Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Distilling large language model embeddings of annual report narratives across disclosure categories for corporate financial distress prediction

Leping Jiang

jiangleping1985@163.com

Guangdong University of Science and Technology

Huiping Huang

Guangdong University of Science and Technology

Youyu Lu

Suqian University

Research Article

Keywords: financial distress prediction, large language model, text mining, multi-modal fusion, MD, &A analysis, gradient boosting, explainable AI, Chinese A-share market

Posted Date: May 26th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9560505/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Distilling large language model embeddings of annual report narratives across disclosure categories for corporate financial distress prediction

Leping Jiang[1*], Huiping Huang[1] and Youyu Lu[2]

[1*]Guangdong University of Science and Technology, Dongguan, 523083, Guangdong, China.

[2]Suqian University, Suqian, 223800, Jiangsu, China.

*Corresponding author(s). E-mail(s): jiangleping1985@163.com;
Contributing authors: 15016855230@139.com; luyouyu2011@163.com;

Abstract

Financial distress prediction (FDP) has traditionally relied on numerical financial ratios, overlooking the