting on post-default outcomes.

Applying our domain-informed dictionary, we identified 459,267 credit risk-related articles, accounting for 16.2% of the total news volume. This corresponds to a broad coverage rate of 74.79% across all issuers (averaging 121 articles per issuer). We observe significant cross-sectional variation in media attention: coverage is nearly universal for listed firms (99.10%) relative to non-listed entities (65.58%), as reported in panel B of Table 1. Crucially, the dictionary captures information effectively for distressed firms, covering 91.91% of defaulting issuers overall and 80.43% prior to their initial default. Even

at a 12-month horizon before default, 46.38% of issuers receive credit risk-related coverage. These findings suggest that the news media function as an active information intermediary, providing ex-ante warning signals rather than merely reporting ex-post outcomes.

We present topic-level word clouds capturing news content over a 3-month prediction horizon in Figure 2. Notably, defaulting issuers exhibit non-zero news coverage across all risk topics prior to the first default event, confirming that our domain-informed dictionary effectively extracts representative early warning signals. Crucially, the dictionary uncovers predictive information about qualitative dimensions that are often absent from standard financial reports. For instance, news media actively disseminates critical details on development strategy, production and operational status, shareholding pledges and reductions, market risk, and asset disposals and financing capacity, as evidenced by an average of more than 10 risk-related articles per defaulted issuer prior to the initial default. Collectively, these results highlight the dictionary's value in revealing the media's role in processing pivotal, non-financial information essential for assessing the credit risk landscape.

[Table 1**:** Descriptive statistics on the firm-level news and risk-related news]

[Figure 2. Word cloud figures for 21]

C.3. Correlation between News Metrics and Credit Risk Indicators

To evaluate the information content of our news metrics (*NTITF*), we examine their correlation with a comprehensive set of credit risk benchmarks. First, we construct three dummy variables to capture signals from credit rating agencies: *NonInvest* indicates non-investment-grade status; *ISDowngrade* denotes rating downgrades; and *FallenAngel* flags issuers transitioning from investment to speculative grade. Second, our comparison set includes Altman (1968)'s Z-score value (*ZS*) and Ohlson (1980)'s O-score value (*OS*) that calculated directly using the function reported in their original specifications[5]. In addition, we examine the underlying components of these models, along with 16 other financial ratios and accounting measures widely used in the literature. Finally, to control for

[5] The ZS and OS are specified as follows: $ZS = 1.2WCOA + 1.4REOA + 3.3EBITOA + 0.6EOD + 1.0TATR$. $OS = -1.32 - 0.407RSIZE + 6.03DR - 1.43WCOA + 0.0757CLCA - 2.37OENEG - 1.83ROA + 0.285NITA - 1.72INTWO - 0.521CHIN$.

macroeconomic conditions, we include the year-on-year growth rates of three key economic indicators: the Producer Price Index (PPI), fixed-asset investment, and aggregate financing to the real economy (following Da et al., 2015; Cong et al., 2019).

To facilitate correlation analysis, we employ a matching strategy that pairs each defaulting issuer with all available non-defaulting issuers at the time of the initial default[6]. This approach ensures that the control group is subject to the same macroeconomic and industry fundamentals as the distressed firms. This procedure yields a training sample of 43,505 firm-day observations, comprising 131 defaulting issuers and 3,178 non-defaulting counterparts.

We conduct our correlation analysis at the time t^*, defined as h months prior to the default event (i.e., $t^* = T_D - h$). To strictly avoid look-ahead bias, the predictor set is constructed using only information publicly available as of t^*. We carefully account for reporting lags, which vary by document type[7]. For example, audited annual reports typically exhibit longer delays than interim filings and are often released in late April, coincident with the first-quarter reports of the subsequent year. Given the superior reliability of audited annual data, we utilize the prior year's annual report for predictions generated between May and August[8]. We apply similar adjustments to macroeconomic variables, accounting for the typical release lag of approximately 20 days following the quarter's end[9].

Table 2 reports the Pearson correlation coefficients between our news metric (*NTITF*), credit risk indicators (*ZS, OS, NonIvest, ISDowngrade, FallenAngel*), and macroeconomic control variables. Correlations between *NTITF* and specific financial ratios are detailed in Table IA.IX of Internet Appendix. In the training sample, *NTITF* exhibits negligible correlations with third-party credit ratings ($|\rho| < 0.04$). Similarly, correlations between *NTITF* and accounting-based credit scores (*ZS* and *OS*) remain low, despite the high

---

[6] In cases where a non-defaulting issuer has multiple outstanding bonds, we select a single representative bond, as our analytical focus is at the issuer level.

[7] Following standard disclosure regulations, we assume the following availability schedule: third-quarter reports from the prior year are available from January to April; prior-year annual reports and current-year first-quarter reports are available from May to August; semi-annual reports are available from September to October; and third-quarter reports are available from November to December.

[8] Alternatively, using the first-quarter report of the current year yields qualitatively similar results.

[9] Based on release schedules, we map macroeconomic data as follows: January incorporates third-quarter data from the prior year; February to April incorporates fourth-quarter data from the prior year; May to July incorporates first-quarter data of the current year; August to October incorporates second-quarter data; and November to December incorporates third-quarter data.

correlation observed between ZS and OS themselves. Notably, rating downgrades (*ISDowngrade*) exhibit a more pronounced association with *ZS* (magnitude > 0.1). In terms of financial fundamentals, *NTITF* displays minimal correlations with most financial ratios. In contrast, third-party ratings and downgrades exhibit significantly stronger correlations with financial ratios (magnitudes > 0.1). Furthermore, *NTITF* reflects shifts in economic conditions; specifically, its correlation with fixed-asset investment (*FAI*) exceeds 0.1, whereas traditional credit scores and external ratings exhibit very weak correlation with macroeconomic variables. Collectively, these findings suggest that *NTITF* captures idiosyncratic, soft information that is largely orthogonal to traditional credit metrics.

[Table 2: Statistics of news metric and its correlations with credit risk and economic measures]

V. Incremental Value of Domain-informed News Extraction

A. Incremental Explanatory Power of Domain-informed News Metric

A.1. Conditional Logistic Regression (LR) Model

To assess the incremental predictive power of our domain topic-driven news metric, we adopt static prediction models to explore its incremental explanatory and predictive power. For incremental explanatory power, we adopt a conditional logistic regression (LR) model to predict the corporate bond issuers in default for the first time, as specified in Eq.(2) following Campbell et al. (2008), which is controlled by available information when forecasts are made with a *h* month forecasting horizon.

$$P\left(Y_{i,T_D}=1\middle|\Omega_{i,t^*}\right)=\frac{1}{1+e^{-\left(\beta_0+\beta_1 NTITF_{i,t^*}+\sum_{j=1}^{p}\gamma_j Controls_{i,j,t^*}\right)}}, \quad (2)$$

where Y_{i,T_D} is an indicator that equals one if the issuer goes into default for the first time on the defaulted date T_D, $x_{i,t}$ is a vector of explanatory variables from the available information set Ω_{i,t^*} known at the prediction time t^* ($t^*=T_D-h$). To evaluate the incremental information value of *NTITF*, we choose the forecasting horizon of 3 months as the benchmark analysis, due to the information disclosure frequency requirement for listed companies in China is quarterly, i.e., *h*=90 days. $Controls_{i,j,t^*}$ is the control

variables used to evaluate the incremental information provided by our news measure beyond financial ratios released by firms and macroeconomic indicators. We consider the financial ratios used in two classical models (Altman's Z-score model and Ohlson's O-score model), 16 financial ratios and accounting terms mostly used in prior studies, and three economic indicators, as listed in Table A. I of the Appendix.

To address the incremental information of news topic measures beyond financial ratios released by firms and macroeconomic indicators, we firstly compare the models with and without news variables $NTITF_{it}$. When beta coefficient of $NTITF_{it}$ is significant and Pseudo R^2 of conditional logistic regression (LR) model are improved with news variables $NTITF_{it}$ compared with that of LR model without $NTITF_{it}$, news variable contain incremental information. Then, we examine the out-of-sample forecasting performance of the model with $NTITF_{it}$ compared to that of model without $NTITF_{it}$. When the forecasting performance especially recall measure is improved, we can confirm the usefulness of news and our proposed credit risk dictionary.

To address the imbalance-sample problem in prediction model construction, this study adopts a random-matching approach with a fixed ratio, a downsampling strategy commonly used in bankruptcy and default prediction, financial fraud, and so on. A 1:3 matching ratio for defaulted and non-defaulted issuers is adopted, following Campbell et al. (2008). The random matching rules are as follows: for the first time, based on the default timestamp associated with the initial defaulted corporate bond of the defaulted issuer, a specific number of non-defaulted issuers with the same type of outstanding corporate bonds are randomly selected around that timestamp. This process is implemented through firm-level non-replacement sampling, which results in defaulted vs. non-defaulted issuers: 131:393 in the training sample, 104:312 in the testing sample.

For robustness, we employ an alternative matching strategy that pairs each defaulting issuer with the universe of available non-defaulting issuers at the time of the initial default. We define non-defaulters as firms with outstanding bonds that subsequently matured without default. Given our focus on issuer-level risk, we select a single representative bond for firms with multiple issues. This approach ensures that the control group shares identical macroeconomic and industry fundamentals with the distressed firms. The resulting training sample comprises 43,505 firm-day observations, matching 131 defaulting issuers with

3,178 non-defaulting counterparts.

A.2. Results of Incremental Explanatory Power of NTITF

To examine the incremental information value of news measure, we examine whether the Pseudo R^2 of the conditional LR model can be improved by incorporating news measure *NTITF* into the LR model with only control variables from three different models: Z-score, O-score and our benchmark model recorded as LRSFM, the last of which is obtained using bidirectional stepwise regression from a pool of 16 financial ratios and three year-on-year growth rates of quarterly economic indicators usually used in previous studies.

The estimation results of the LR model for a 3-month prediction horizon under the randomly matching sample construction method estimated using the training sample are reported in Panel A of Table 3.[10] The empirical results confirm the incremental information value of credit risk-related news measures beyond financial ratios and macroeconomic variables. Specifically, the news-based measure *NTITF* is statistically significant at 1% level. The conditional LR model incorporating the news measure solely attains a Pseudo R2 of 9.06%, while the inclusion of the news measure yields an improvement in Pseudo R2 ranging from 7.25% to 8.33% across three benchmark models relative to their counterparts excluding the news variable.

According to the results shown in Panel B of Table 3, even based on the sample construction method using all available matches, the news measure NTITF improves Pseudo R^2 by about 3.79% to 4.15% across benchmark models. The incremental value of the news measure is still robust[11]. These results indicate the marginal utility of news analytics and affirm the efficacy of the proposed credit risk dictionary.

To further examine the early-warning ability of the topic-driven news metric, we re-estimate the conditional logistic regression model across multiple forecasting horizons: 4, 6, 9, and 12 months. The results in Panel C of Table 3 confirm the incremental predictive value of the news measure *NTITF*. The news metric *NTITF* consistently enhances the

---

[10] To save space, we only provide the coefficient results of *NTITF* of the conditional LR model in Table 3. The full estimation results are presented in Table IA.X of Internet Appendix E.

[11] We also examine the robustness from several aspects: an alternative training sample period with a new cut-off point on December 31, 2018, a sub-sample of good financial quality issuers that have no missing data for all 16 financial ratios selected in this study, or more control variables of 93 financial indicators. More results are available in Internet Appendix F.

goodness-of-fit of the conditional LR model, although the improvement in Pseudo R^2 diminishes as the horizon extends from 4 to 12 months.

[Table 3: Incremental explanatory power of *NTITF* across horizons]

A.3. Robustness Checks of Alternative News Measures

To ensure the robustness of our results, we constructed other alternative news measures using our domain-informed credit risk dictionary. Besides the inverse topic frequency weighting scheme, another weighting scheme, "the topic score," is adopted, which is the degree of inverse topic frequency. Based on the ranking of inverse topic frequency, $\ln\left(\frac{N}{N_k}\right)$, topics in the top tercile were assigned a score of 3, those in the second tercile a score of 2, and those in the bottom tercile a score of 1. Thus, the measure of topics counts with topic- score weighting scheme ($NTTDW_{it}$) is defined as Eq.(3).

$$NTTDW_{it} = \sum_{j_t \in D(t)} \sum_k TMentioned_{ijk} \cdot Score_k. \quad (3)$$

We also examine a straightforward measure of news topics, denoted as topic counts (NT_{it}), which is defined as the number of distinct topics to which risk-related words or phrases belong for each issuer, based on a historical corpus of news headlines in $D(t)$, as specified in Eq.(4), is also examined.

$$NT_{it} = \sum_{j_t \in D(t)} NT_{ij}, \quad (4)$$

where NT_{ij} is the number of topics mentioned by news j in its headline for issuer i by searching the words/phrases of the credit risk dictionary.

Similarly, we also construct and examine the credit risk-related words count measure with/without a weighting scheme. The word count variable is commonly used in previous literature, e.g., Tetlock (2007). We adopt a credit risk-related words count measure (NW_{it}), as well as the variables of credit risk words count under *ITF* weighting ($NWITF_{it}$), and that under topic score weighting ($NWTDW_{it}$), as shown in Eq.(5) to Eq.(7), respectively.

$$NW_{it} = \sum_{j_t \in D(t)} NW_{ij}, \quad (5)$$

$$NWITF_{it} = \sum_{j_t \in D(t)} \sum_k NW_{ijk} \cdot \ln\left(\frac{N}{N_k}\right), \quad (6)$$

$$NWTDW_{it} = \sum_{j_t \in D(t)} \sum_k NW_{ijk} \cdot Score_k, \quad (7)$$

where NW_{ij} represents the number of risk words/phrases in our credit risk dictionary mentioned by news headline j of issuer i released on a day within a historical period of

$D(t)$ until day *t*.

Besides, we also adopt the variable of credit risk-related newspaper counts (NN_{it}), i.e., the number of news that contains risk words of our credit risk dictionary in the news headlines for issuer *i* until day *t*, as shown in Eq.(8).

$$NN_{it} = \sum_{j_t \in D(t)} IsNeg_{ij}, \tag{8}$$

where $IsNeg_{ij}$ is a dummy variable that equals 1 if news headline *j* for issuer *i* contain the risk words/phrases for issuer *i*'s news in a historical period $D(t)$ until day *t*. The descriptive statistics of these alternative news measures are reported in Table IA.VIII of Internet Appendix.

The results of the robustness check using alternative new measures under the LR model with control variables from different models are reported in Table IA.XI of Section F in the Internet Appendix. Regardless of the specific measure employed, the incremental information content of news remains robust: the coefficients on the news variables are consistently positive and statistically significant, and they increase Pseudo R2 beyond that explained by the control variables of financial ratios and macroeconomic variables. However, the magnitude of this incremental power varies. Notably, among the alternative candidates, the word-based metric (*NWITF*) yields the largest improvement in Pseudo R^2 (rising from approximately 7.41% to 8.41%), although this improvement is slightly lower than that achieved by our primary metric *NTITF*.

We report a battery of additional robustness checks in Section F of the Internet Appendix (Tables IA.XII-IA.XIV). These tests consider: (1) an alternative training window ending on December 31, 2018; (2) a subsample restricted to issuers with complete data for all 16 financial ratios; and (3) an expanded model controlling for 93 financial indicators. Across all specifications, the results consistently confirm the incremental information value of our domain-informed news metric (*NTITF*).

B. Incremental Forecasting Power of Domain-informed News Metric

To carefully evaluate the forecasting performance of the news variables, we compare the out-of-sample forecasting performance under a 3-month forecasting horizon based on the testing sample. This study examines the predictive power of the risk topic-related news variable *NTITF* using both conditional logistic regression and multiple machine learning

models, specifically XGBoost, support vector machines (SVM), and decision trees (DT), to determine its sensitivity to model choice. We adopt several evaluation measures to assess the performance of default prediction models, including recall, F_1 score, precision, accuracy, AUC. For corporate bond investors, the recall measure is more important than other measures for identifying risks, compared with the most commonly used measures of accuracy and AUC. F_1 score is the harmonic mean of precision and recall. Hence, we focus more on the performance evaluation measures of recall and F_1 score for comparison.

The parameters and hyperparameters of machine learning models are trained and selected using five-fold cross-validation tests based on the training sample. For example, the optimal probability threshold hyperparameter for the conditional logistic regression model is selected by maximizing recall while maintaining precision above 0.4 in five-fold cross-validation. This value is determined via a grid search over thresholds from 0.3 to 0.7, in increments of 0.02. Also, we estimate XGBoost, SVM, and DT models and select optimal models based on out-of-sample recall under five-fold cross-validation. The selection criterion targets the highest average recall while ensuring precision remains above 0.4 across all folds. For each algorithm, we explore a range of hyperparameters to identify the best-performing specification. More details on the approach to hyperparameter selection can be found in Internet Appendix G.

The out-of-sample 3-month horizon forecasting performance of several machine learning models based on the testing sample is reported in Table 4. The results indicate that incorporating the news measure *NTITF* consistently improves predictive accuracy across all benchmark models. Specifically, including NTITF leads to a substantial increase in recall, with improvements ranging from 0.096 to 0.337, resulting in final recall values of 0.654-0.808. A similar improvement can be observed in the F1 score. Although the XGBoost model delivers the strongest overall performance, the predictive power of *NTITF* remains robust across different models, demonstrating that its effectiveness is not contingent upon the choice of model.

We then evaluate the out-of-sample forecasting performance of these conditional logistic regression (LR) models over long-term horizons using a testing sample. To assess the early-warning capability of the news measure, we also test multiple machine learning models—XGBoost, SVM, and DT. As summarized in Table 5, both recall and F_1 score are

improved consistently by incorporating *NTITF* across forecasting horizons. The predictive performance at 6-month horizons is notably strong, achieving a recall of 0.798 under both XGBoost and SVM models for a 4-month forecasting horizon, 0.75 under DT, and 0.74 under LR for a 6-month forecasting horizon. Even at the 12-month horizon, recall remains substantial at 0.625 under XGBoost model. These results confirm the robust early-warning ability of the news measure, which is invariant to model selection. Thus, these results confirm the early-warning ability of news for predicting defaults among corporate bond issuers.

[Table 4: Out-of-sample 3-month horizon forecasting performance of news metric]

[Table 5: Out-of-sample long-term prediction performance of news metric]

VI. Comparison of Expert-driven versus Data-driven Methods

A. Comparison to Topics from BERTopic Models

To evaluate the efficacy of our framework, which introduces domain-specific knowledge into textual analysis to construct a specialized dictionary based on specific topics, we compare the performance of expert-driven topics derived from domain-specific expertise with that of data-driven topics constructed without incorporating experts' logic and domain-specific knowledge. For data-driven topic analysis, we employ the BERTopic model (Grootendorst, 2022), which leverages clustering techniques and a class-based variant of TF-IDF to generate coherent topic representations from a pre-trained language model (e.g., BERT; Devlin et al., 2019) to obtain document-level information. The architectural details and procedural flow of our BERTopic model implementation are delineated in Figure 3.

[Figure 3. BERTopic architecture for identifying topics of news]

The use of pre-trained embeddings of BERTopic model facilitates the capture of generalized linguistic features while mitigating variance in data-scarce environments through transfer learning. Recent applications in financial text analysis have demonstrated the efficacy of BERTopic. In this study, we implement a BERTopic framework initialized with word embedding from a pre-trained LLMs such as BERT embedding of Devlin et al. (2019) and FinBERT embedding of Araci (2019). The baseline training corpus of BERT

model is generic (not finance-specific) [12], while FinBERT is a language model based on BERT to tackle financial sentiment analysis using financial domain datasets[13].

In this study, we train a BERTopic model using the set of news headlines of the first-time defaulted issuers in our training sample, which comprises news articles covering 131 defaulted issuers. The optimal number of topics, *K*, is selected from a candidate set ranging from 50 to 150 in increments of 25, by maximizing the topic coherence score following the method of Röder et al. (2015). Each topic's document embedding and each topic's word representation based on *c-TF-IDF* in Grootendorst (2022)[14] can be calculated.

Then we can identify news topics using the trained BERTopic model. For each news headline, we calculate its document embedding using the word embedding of a pre-trained model. The topic probability is then derived based on the cosine similarity between the document embedding of this news and the topic embedding of each topic obtained by the BERTopic model. This probability of news j belonging to topic k is formally defined as $TopicProb_{jk} = \frac{ENews_j \cdot ETopic_k}{\|ENews_j\| \cdot \|ETopic_k\|}$, where $ENews_j$ is document embedding of news j, $ETopic_k$ is the topic embedding of topic k. Based on the topic probability ($TopicProb_{jk}$) that a news headline belongs to various topics, the topic with the highest probability is assigned as the article's primary topic. Using historical news for each issuer over a specific period, we then calculate the topic-level number of topics covered by news (NT) using data-driven primary topic identification. This enables a comparison of the effectiveness of expert-driven versus data-driven topics using topic-level *NT* as a common metric.

To compare the predictive power of expert-driven topics versus data-driven topics

---

[12] BERT is designed to pretrain deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. The code and pre-trained BERT models of Devlin et al. (2019) are available at https://github.com/google-research/bert. This study adopts the "BERT-Base, Chinese" model. The English version of BERT models was pre-trained on BookCorpus and English Wikipedia, while there is not an official statement from the Google BERT release that shows the exact Chinese training corpus. The Chinese version of BERT models was pre-trained possibly based on Chinese Wikipedia only, or also other corpora such as news, Common Crawl, etc. The baseline training corpus is generic (not finance-specific).

[13] Araci (2019) further pre-train a BERT language model on a financial corpus. FinBERT is developed by ProsusAI company, which can be downloaded through https://github.com/ProsusAI/finBERT.

[14] A class-based variation of *TF-IDF*, c-TF-IDF$_{x,k} = tf_{x,k} \times \log\left(1 + \frac{A}{f_x}\right)$, where $tf_{x,k}$ is the frequency of word x in topic k, f_x is the frequency of word x in all topics and A is the average number of words per topic. In general, the representative terms of each topic are identified by ranking the words according to their *c-TF-IDF* scores.

extracted by BERTopic models for first-time default issuers, we adopt the LASSO model, following Chen, Kelly and Xiu (2022), to select powerful topics or the same number of topics as our credit risk dictionary used. The conventional LASSO and size-21-constrained LASSO are adopted for topic selection, based on the topic-level variable of each issuer, *NT*, based on topics mentioned in the news of bond issuer i at the prediction date t^*. The penalty parameter of LASSO is selected using a five-fold cross-validation procedure, with the value chosen to maximize recall. Based on the training sample constructed by a 1:3 randomly matching method, we estimate the LASSO model for a 3-month prediction horizon and report the coefficients of selected topics and their p-value estimated according to the post-selection inference procedure proposed by Tibshirani et al. (2016) in Table 6.

For topic selection by conventional LASSO[15], four topics from topics clustered by FinBERT embedding-based BERTopic model can provide a Pseudo R^2 of 0.0759, while three topics from topics clustered by BERT embedding-based BERTopic model can provide a Pseudo R^2 of 0.0436, respectively. For the size-21-constrained LASSO models, 21 topics from the topics clustered by FinBERT embedding-based BERTopic model can provide a Pseudo R^2 of 0.1950, while 21 topics from topics clustered by BERT embedding-based BERTopic model can provide a Pseudo R^2 of 0.1601, respectively. On the other hand, 6 expert-driven topic-level NT variables selected by conventional LASSO achieve a Pseudo R2 of 0.1347, while the 21 expert-driven topics achieve a Pseudo R^2 of 0.2008. From the goodness-of-fit, our expert-driven topics provide more information.

We evaluate the out-of-sample performance of default prediction models that incorporate topic-level variables *NT*, either expert-driven or data-driven topics derived from BERTopic, into algorithm-specific benchmark prediction models (Table 7). We find two main results. First, expert-driven topics outperform data-driven topics in forecasting accuracy (e.g., recall and F1 score), both when using the full set of 21 topics and when focusing on a selected subset. Second, expert-driven topics provide incremental predictive power beyond financial ratios and macroeconomic indicators, and this improvement is robust across model specifications. By contrast, the incremental value of data-driven topics relative to financial and macro variables is less stable—although generally positive, their

[15] The topic representation of topics clustered by different BERTopic models and selected by conventional LASSO can be found in Figure IA.II and Figure IA.III of Internet Appendix.

predictive gains disappear in certain models. Specifically, the refined set of data-driven topics outperforms the full set of 21 topics, yielding stronger and more stable incremental predictive power. The BERTopic model, based on BERT embeddings trained on larger general corpora, exhibits more stable incremental predictive power than the model using FinBERT embeddings that was trained specifically for sentiment analysis tasks. Our findings show that embedding domain-specific expertise into text analysis yields greater value for specialized default-prediction tasks than relying on either general linguistic knowledge or task-specific knowledge from other domains.

[Table 6: Predictive power of topics extracted from domain-informed dictionary and data-driven BERTopic models]

[Table 7: Out-of-sample prediction performance of model with data- and expert-driven topics]

B. Comparison to ChatGPT

To further assess the value of domain-specific knowledge, we compare our approach with ChatGPT-4.1, a large language model trained on broader, though not necessarily more specialized, corpora. For this comparison, we employ the issuer-level measure (NN_{it}), the number of credit risk–related news headlines associated with each issuer at prediction time using a 3-month historical data period, to conduct a direct performance evaluation.

With an instruction-based prompting approach, we employ ChatGPT-4.1 to assess whether a news headline contains predictive signals relevant to default prediction. The specific prompt is as follows: "You are a professional credit risk analyst. Please determine whether each news headline contains information that can be used for positively predicting corporate credit risk events. If it does, label it as 1; otherwise, label it as 0. Please output the credit label for each news headline." Using all news headlines that have been anonymized through the concealment of company names, stock tickers, bond codes, and other identifiers, we obtain the assessment of news information and obtain the measure of *NN* calculated by ChatGPT.

We provide the out-of-sample default prediction performance of models with *NN* as well as control variables for a 3-month forecasting horizon in Panel A of Table 8. The

results show that the prediction model based on our domain-specific dictionary–constructed news variable (*NN*) outperforms the model using ChatGPT-generated *NN*. Moreover, the domain-specific *NN* exhibits stable incremental predictive power, consistently improving recall across all forecasting models. In contrast, the ChatGPT-based *NN* delivers incremental gains in most models, but its predictive power deteriorates and becomes insignificant under the XGBoost specification.

[Table 8: Out-of-sample performance of models with *NN* constructed by different methods]

C. Comparison to other General or Financial Domain Dictionaries

Besides, we compare the performance of extracting useful information in news by our special-purpose dictionary with embedded domain-specific expertise versus that of a general dictionary with larger words, as well as financial domain knowledge for other tasks. We select one general dictionary often used in the Chinese text analysis, developed by HowNet company[16], which contains 1,254 negative words out of 2,090 total words. For a dictionary with financial domain knowledge, we employ a financial media tone dictionary developed by Jiang et al. (2021), which is based on Chinese news and contains 5890 negative words out of 9228 total words. We only use negative words from the dictionary to compare with our credit risk dictionary.

To examine the impact of different dictionaries on information extraction, we compare the out-of-sample forecasting performance of the news measure NN constructed with each dictionary. The results are reported in Panel B of Table 8. The comparison results show that our domain-specific credit risk dictionary outperforms not only a general-purpose dictionary with a much larger vocabulary but also a media tone dictionary developed from Chinese news. Moreover, the *NN* constructed from the general dictionary exhibits lower stability than that based on our specialized dictionary, because the *NN* derived from the general dictionary loses predictive significance under XGBoost.

Overall, the three sets of comparisons demonstrate that news variables constructed

[16] Source of OpenHowNet: https://openhownet.thunlp.org/.

based on our domain-specific credit risk dictionary consistently deliver superior and more stable predictive performance than alternative approaches, including BERTopic models based on BERT or FinBERT embeddings, ChatGPT-4.1, and other dictionaries for general-purpose or other tasks. These findings underscore that the depth of knowledge outweighs breadth in tackling specialized problems. The results highlight the substantial value of embedding human domain expertise into text analysis. The depth and specialization of knowledge prove more valuable than its breadth or the transfer of task-specific knowledge from other domains. The proposed domain-knowledge embedding approach offers a cost-efficient solution: a specialized dictionary containing only 323 curated terms and phrases not only outperforms large general-purpose dictionaries with thousands of entries but also surpasses state-of-the-art large language models such as ChatGPT and BERTopic-based models.

VII. Source and Incremental Content of Media News

A. The Different Roles of State-controlled and Market-oriented Media

A.1. Hypotheses about the Role of Different Types of Media

The media play a central role in capital markets as both information intermediaries and external monitors. While a growing literature documents the informational and governance functions of media, existing empirical evidence largely focuses on market-oriented outlets operating under private ownership (e.g., You et al., 2018). Much less is known about whether, and how, state-controlled media differ from market-oriented media in transmitting economically relevant information—particularly negative information related to corporate credit risk. This gap is especially salient in the context of early warning signals for first-time bond defaults, where traditional financial indicators often fail to provide timely information.

We examine this issue using the Chinese market, where state-controlled and market-oriented media coexist and jointly shape the information environment. China offers a uniquely rich setting: state ownership remains prevalent in media organizations, while market-oriented outlets have expanded rapidly; at the same time, corporate and enterprise bond issuers exhibit substantial heterogeneity in ownership structures, including central

state-owned enterprises (CSOEs), local state-owned enterprises (SOEs), and non-state-owned enterprises. Our empirical analysis draws on a comprehensive dataset covering all media outlets and corporate and enterprise bond issuers in China from 2015 to 2022, including both listed and unlisted firms. In our sample of 3,928 media sources, 38 are state-controlled. State-controlled outlets account for a nontrivial share of influential coverage: seven appear among the top 20 outlets by total news volume (Table C1 of Internet Appendix), and five rank among the top 20 by credit-risk-related coverage prior to issuers' first defaults (Table C2 of Internet Appendix). These features make China an ideal laboratory for studying the roles of media in credit risk disclosure and monitoring.

Our analysis builds on and extends the framework of You et al. (2018), who propose two non-mutually exclusive hypotheses. Under the *informative state hypothesis*, state-controlled media are viewed as more credible and less prone to sensationalism, owing to weaker commercial incentives and government oversight. Under the *informative market hypothesis*, market-oriented media are expected to outperform state-controlled media in information production efficiency and independence, driven by incentive-based compensation and promotion mechanisms that encourage information acquisition through diverse channels. Prior work also suggests that self-censorship may impair the reporting quality of state-controlled media, particularly for politically sensitive topics (Houston et al., 2011; You et al., 2018). We revisit these hypotheses in the context of credit risk news, which represents economically meaningful yet potentially sensitive negative information for issuers.

Beyond these perspectives, we propose a media governance hypothesis that highlights the trade-off between informational access and strategic selectivity. We argue that state-controlled media possess a unique monitoring advantage derived from privileged access to government channels—an advantage that remains robust even in weak disclosure environments. However, this governance role is likely exercised selectively: for issuers of national strategic importance, state outlets may suppress negative signals to prioritize social stability over transparency. In contrast, market-oriented media function as independent information substitutes. Their monitoring effectiveness is most pronounced in opaque settings characterized by high information acquisition costs, yet they remain relatively unconstrained by the political incentives that drive selective disclosure.

A.2. The Roles of Different Types of Media under Different Information Environments

To examine the informational and corporate governance roles of the two media types, we first compare the predictive power of news published by each type across differing information disclosure environments. This approach helps illuminate the distinct information channels characteristic of each media type. In China, listed companies are subject to stricter and more frequent disclosure requirements than unlisted firms, as stipulated by relevant regulations. Listed firms must regularly release comprehensive financial and operational details, including annual, interim, and quarterly reports—the latter mandated by Chinese stock exchanges. In contrast, enterprises that issue bonds are primarily required to disclose annual and interim reports, with an emphasis on bond-related financial information and the use of the capital raised. This divergence in disclosure intensity between listed and non-listed bond issuers provides a natural setting for analyzing how media types function under varying information conditions.

There are two kinds of views related to the information environment. From the view of information channels, two types of media become more apparent in an information-sparse environment characterized by less information disclosed by companies. Because state-controlled media has unique access to official information about companies, its predictive power should not be greatly affected by the broader information environment, and it may provide more accurate information due to the credibility of its information source. Market-oriented media may offer diverse channels of information, but the accuracy of that information can be affected by the broader information environment. On the other hand, incentive-based compensation and promotion systems enable market-oriented media to demonstrate greater initiative in providing higher-quality information. In an information-sparse environment, when media, especially market-oriented media, face high information acquisition costs, the difference in motivation between the two types of media becomes more pronounced. Market-oriented media, which typically operate under high-powered, performance-lined incentive schemes, show a stronger tendency to actively seek out and collect information. In an information-rich environment where information is readily available, differences in incentive structures across media types are most likely to manifest in analytical capabilities: the higher the incentives, the stronger the analytical capabilities exhibited by the media.

Following the conditional logistic regression framework for a 3-month forecasting horizon used is this study, we conduct an empirical analysis using a new media-level news measure, credit risk information ratio, separately computed for state-controlled and market-oriented media (SCR_{it} and MCR_{it})[17]. This measure allows us to control for differences in the number of units/outlets across media types. Table 9 presents the estimation results ofor the media-level credit risk information ratio measures (SCR and MCR) based on a subsample of listed and non-listed issuers for the fullsample period from Nov. 11, 2015, to Dec. 31, 2022. The sample includes both defaulted and non-defaulted issuers, constructed using the matching method that matches all available non-defaulted issuers to defaulted issuers to leverage more data for building an explanatory model. The results show that in an information-rich environment, both state-controlled and market-oriented media individually contribute incremental information value to credit risk prediction. However, when compared, state-controlled media exhibit a more significant incremental information value for listed firms' issuers. Conversely, under an information-sparse environment with less disclosed information for non-listed firms' issuers, both media-level news measure of credit risk information ratio (SCR_{it} and MCR_{it}) are significantly positive, irrespective of whether they are included separately or jointly in the model, with coefficients of comparable magnitude.

Our findings are consistent with both the informative state hypothesis and the informative market hypothesis, and provide supportive evidence for our media governance hypothesis. We document that state-controlled media exert monitoring effects that shape the information environment, consistent with their access to stable and authoritative information channels. We further show that state-controlled and market-oriented media rely on distinct information channels, with these differences becoming particularly pronounced in environments characterized by limited disclosure. In addition, our results lend support to a quality-oriented incentive hypothesis for market-oriented media: when

---

[17] $SCR_{it} = SNTITF_{it}/SN_{it}$ and $MCR_{it} = MNTITF_{it}/MN_{it}$, where $SNTITF_{it}$ or $MNTITF_{it}$ are topic related news measure calculated separately for state-controlled and market-oriented media using their published news respectively. Similar to Eq.(1), $SNTITF_{it} = \sum_k \sum_{j_t \in D(t)\ \&\ j_{mediatype}=S} TMentioned_{ijk} \cdot \ln\left(\frac{N}{N_k}\right)$, $MNTITF_{it} = \sum_k \sum_{j_t \in D(t)\ \&\ j_{mediatype}=M} TMentioned_{ijk} \cdot \ln\left(\frac{N}{N_k}\right)$, where SN_{it} or MN_{it} represent total number of news articles about enterprise *i* published by state-controlled media or market-oriented media during period *D*(*t*). *D*(*t*) is 3-month period in this study.

disclosure is limited and information acquisition costs are high, market-oriented outlets provide incremental and qualitatively different information relative to state-controlled media. However, we find no evidence that market-oriented media exhibit superior analytical capacity in information-rich environments.

[Table 9: Predictive power of media-level news under different information density]

A.3. The Impact of Issuers' Ownership Nature on the Roles of Media

Next, we further explore the media governance hypothesis by using the unique corporate ownership structure inherent in China. The Chinese central state-owned enterprises (CSOEs) are typically large-scale entities with nationwide and international operations and play a critical role in national economic and social development. State-owned enterprises (SOEs) share similarities with CSOEs but have somewhat lower national strategic significance. Both CSOEs and SOEs occupy an important position in China's economy and serve as key drivers of national economic growth. We argue that differences in national importance among enterprises may lead to differential media behavior, particularly among state-controlled media that practice self-censorship. Consequently, issuers' ownership type is likely to influence the news media's reporting strategies. Using the LR framework and employing media-level news measure of credit risk information ratio, we analyze the differences in predictive performance across media types and issuer ownership categories. The results are reported in Table 10.

Our results uncover that the issuers' ownership nature has a significant impact on media's performance, and the degree of influence on state-controlled media and market-oriented media differs. For non-SOEs issuers, both state-controlled and market-oriented media provide incremental information for predicting first-time defaults, with state-controlled media playing a more pronounced role. For SOEs but non-CSOEs issuers, although SCR_{it} and MCR_{it} are significant at 5% significant level when entered separately into the model with control variables, only MCR_{it} remains significant when they are jointly entered into the model with control variables. For CSOEs issuers, neither SCR_{it} nor MCR_{it} is significant at 5% significant level when jointly entered into the model with control variables of our benchmark model, although MCR_{it} is only significant at 10% significant level when controlled by variables from Z-score or O-score models. Our

findings suggest that the greater an enterprise's national importance, the more cautious media outlets become in reporting negative credit-risk–related information, and in some cases may refrain from such reporting altogether to avoid adverse effects on strategically important firms. This pattern is particularly pronounced for state-controlled media, which exhibit heightened restraint in covering credit risk associated with central state-owned enterprises (CSOEs) and state-owned enterprises (SOEs).

These results provide direct support for the media governance hypothesis by documenting strategically selective disclosure of credit-risk information, especially by state-controlled outlets. The evidence is consistent with state-controlled media being guided by broader policy objectives—such as preserving social stability, supporting economic development, and aligning public narratives with state priorities—rather than solely by informational considerations. By contrast, our findings indicate that market-oriented media play a comparatively stronger role in external corporate governance by supplying critical information on firm-level credit risk.

[Table 10: Predictive performance of different media across issuers' ownership nature]

B. Topic-level Analysis of Media News

In this section, we employ topic-level analysis to examine the quality-incentive effects of market-oriented media and the strategic, selective reporting behaviors of both state-controlled and market-oriented outlets as articulated in the preceding hypotheses. We also analyze the types of information disclosed by different media outlets for firms with different ownership structures.

We further examine whether media coverage conveys incremental information beyond firms' observable financial fundamentals, rather than merely repackaging, extracting, or analyzing existing disclosures, or revealing information that has not yet been formally disclosed. We construct a comprehensive set of 93 financial indicators to proxy for issuers' financial information[18]. We then assess the informational overlap between

[18] We select 31 financial ratios from 66 candidate financial ratios from the quarterly financial reports by retaining only one financial ratio from a pair or a group with a Pearson correlation coefficient exceeding 0.4, to control for collinearity. A total of 93 indicators, comprising 31 selected financial ratios and their respective year-over-year and quarter-over-quarter changes, are employed as proxy variables for the financial information of issuers in this section.

media coverage and financial fundamentals by regressing the 21 topic-level news measures for state-controlled and market-oriented media ($SNTITF_{itk}$, $MNTITF_{itk}$) on these current, ongoing, and future financial indicators[19]. Specifically, these indicators are drawn from financial ratios reported in the most recently available quarterly report at time *t*, the subsequent quarterly report, and the second subsequent quarterly report. The extent to which news content can be explained by financial information can be measured by the R^2 of the regression model. Then, utilizing the portion of news content not explainable by financial information ($SNTITF_{itk}^{\perp}$, $MNTITF_{itk}^{\perp}$), measured by the residuals of the first step regression, this study examines its incremental informational value in predicting the first-time default of bond issuers.

Figure 4 report the R^2 of the first step regression for two types of media under the subsample of SOEs & non-SOEs bond issuers, as well as listed & non-listed firms. The extent to which news content can be explained by financial information is low, as all R2 values are below 9.53% for SOE and non-SOE bond issuers, and 10.23% for listed & non-listed firms. The extent of news content can be explained by the fact that financial information is much lower for SOEs and non-listed firms than for their counterparty. The state-controlled media's coverage of non-listed and SOE issuers has the lowest proportion of information explained by corporate financial information. These results imply the information channels of state-controlled media are not limited to corporate sources. In contrast, we find that market-oriented media are more influenced by the information environment than the ownership nature of issuers. This is evidenced by the results that the explanatory power of financial information is lowest for market-oriented media's news coverage of non-listed firms.

Moreover, the topics reported by different media types are distinct. Market-oriented media primarily serve to disseminate financial information disclosed by issuers, particularly non-SOEs and listed firms. This is evidenced by the high R^2 values observed in the coverage of corporate financial status, as well as market risk and distress by market-oriented media. Market-oriented media can even uncover future financial changes in firms through coverage of topics such as credit ratings, debt crises, production and operational

[19] Topic-level news measure in the regression is 99.95% winsorization of topic-level news from different media to avoid the effects of abnormal observations.

status, legal risks, violations of laws and regulations, regulatory inquiries, and financial fraud. Compared to market-oriented media, state-controlled media's coverage of corporate misconduct and risk-related topics, such as debt crises, shareholder misappropriation, violations of laws and rules, regulatory inquiries, and financial fraud, can predict future changes in firms' financial condition, especially non-SOEs.

[Figure 4. Extent of topic-level news explained by financial ratios for different firms]

Next, we examine the incremental information from orthogonal topic-level news across different media, removing the portion that can be explained by all available, ongoing, and future financial information. The estimation coefficients of the standardized orthogonal variables of the media's topic-level news under the LR framework are reported in Figure 5. We find significant differences in topic-efficacy across issue types between the two media. For listed firms and non-SOEs, most of the orthogonal variables in the media's topic-level news measures remain significant at the 5% level. For SOEs, state-controlled media provide more topic-level information than market-oriented media, especially on audit, risks, most aspects of corporate operations, most aspects of corporate development strategies, and regulatory inquiries. It seems that the information uncovered by market-oriented media complements that disseminated by state-controlled media for SOEs, focusing on the debt crisis, production status, strategic investors, personnel management, and deception. For unlisted firms, the difference in new information uncovered by state-controlled media and market-oriented media stems primarily from corporate development-related news. These results imply the differences in information channels of different media. These findings support our hypothesis that state-controlled media typically rely on relatively stable, official information sources, such as government departments and official institutions, and can obtain authoritative policy interpretations, official data, and other information. Our results highlight the unique position of state-controlled media relative to market-oriented media in China.

[Figure 5. Incremental value of orthogonalized topic-level news measure.]

VIII. Conclusion

This study evaluates a new domain-specific, expertise-informed approach to textual

analysis and demonstrates its effectiveness in forecasting first-time corporate bond defaults—a particularly challenging prediction task. We show that news headlines contain incremental, economically meaningful information about default risk. The proposed domain-informed credit-risk measure delivers strong early-warning signals and robust out-of-sample performance, with predictive accuracy extending to at least six months prior to default. These results highlight the value of embedding specialized financial expertise into textual analysis and establish news headlines as an informative yet underutilized source of forward-looking credit-risk indicators, especially for identifying first-time defaulters.

We further examine whether the method of extracting information from news matters. Specifically, we compare domain-specific expertise with broader but less targeted forms of knowledge in news analytics. At the topic level, we benchmark expert-driven topics derived from our credit-risk dictionary against data-driven BERTopic models built on BERT and FinBERT embeddings and selected via LASSO. At the firm level, we compare our approach with ChatGPT using a firm-level measure. Across both dimensions, our expert-driven method consistently outperforms LLM-based alternatives. Collectively, these findings demonstrate that while news headlines contain substantial predictive content for first-time defaults, extracting this information effectively requires domain-tailored structure. In specialized applications, depth of domain knowledge ultimately dominates breadth of general language understanding.

Furthermore, this study investigates the sources and nature of the incremental information embedded in media coverage by testing hypotheses on the roles of state-controlled and market-oriented outlets. Revisiting the debate between the informative-state and informative-market hypotheses (You et al., 2018), we introduce a governance hypothesis for state-controlled media, positing that their privileged access to governmental channels enhances monitoring but may lead to selective reporting for issuers of strategic importance. Empirically, we find that state-controlled media exhibit greater stability and stronger predictive power for first-time defaults, while market-oriented media provide complementary value, particularly when disclosure is constrained and information acquisition costs are high. These results provide further evidence consistent with the quality-oriented incentive hypothesis for market-oriented media. Also, we show that issuers' ownership structure significantly affects media's performance in predicting

defaults, with differing effects across state-controlled and market-oriented outlets. Specifically, the evidence supports the governance hypothesis for state-controlled media, reflecting their strategically selective reporting of credit risk information related to central state-owned enterprises (CSOEs) and state-owned enterprises (SOEs).

We also examine differences in the content of incremental information conveyed by these media outlets. By comparing topic-level news measures with issuers' financial indicators and analyzing the orthogonal components unexplained by current or forthcoming financial data, our findings confirm that news media do not merely disclose accessible or predictable corporate financial information; instead, they deliver a more diverse set of forward-looking insights, including perspectives on risk and distress, business operations, development strategies, and corporate ethics. For state-owned enterprises, state-controlled media deliver especially rich incremental topic-level information, consistent with their role as authoritative information channels.

An important avenue for future research lies in further advancing how domain-specific financial knowledge—particularly the fragmented, experience-based, and non-formalizable components that escape traditional modeling—can be systematically embedded into machine learning frameworks. This study offers an initial, transparent, and interpretable approach for incorporating such expert knowledge into textual analysis, but much remains to be explored. Developing more generalizable mechanisms for integrating economic reasoning with modern NLP architectures, understanding how and when embedded expertise improves predictive performance, and extending these methods to broader financial applications represent promising directions for continued investigation.

REFERENCE

Agarwal, S., Chen. V. Y., Zhang, W. (2016). The information value of credit rating action reports: A textual analysis. *Management Science*, 62(8), 2218-2240.

Altman, E. I. (1968). Financial ratios, discriminant analysis, and the prediction of corporate bankruptcy. *The Journal of Finance*, 23(4), 589-609.

Altman, E. I., Sabato, G., and Wilson, N. (2010). The value of non-financial information in SME risk management. *Journal of Credit Risk*, 6(2), 95-127.

Araci, D. (2019). Finbert: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021, March). On the dangers of stochastic parrots: Can language models be too big?. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency* (pp. 610-623).

Bie, S. Y., Feng, G. H., Guo, N. X., & He, J. Y. (2024). Can news predict firm bankruptcy? *Available at SSRN 5022837.*

Bochkay, K., Brown, S. V., Leone, A. J., & Tucker, J. W. (2022). Textual Analysis in Accounting: What's Next?. *Contemporary Accounting Research*, 40(2), 765-805.

Campbell, J. Y., Hilscher, J., Szilagyi, J. (2008). In search of distress risk. The Journal of Finance, 63(6), 2899-2939.

Chen, H., Cheng, Y., Liu, Y., & Tang, K. (2023). Teaching economics to the machines. *Available at SSRN, 4642167.*

Chen, Y., Kelly, B. T., & Xiu, D. (2022). Expected returns and large language models. *Available at SSRN 4416687*.

Cong, L. W., Gao, H., Ponticelli, J., & Yang, X. (2019). Credit allocation under economic stimulus: Evidence from China. *The Review of Financial Studies*, *32*(9), 3412-3460.

Cong, L. W., Liang, T., Zhang, X., & Zhu, W. (2025). Textual factors: A scalable, interpretable, and data-driven approach to analyzing unstructured information. *Management Science.*

Da, Z., Engelberg, J., & Gao, P. (2015). The sum of all FEARS investor sentiment and asset prices. *The Review of Financial Studies*, *28*(1), 1-32.

Davis, A. K., Ge, W., Matsumoto, D., & Zhang, J. L. (2015). The effect of manager-specific optimism on the tone of earnings conference calls. *Review of Accounting Studies*, 20(2), 639-673.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019, June). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies,* volume 1 (long and short papers) (pp. 4171-4186).

Dong, Y., Jiang, X., Liu, H., Jin, Z., Gu, B., Yang, M., & Li, G. (2024). Generalization or memorization: Data contamination and trustworthy evaluation for large language models. *arXiv preprint arXiv:2402.15938*.

Fan, J., Liu, Q., Song, Y., & Wang, Z. (2024). Measuring misinformation in financial markets. *Available at SSRN* 4922648.

Fan, J., Liu, Q., Wang, B., & Zheng, K. (2025). Unearthing financial statement fraud: Insights from news coverage analysis. *Management Science*.

Gao, H., Wang, J., Wang, Y., Wu, C., & Dong, X. (2020). Media coverage and the cost of debt. *Journal of Financial and Quantitative Analysis*, *55*(2), 429-471.

Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794.*

Huang, A. H., Wang, H., & Yang, Y. (2023). FinBERT: A large language model for extracting information from financial text. *Contemporary Accounting Research*, *40*(2), 806-841.

Houston, J. F., Lin, C. and Y. Ma. (2011). Media ownership, concentration and corruption in bank lending. *Journal of Financial Economics*, 100, 326–50.

Jiang, F., Meng, L. and Tang, G. (2021). Media textual sentiment and Chinese stock return predictability. *China Economic Quarterly*, 21(04), 1323-1344. (In Chinese)

Jiang, M., Liu, K. Z., Zhong, M., Schaeffer, R., Ouyang, S., Han, J., & Koyejo, S. (2024). Investigating data contamination for pre-training language models. *arXiv preprint arXiv:2401.06059*.

Li, Y. L., Bu, H., Li, J. H., Wu, J. J. (2020). The role of text-extracted investor sentiment in Chinese stock price prediction with the enhancement of deep learning. *International Journal of Forecasting*, 36(4), 1541-1562.

Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of Finance*, 66(1), 35-65.

Loughran, T., & McDonald, B. (2014). Measuring readability in financial disclosures. *The Journal of Finance,* 69(4), 1643-1671.

Loughran, T., & McDonald, B. (2016). Textual analysis in accounting and finance: A survey. *Journal of Accounting Research*, 54(4), 1187-1230.

Manela, A. & Moreira A. (2017). News implied volatility and disaster concerns. *Journal of Financial Economics*, 123(1), 137–162.

Ohlson, J. A. (1980). Financial ratios and the probabilistic prediction of bankruptcy. *Journal of Accounting Research*, 18(1), 109-131.

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

Price, S. M., Doran, J. S., Peterson, D. R., & Bliss, B. A. (2012). Earnings conference calls and stock returns: The incremental informativeness of textual tone. *Journal of Banking & Finance*, 36(4), 992-1011.

Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the space of topic coherence measures. In *Proceedings of the eighth ACM international conference on Web search and data mining*, 399-408.

Smales, L. A. (2016). News sentiment and bank credit risk. *Journal of Empirical Finance*, 38, 37-61.

Tian, S., Yu, Y., Guo, H. (2015). Variable selection and corporate bankruptcy forecasts, *Journal of Banking and Finance,* 52(3), 89-100.

Tsai, F. T., Lu, H. M., & Hung, M. W. (2016). The impact of news articles and corporate disclosure on credit risk valuation. *Journal of Banking & Finance*, 68, 100-116.

Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *Journal of Finance*, 62(3), 1139-1168.

Tetlock, P. C., Saar-Tsechansky, M., & Macskassy, S. (2008). More than words: Quantifying language to measure firms' fundamentals. *Journal of Finance*, 63(3), 1437-1467.

Tibshirani, R. J., Taylor, J., Lockhart, R., & Tibshirani, R. (2016). Exact post-selection inference for sequential regression procedures. *Journal of the American Statistical Association*, *111*(514), 600-620.

You, J., Zhang, B., & Zhang, L. (2018). Who captures the power of the pen? *The Review of Financial Studies*, *31*(1), 43-96.

Zhou, Y., Fan, J., & Xue, L. (2024). How much can machines learn finance from Chinese text data?. *Management Science*, *70*(12), 8962-8987.

Figures and Tables

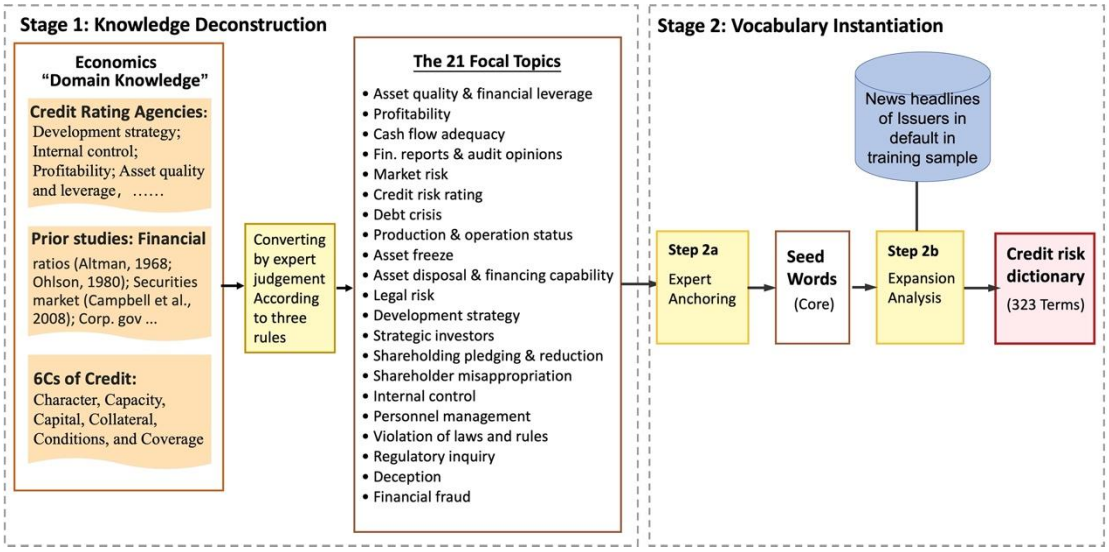


Figure 1. Framework for constructing a domain-specific credit risk dictionary.
The top eight external credit rating agencies in the Chinese bond market, adopted and reviewed in this study, include S&P, Fitch, Moody's, Dongfang Jincheng, Zhongchengxin, Dagong International, United Credit, and Shanghai New Century. A summary of the perspectives and topics used by these external credit risk rating agencies in the Chinese bond market is provided in Table B.1 of Internet Appendix B, while the variables used in some previous studies are listed in Table B.2 of Internet Appendix B. In the first step, a systematic set of filtering rules is used to synthesize the selected domain knowledge into coherent risk themes. Specifically: (1) we incorporate all topics and evaluative perspectives utilized by major credit rating agencies; (2) we retain the dimensions of the "6Cs" framework that are observable to external parties; and (3) we include key predictors validated in prior credit risk literature. In the second step, expert-anchored seed words are identified directly from domain materials to serve as the conceptual core of each theme, and we then expand this seed set by analyzing collocation patterns within the news headline corpus. The news collection period corresponding to the training sample is from May 1, 2015, to Dec. 31, 2019. The final constructed credit risk dictionary adopts words or phrases if they are explicitly identified as seed words by experts or high-frequency tokens that exhibit strong semantic links to the seed words in our training sample.



(A)Asset quality and leverage (B) Profitability (C) Cash flow adequacy (D) Audit opinions (E) Market risk (F) Credit rating

(G) Debt crisis (H) Production and operation status (I) Asset freeze (J) Asset disposal and financing capacity (K) Legal risk (L) Development strateg y

(M) Strategic investors (N) Shareholding pledge and reduction (O) Shareholder misappropriation (P) Internal control (Q) Personnel management (R) Violation of laws and rules

(S) Regulatory inquiry (T) Deception (U) Financial fraud

Figure 2. Word cloud figures for 21 credit risk dictionary topics.

The word cloud subfigures are plotted based on the news headlines of defaulted bond issuers during the news collection 3-month period for a 3-month prediction horizon $[T_D - 180,\ T_D - 90]$. Here, we provide the English translation of our credit risk dictionary for easier understanding, while the word cloud of our original Chinese version is shown in Figure IA.I of Internet Appendix.

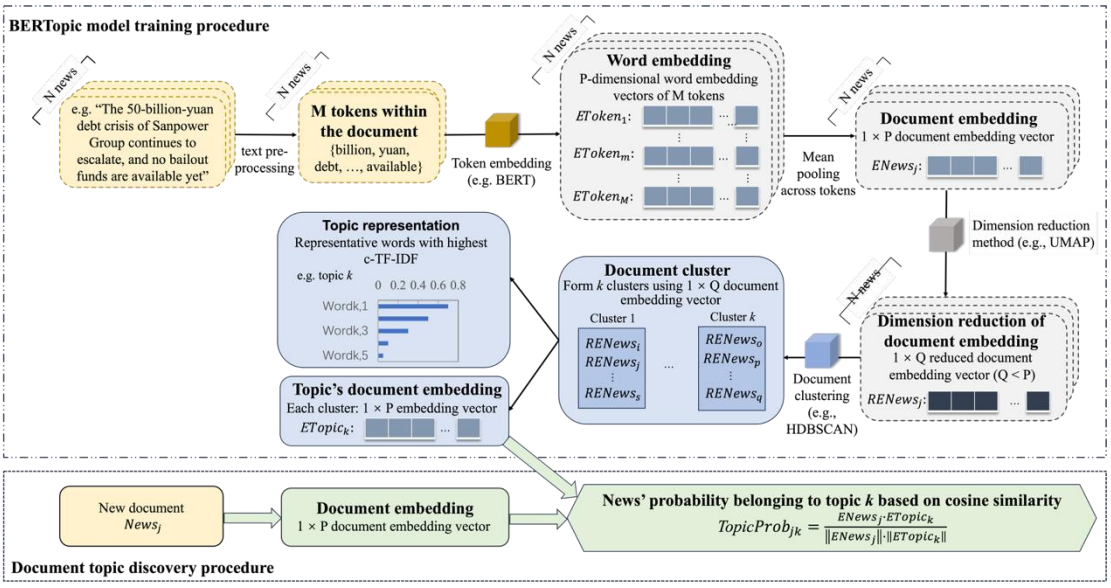
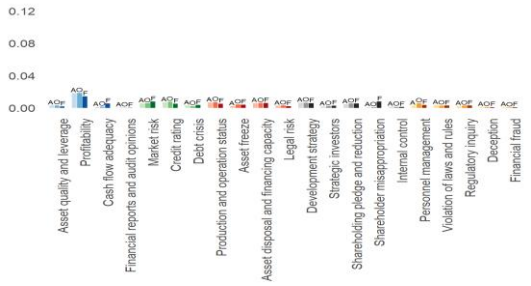
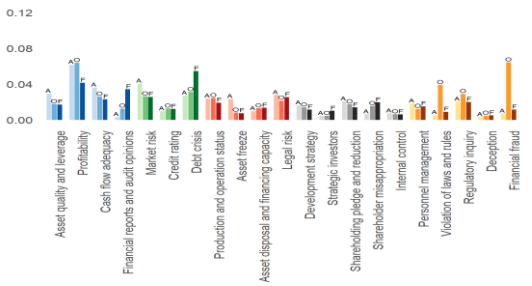


Figure 3. BERTopic architecture for identifying topics of news.

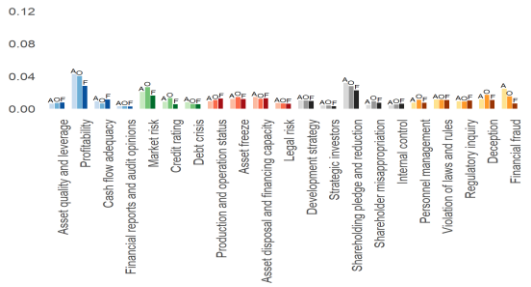
For text preprocessing, non-Chinese characters (e.g., punctuation, numbers) and bond issuer names are removed using regular expressions for each news headline. The tokenization is performed using the Jieba toolkit and eliminates function words (i.e., stop words) that carry limited economic meaning. The pre-trained LLMs, such as BERT (Devlin et al., 2019), FinBERT (Araci, 2019), are adopted for word embedding. It generates a 300-dimensional word vector for each token. Document embedding is calculated by averaging all token embeddings within the document. UMAP method proposed in McInnes et al. (2018) is used to reduce the dimensionality of document embeddings. UMAP employs a cross-entropy-based loss function to quantify the discrepancy between the high-dimensional and low-dimensional representations. HDBSCAN (McInnes et al. 2017) is employed in BERTopic to cluster document embeddings. HDBSCAN is a density-based clustering algorithm that performs well in the low-dimensional space constructed by UMAP. It constructs a hierarchy of clusters using mutual reachability distances and extracts the most stable clusters without requiring the number of topics to be specified in advance. The resulting document clusters represent distinct topics. When a specific number of topics is desired, HDBSCAN merges clusters using a hierarchical clustering structure to achieve the target number. Based on these clusters, we can obtain each topic's document embedding based on all documents in each cluster. Meanwhile, we can obtain topic representations for each generated topic using a new class-based variant of TF-IDF proposed by Grootendorst (2022). Topic representation helps interpret the economic content of topics.



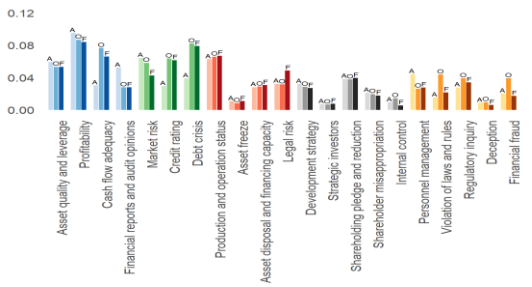
(A) State-controlled media on SOEs



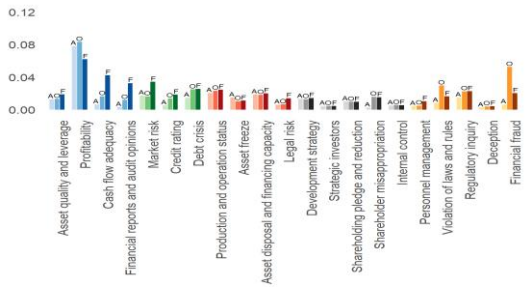
(B) State-controlled media on non-SOEs



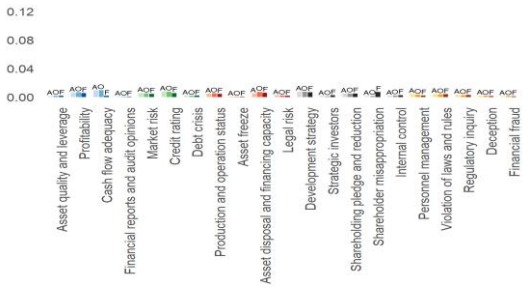
(C) Market-oriented media on SOEs



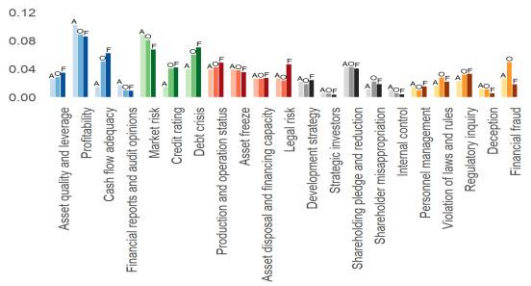
(D) Market-oriented media on non-SOEs



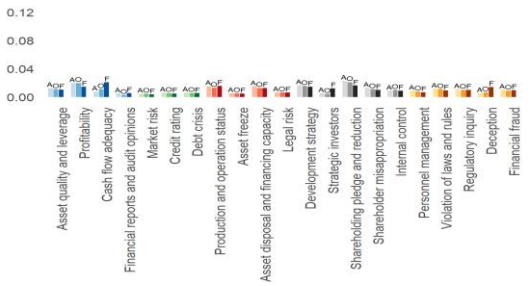
(E) State-controlled media on listed firms



(F) State-controlled media on non-listed firms



(G) Market-oriented media on listed firms



(H) Market-oriented media on non-listed firms

Figure 4. Extent of topic-level news explained by financial ratios for different firms. This figure displays the R^2 of regression models that regress standardized topic-level news measures of state-owned and market-oriented media $SNTITF_{itk}$ and $MNTITF_{itk}$ on 93 financial variables of issuers based on the different subsamples, respectively. We construct available, on-going, and future financial information using the financial ratios from the current available quarterly report at time *t*, the next quarterly report following it, and the second subsequent quarterly report, respectively. The results for available, on-going, and future financial information are recorded as "A", "O", "T". The results of topics in different perspectives are displayed in different colors.

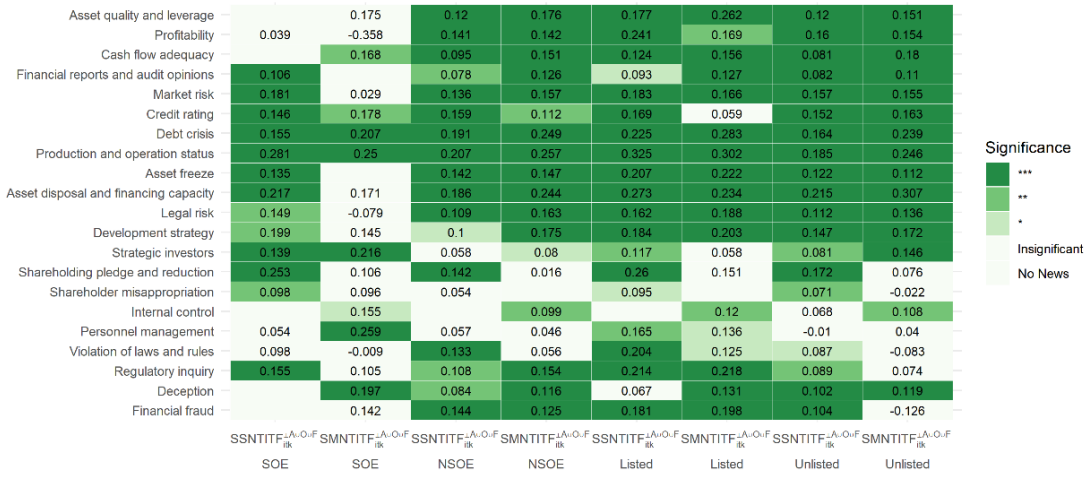


Figure 5. Incremental value of orthogonalized topic-level news measure.
This figure reports the coefficient estimation of the topic-level news measure reported by state-controlled media and market-oriented media for defaulted issuers for the first time, using Eq. (2) without control variables. $SSNTITF_{itk}^{\perp AUOUF}$ and $SMNTITF_{itk}^{\perp AUOUF}$ represent the standardized orthogonal variables for topic-level news measures of state-owned and market-oriented media, removing the portion explained by available, ongoing, and future financial information. The results are reported under different sub-samples. "SOE", "NSOE", "Listed" and "Unlisted" indicate the subsamples of state-owned enterprises, non-state-owned enterprises, listed firms and unlisted firms. Four shades of green, from dark to light, indicate the significance levels of the estimated coefficients: 1%, 5%, 10%, and non-significant. Three shades of green, from dark to light, indicate the significance levels of the estimated coefficients: 1%, 5%, and 10%. White color denotes insignificant topics. The blank parts indicate topics for which no media news coverage was available from the respective media source for the sampled defaulted bond issuers.

Table 1. Descriptive statistics on the firm-level news and risk-related news

This table reports statistics using data from May. 1, 2015, to Dec. 31, 2022. The "news coverage" reports the proportion of firms that have historical news. "Issuers in default" refers to issuers that were in default during the news collection period. The minimum number of firm-level news for all kinds of issuers are zero.

For bond issuers [at a specific period]	New coverage	Statistics					
		Mean	P25	Median	P75	Max	Std.
Panel A: Descriptive statistics on the firm-level news							
All issuers	87.37%	778	4	83	1,149	29,266	1,605
Listed companies	99.42%	1,426	573	1,203	1,623	15,423	1,485
Non-listed firms	82.57%	520	2	13	273	29,266	1,578
SOEs	86.80%	736	3	33	1,013	29,266	1,692
non-SOEs	88.24%	841	8	258	1,278	15,423	1,461
Issuers in default	97.02%	969	54	314	1,212	13,546	1,647
Before 1st default	93.19%	376	13	82	274	5,971	777
$[T_D - 90,\ T_D]$	82.98%	50	2	12	41	658	99
$[T_D - 180,\ T_D - 90]$	77.45%	45	1	7	27	866	106
$[T_D - 270,\ T_D - 180]$	74.47%	34	0	6	23	455	75
$[T_D - 360,\ T_D - 270]$	73.19%	29	0	6	21	479	61
$[T_D - 450,\ T_D - 360]$	67.66%	27	0	4	17	512	68
After 1st default within 3m	78.41%	48	1	9	46	609	95
Panel B: Statistics of firm-level credit risk-related news using domain-informed dictionary							
All issuers	74.79%	121	0	15	187	3954	236
Listed companies	99.10%	222	68	198	278	1751	218
Non-listed firms	65.58%	83	0	3	48	3954	231
SOEs	73.08%	107	0	7	143	3954	236
non-SOEs	77.15%	139	1	39	217	2460	235
Issuers in default	91.91%	235	11	98	286	2460	381
Before 1st default	80.43%	89	1	21	83	1509	189
$[T_D - 90,\ T_D]$	68.94%	14	0	4	14	303	31
$[T_D - 180,\ T_D - 90]$	57.02%	13	0	1	8	384	40
$[T_D - 270,\ T_D - 180]$	54.04%	9	0	1	8	227	25
$[T_D - 360,\ T_D - 270]$	50.21%	7	0	1	5	126	16
$[T_D - 450,\ T_D - 360]$	46.38%	6	0	0	3	222	18
After 1st default within 3m	72.73%	18	0	5	18	208	32

Table 2. Statistics of news metric and correlations

Panel A reports the descriptive statistics of the news metric *NTITF*. Panel B reports its Pearson correlation coefficients with credit risk measures and economic measures, with significance levels at 1%, 5%, and 10% denoted by ***, **, and *, respectively. The analysis is based on a training sample (from May 1, 2015, to Dec. 31, 2019) constructed by matching all available non-defaulted to each defaulted issuer method. This matching procedure yields 43,505 unique firm-day observations from 3,178 non-defaulted firms, serving as the control group for 131 defaulted issuers. We construct three dummy variables to capture key rating information from credit rating agencies: *NonInvest* indicates non-investment-grade status; ISDowngrade denotes rating downgrades; and *FallenAngel* flags issuers transitioning from investment to speculative grade. *ZS* and *OS* are values of Z-score and O-score models calculated directly using the function reported in their paper. *OS* represents the log-odds of credit risk, i.e., ln(*P*/(1-*P*)), where *P* is the probability of default. Observations of *ZS* with missing values are excluded only for the specific pair of variables being analyzed. To save space, the correlation analysis of *NTITF*, credit risk indicators, and financial ratios is reported in Table IA.IX in the Internet Appendix.

Panel A: Descriptive statistics of news metric *NTITF*						
Variable	# of obs.	Mean	SD	Q50	Q75	Q90
NTITF	65913	79.539	330.489	0.000	30.338	190.564
Panel B: Correlations of *NTITF* with credit risk and economic measures						
	NTITF	*ZS*	*OS*	*NonInvest*	*ISDowngrade*	*FallenAngel*
ZS	-0.031***					
OS	0.013***	-0.372***				
NonInvest	0.034***	-0.035***	-0.012**			
ISDowngrade	0.029***	-0.102***	0.040***	0.329***		
FallenAngel	0.040***	-0.022***	-0.008	0.626***	0.298***	
PPI	0.031***	-0.028***	0.014***	0.031***	-0.001	-0.048***
FAI	-0.113***	0.051***	-0.059***	-0.011**	-0.010*	0.037***
AFRE	-0.025***	-0.007	-0.044***	0.026***	-0.009*	-0.019***

Table 3. Incremental explanatory power of *NTITF* across horizons

This table reports the estimation results for the training sample at 3-month and long-term horizons. Panel A reports the results based on a sample constructed by a 1:3 random matching method for a 3-month forecasting horizon. Panel B reports the results based on a sample constructed by matching all available non-defaulted issuers using the method for a 3-month forecasting horizon. Panel C reports the results for long-term forecasting horizons based on an ample constructed using a 1:3 random-matching method. "Z-score", "O-score", and "LRSFM" in columns 2 to 4 of Panels A and B report the results of LR models with control variables from the Z-score model, O-score model, and LRSFM. LRSFM is our new benchmark model obtained using bidirectional stepwise regression from a pool of 16 financial ratios and three year-on-year growth rates of quarterly economic indicators, as listed in Table A.I of Appendix. All financial ratios and economic variables are available for each issuer when the prediction is made. To enhance readability, we report the coefficients of *NTITF*/10 in the table. To The z-statistics of coefficients are reported in parentheses. *** (**, *) represents the significant coefficient at a 1% (5%, 10%) significance level.

Panel A: 3-month forecasting based on sample constructed with 1:3 randomly matching method

	(1)	(2) Z-score	(3) O-score	(4) LRSFM
$NTITF/10$	0.023	0.022	0.023	0.021
	(5.291)***	(5.247)***	(5.104)***	(4.935)***
Controls	N	Y	Y	Y
Intercept	-1.414	-1.514	-1.975	-2.181
	(-12.018)***	(-7.459)***	(-4.446)***	(-5.428)***
AIC	539.963	523.119	532.528	509.743
Log likelihood	-267.981	-254.559	-255.264	-242.872
Pseudo R^2	0.091	0.136	0.134	0.176
Benchmark Pseudo R^2		0.053	0.051	0.102
Improved Pseudo R^2		0.083	0.082	0.074

Panel B: 3-month forecasting based on sample constructed with all available matching method

	(1)	(2) Z-score	(3) O-score	(4) LRSFM
$NTITF/10$	0.011	0.011	0.011	0.010
	(10.864)***	(10.365)***	(10.413)***	(9.723)***
Controls	N	Y	Y	Y
Intercept	-5.941	-5.807	-5.277	-5.586
	(-64.272)***	(-38.322)***	(-13.990)***	(-22.447)***
AIC	1704.516	1687.911	1685.826	1641.461
Log likelihood	-850.258	-836.955	-831.913	-808.730
Pseudo R^2	0.047	0.061	0.067	0.093
Benchmark Pseudo R^2		0.020	0.028	0.055
Improved Pseudo R^2		0.042	0.040	0.038

Panel C: Long-term forecasting based sample constructed with 1:3 randomly matching method

	h=4m	h=6m	h=9m	h=12m
$NTITF/10$	0.016	0.010	0.011	0.012
	(4.246)***	(3.721)***	(2.750)***	(2.240)**
Controls	Y	Y	Y	Y
Intercept	-2.213	-2.111	-1.325	-2.007
	(-5.651)***	(-4.848)***	(-3.601)***	(-3.566)***
AIC	528.982	545.775	684.458	543.403
Log likelihood	-253.491	-261.888	-332.229	-260.701
Pseudo R^2	0.140	0.111	0.077	0.115
Baseline Pseudo R^2	0.084	0.078	0.056	0.100
Improved Pseudo R^2	0.056	0.033	0.021	0.015

Table 4. Out-of-sample 3-month horizon forecasting performance of the news metric
This table presents the out-of-sample predictive performance of the news variable, *NTITF*, using several machine learning models trained on a training sample generated by a 1:3 random matching method. The default probability threshold for the logistic regression model is optimized over 0.3-0.7 to maximize recall while ensuring precision exceeds 0.4, using 5-fold cross-validation. "LR + Z-score" and "LR + O-score" are LR models with variables from Altman Z-score or Ohlson O-score. Using the same 16 financial ratios and 3 economic variables as inputs, this paper further applies machine learning models, including XGBoost, SVM, and DT, to examine the incremental predictive power of the *NTITF* news variable. Similarly, the machine learning models set their threshold parameters to maximize recall while constraining precision to above 0.4. To save space, we only report the performance measures of recall, F_1 score, precision, and accuracy here. The "Algorithm-specific benchmark models" are models with only control variables under different prediction models. The improved recall and F_1 scores compared to the respective benchmark models are reported in parentheses.

	Adding *NTITF* into benchmark models				Algorithm-specific benchmark models	
	Accuracy	Precision	Recall	F_1 score	Recall	F_1 score
LR + Z-score	0.613	0.352	0.654 (0.375)	0.458 (0.144)	0.279	0.314
LR + O-score	0.683	0.683	0.683 (0.443)	0.518 (0.178)	0.240	0.340
LRSFM	0.649	0.394	0.750 (0.135)	0.517 (0.017)	0.615	0.500
XGBoost	0.620	0.378	0.808 (0.096)	0.515 (0.013)	0.712	0.502
SVM	0.714	0.452	0.673 (0.336)	0.541 (0.223)	0.337	0.318
DT	0.651	0.387	0.673 (0.173)	0.491 (0.097)	0.500	0.394

Table 5. Out-of-sample long-term prediction performance of the news metric
This table reports the out-of-sample early warning performance of the news measure across different default prediction models, trained on a 1:3 randomly matched training sample. The LR benchmark model is obtained via stepwise regression of 16 financial ratios and 3 economic variables across different forecasting horizons. The LR model's estimation results on the training sample are reported in Table 3. The measures of recall and F_1 score improved relative to benchmark models and are displayed in bold.

	Adding *NTITF* into benchmark models				Corresponding benchmark models	
	Accuracy	Precision	Recall	F_1 score	Recall	F_1 score
Panel A: 4-month forecasting horizon						
LR	0.647	0.393	**0.760**	**0.518**	0.587	0.492
XGBoost	0.625	0.381	**0.798**	**0.516**	0.692	0.497
SVM	0.731	0.477	**0.798**	**0.597**	0.510	0.463
DT	0.589	0.340	**0.683**	**0.454**	0.510	0.417
Panel B: 6-month forecasting horizon						
LR	0.661	0.403	**0.740**	**0.522**	0.587	0.494
XGBoost	0.649	0.389	**0.712**	**0.503**	0.663	0.474
SVM	0.666	0.374	0.500	0.428	0.587	0.477
DT	0.550	0.326	**0.750**	**0.455**	0.471	0.410
Panel C: 9-month forecasting horizon						
LR	0.745	0.492	**0.567**	**0.527**	0.327	0.378
XGBoost	0.649	0.365	**0.548**	**0.438**	0.404	0.377
SVM	0.675	0.358	**0.375**	**0.366**	0.279	0.266
DT	0.627	0.333	**0.490**	**0.397**	0.385	0.336
Panel D: 12-month forecasting horizon						
LR	0.704	0.417	**0.462**	**0.438**	0.240	0.298
XGBoost	0.651	0.380	**0.625**	**0.473**	0.500	0.432
SVM	0.692	0.387	**0.394**	**0.390**	0.288	0.286
DT	0.685	0.369	**0.365**	**0.367**	0.250	0.287

Table 6. Predictive power of topics extracted from domain-informed dictionary and data-driven BERTopic models
Panels A and B present estimation results for conventional and constrained (size-21) LASSO models using topics derived from BERTopic and our domain-informed dictionary. The models predict defaults over a 3-month horizon using a training sample generated via a 1:3 random-matching procedure. Topic exposure variables (*NT*) are calculated based on news coverage in the 3-month window ending at $t^* = T_D - 90$. For data-driven topics, we employ BERTopic with embeddings from FinBERT (Araci, 2019) and BERT (Devlin et al., 2019). The LASSO penalty parameter (λ) is selected via five-fold cross-validation to maximize recall, searching a grid from e^{-2} to e^{-10} with increments of $e^{-0.1}$. Figures IA.II and IA.III in the Internet Appendix detail the selected topics, indexed by cluster IDs derived from the data-driven BERTopic model. Statistical significance for data-driven models is assessed using the post-selection inference procedure of Tibshirani et al. (2016), whereas p-values for domain-informed models are derived from standard logistic regression.

FinBERT embedding		BERT embedding		Domain-informed credit risk dictionary	
Topic	Coef.	Topic	Coef.	Topic	Coef.
Panel A: Topic selection of conventional LASSO for *NT* from data-driven BERTopic models or from domain-informed credit risk dictionary					
Topic 2	0.058 (0.198)	Topic 7	0.106 (0.000)	Cash flow adequacy	0.204 (0.979)
Topic 15	0.280 (0.000)	Topic 21	0.119 (0.000)	Reports and audit opinions	0.027 (0.267)
Topic 45	0.191 (0.000)	Topic 93	0.055 (0.000)	Credit rating	0.348 (0.012)
Topic 65	0.030 (0.000)			Debt crisis	0.250 (0.050)
				Asset freeze	0.659 (0.004)
				Strategic investors	0.110 (0.131)
Pseudo R^2	0.076	Pseudo R^2	0.044	Pseudo R^2	0.135
Panel B: Topic selection of size-21-constrained LASSO for *NT* from data-driven BERTopics models or from domain-informed credit risk dictionary					
Pseudo R^2	0.195	Pseudo R^2	0.160	Pseudo R^2	0.201

Table 7. Out-of-sample prediction performance of the model with data- and expert-driven topics

The 3-month horizon out-of-sample prediction performance is examined using several prediction models based on the testing sample constructed by 1:3 matching method. The settings of machine learning models are similar to the notes of Table 4. We compare our expert-driven topics with the data-driven topics obtained from BERTopic models with different word embedding and selected using the LASSO models reported in Table 6, by incorporating topic-level *NT* into algorithm-specific benchmark models with 16 financial ratios and 3 economic variables as control variables. The benchmark model of LR model is LRSFM, which is obtained by the stepwise procedure. To save space, we only report recall and F_1 score here. When the performance of the prediction model by incorporating topic-level variables into the benchmark model with control variables of financial ratios and economic variables is improved, the non-improved recall and F1 score measures are illustrated in italics. The 3-month horizon out-of-sample prediction performance of algorithm-specific benchmark models with only control variables can be found in the last two columns in Table 4. The best performance is highlighted in bold.

	Expert-driven topics from our dictionary		Data-driven topics from BERTopic based on FinBERT embedding		Data-driven topics from BERTopic based on BERT embedding	
	Recall	F_1 score	Recall	F_1 score	Recall	F_1 score
Panel A: Using topic variables selected from size-21-constrained LASSO models						
	21 topic-level *NT*		FinBERT embedding		BERT embedding	
LR	**0.788**	**0.788**	*0.615*	*0.478*	*0.606*	0.508
XGBoost	**0.760**	**0.760**	*0.654*	*0.496*	*0.683*	*0.485*
SVM	**0.529**	**0.529**	0.481	0.450	**0.529**	0.495
DT	**0.663**	**0.663**	0.577	0.469	0.519	0.429
Panel B: Using optimal topic variables selected by conventional LASSO						
	6 topic-level *NT*		4 topics based on FinBERT embedding		3 topics based on BERT embedding	
LR	**0.808**	**0.808**	0.731	*0.486*	0.683	0.538
XGBoost	**0.731**	**0.731**	*0.663*	*0.479*	0.721	0.505
SVM	**0.760**	**0.760**	0.692	0.505	0.404	0.402
DT	**0.731**	**0.731**	0.615	0.404	0.654	0.455

Table 8. Out-of-sample performance of models with *NN* constructed by different methods

This table reports the out-of-sample default prediction performance for a 3-month forecasting horizon based on the test sample. The forecasting model is trained and tested based on the sample constructed with 1:3 matching method. With an instruction-based prompting approach, we employ ChatGPT-4.1 to assess whether a news headline contains predictive signals relevant to default prediction. The specific prompt is as follows: "You are a professional credit risk analyst. Please determine whether each news headline contains information that can be used for positively predicting corporate credit risk events. If it does, label it as 1; otherwise, label it as 0. Please output the credit label for each news headline." All news headlines have been anonymized by concealing company names, stock tickers, bond codes, and other identifiers. When the prediction performance of incorporating *NN* into the prediction model with control variables is not improved according to different performance metrics (e.g., recall, F_1 scores), the non-improved performance metrics are highlighted in italics. The algorithm-specific benchmark models are those with control variables of 16 financial ratios and 3 economic variables, while the benchmark model of LR model is LRSFM, which is obtained by the stepwise procedure. The 3-month horizon out-of-sample prediction performance of algorithm-specific benchmark models can be found in Table 4. The best performance is highlighted in bold.

Prediction model	Our credit risk dictionary		ChatGPT		General dictionary of HowNet		Media tone dictionary of Jiang et al. (2021)	
	Recall	F_1 score	Recall	F_1 score	Recall	F_1 score	Recall	F_1 score
LR	**0.760**	0.520	*0.702*	*0.479*	0.683	**0.532**	0.683	0.518
XGBoost	**0.770**	0.505	*0.702*	**0.510**	*0.692*	*0.482*	**0.770**	0.486
SVM	**0.567**	**0.480**	0.413	0.371	0.490	0.413	0.519	0.450
DT	**0.683**	0.460	0.673	**0.505**	0.577	0.448	0.607	0.339

Table 9. Predictive power of media-level news under different information density

This table compares the predictive power of news reported by state-owned and market-based media for defaults, using the conditional LR model framework employed in this study at a 3-month forecasting horizon. State-controlled media outlets are identified according to the latest "List of Internet News and Information Source Units" released by the Cyberspace Administration of China on October 20, 2021. This list includes 38 state-owned news units, totaling 241 media outlets, while our sample includes 3,889 market-oriented media outlets. A new media-level news measure of credit risk information ratio is adopted and calculated separately for state-owned and market-oriented media, SCR_{it} and MCR_{it}, to make the scale of these two variables similar. Panels A and B present the estimation results for subsamples of bond issuers from listed and non-listed companies, respectively, covering the whole sample period from Nov. 1 2015, to Dec. 31, 2022. The analysis employs a matching method using all available non-defaulted issuers, which results in an observation ratio of 63 first-time defaulted issuers (Group 1) to 21,369 non-defaulted issuers (Group 2) for bond issuers from listed companies. For non-listed companies, the corresponding ratio is 172 observations of first-time defaulters (Group 1) to 44,309 observations of non-defaulters (Group 2).

	(1) LRSFM	(2) LRSFM	(3) Z-score	(4) O-score	(5) LRSFM
Panel A: Subsample of listed companies as bond issuers					
SCR	0.011***		0.011***	0.010***	0.009***
	(4.782)		(3.638)	(3.675)	(3.319)
MCR		0.005***	0.002	0.003	0.003
		(3.717)	(1.058)	(1.485)	(1.530)
Intercept	-4.823***	-4.736***	-4.392***	-6.436***	-4.821***
	(-18.061)	(-17.986)	(-16.751)	(-4.529)	(-18.033)
Controls	Y	Y	Y	Y	Y
AIC	771.761	778.192	787.224	804.373	771.869
Log likelihood	-373.880	-377.096	-385.612	-390.186	-372.934
Pseudo R^2	0.131	0.123	0.104	0.093	0.133
# of observations in Group 1 vs Group 2	63: 21369	63: 21369	63: 21369	63: 21369	63: 21369
Panel B: Subsamples of non-listed companies as bond issuers					
SCR	0.005***		0.004***	0.005***	0.004***
	(3.890)		(3.346)	(3.903)	(3.201)
MCR		0.006***	0.005***	0.004***	0.005***
		(6.147)	(4.870)	(4.603)	(5.4211)
Intercept	-5.426***	-5.417***	-5.266***	-4.423***	-5.438***
	(-30.863)	(-30.987)	(-38.608)	(-10.015)	(-31.037)
Controls	Y	Y	Y	Y	Y
AIC	2052.209	2044.446	2112.693	2100.476	2039.985
Log likelihood	-1014.104	-1010.223	-1048.346	-1038.238	-1006.993
Pseudo R^2	0.100	0.104	0.070	0.079	0.107
# of observations in Group 1 vs Group 2	172: 44309	172: 44309	172: 44309	172: 44309	172: 44309

Table 10.Predictive performance of different media across issuers' ownership nature
This table presents the results about the incremental information value of media-level news published by state-owned and market-based media for predicting first-time defaults at a 3-month forecasting horizon, under the consideration of bond issuers' different nature of corporate ownership (CSOEs, SOEs, but non-CSOEs, and non-SOEs) in different panels. The number of defaulted issuers for CSOEs, SOEs but non-CSOEs, non-SOE are 9, 30, 196 in the whole sample. The sample is constructed using all available matching methods to make use of more data. LRSFM model is the best model using stepwise regression with input variables of 16 financial ratios and 3 economic variables based on the whole sample from Nov. 1, 2015, to Dec. 31, 2022.

	(1) LRSFM	(2) LRSFM	(3) Z-score	(4) O-score	(5) LRSFM
Panel A: Subsample of bond issuers from CSOEs					
SCR	0.008		0.003	0.004	0.004
	(1.571)		(0.808)	(0.700)	(0.459)
MCR		0.008**	0.006*	0.006*	0.007
		(2.473)	(1.821)	(1.7519)	(1.450)
Intercept	-6.734***	-6.753***	-6.093***	-7.134	-6.773***
	(-8.955)	(-8.920)	(-8.283)	(-1.357)	(-8.868)
Controls	Y	Y	Y	Y	Y
AIC	145.224	143.898	140.693	146.144	145.720
Log likelihood	-60.612	-59.949	-62.346	-61.072	-59.860
Pseudo R^2	0.199	0.207	0.176	0.192	0.208
# of observations in Group 1 vs Group2	9:14758	9:14758	9:14758	9:14758	9:14758
Panel B: Subsample of bond issuers from SOEs but non-CSOEs					
SCR	0.004**		0.004*	0.003	0.003
	(1.996)		(1.773)	(1.535)	(1.471)
MCR		0.006***	0.005***	0.005***	0.005***
		(3.627)	(3.011)	(3.165)	(3.355)
Intercept	-6.227***	-6.230***	-6.082***	-9.063***	-6.255***
	(-16.145)	(-16.030)	(-18.162)	(-3.419)	(-16.062)
Controls	Y	Y	Y	Y	Y
AIC	412.149	408.573	419.668	420.651	409.468
Log likelihood	-194.075	-192.286	-201.834	-198.325	-191.734
Pseudo R^2	0.189	0.196	0.157	0.171	0.199
# of observations in Group 1 vs Group2	30:32094	30:32094	30:32094	30:32094	30:32094
Panel C: Subsample of bond issuers from non-SOEs					
SCR	0.010***		0.011***	0.012***	0.009***
	(5.319)		(5.550)	(6.067)	(4.392)
MCR		0.004***	0.004***	0.004***	0.003***
		(4.150)	(3.273)	(3.327)	(3.014)
Intercept	-4.019***	-3.944***	-3.732***	-4.152***	-4.040***
	(-23.141)	(-22.969)	(-25.018)	(-6.992)	(-23.032)
Controls	Y	Y	Y	Y	Y
AIC	1955.089	1963.864	2031.665	2005.845	1951.053
Log likelihood	-965.545	-969.932	-1007.833	-990.923	-962.526
Pseudo R^2	0.116	0.112	0.077	0.092	0.118
# of observations in Group 1 vs Group2	196:18826	196:18826	196:18826	196:18826	196:18826

Appendix

Table A. I. Definitions of news metrics, financial ratios, and economic controls

Category	Aspects	Variables
Panel A: News variable proposed in this study		
News metric		$NTITF_{it} = \sum_k \sum_{j \in D(t)} TMentioned_{ijk} \cdot \ln\left(\frac{N_m}{N_{k,m}}\right)$
Alternative news metric		$NTTDW_{it} = \sum_{j_t \in D(t)} \sum_k TMentioned_{ijk} \cdot Score_k$
		$NT_{it} = \sum_{j_t \in D(t)} NT_{ij}$
		$NWITF_{it} = \sum_{j_t \in D(t)} \sum_k NW_{ijk} \cdot \ln\left(\frac{N}{N_k}\right)$
		$NWTDW_{it} = \sum_{j_t \in D(t)} \sum_k NW_{ijk} \cdot Score_k$
		$NN_{it} = \sum_{j_t \in D(t)} IsNeg_{ij}$
Panel B: financial ratios from classic models		
Financial ratios	Z-score model	Working Capital/Total Assets (*WCOA*) Retained Earnings/Total Assets (*REOA*) Earnings Before Interest and Taxes/Total Assets (*EBITOA=EBIT/TA*) Market Value of Equity/Book Value of Total Debt (*EOD=MVE/TL*) Sales/Total Assets (Total Asset Turnover Ratio, *TATR=S/TA*)
Financial ratios	O-score model	*RSIZE*=log(total assets/GNP price-level index) *DR*=Total Liabilities/Total Assets *WCOA*= Working Capital/Total Assets *CLCA* =Current Liabilities/Current Assets *OENEG*= One if total liabilities exceed total assets, zero otherwise NITA=Net Income/Total Assets FUTL=Funds provided by Operations/Total Liabilities *INTWO*=1 if net income was negative for the last two periods *CHIN*=$(NI_t - NI_{t-1})/(\lvert NI_t \rvert + \lvert NI_{t-1} \rvert)$, where NI_t is net income for the most recent period.
Panel C: Financial ratios, accounting terms and economic indicators used in benchmark model LRSFM		
Financial ratios	Profitability	Return on equity (*ROE*) Return on assets (*ROA=Net Income/Average Total Assets*) Retained earnings on total asset (*REOA=RE/TA*) Net profit margin on sales (*NPMOS*)
	Liquidity	Working capital on total assets (*WCOA=WC/TA*) Quick ratio (*QR*) Cash ratio (*CR*)
	Solvency	Interest coverage ratio (*ICR*)
	Leverage	Debt ratio (*DR*=Total Liability/Total Assets) Debt to equity ratio (*DER=Debt/Equity)*
	Operating efficiency	Total asset turnover ratio (*TATR=Sales/Average total assets*) Accounts receivable turnover ratio (*ARTR*) Inventory turnover ratio (*ITR*)
Growth rate	Accounting term	Year-on-year growth rate of total assets (*GRA*) Year-on-year growth rate of net profit (*GRNP*) Year-on-year growth rate of revenue (*GRR*)
Economic variables		Year-on-year growth rate of quarterly PPI (*PPI*) (Da et al. 2015) Fixed-asset investment (*FAI*) (Cong et al., 2019) Aggregate financing to the real economy (*AFRE*) (Cong et al., 2019)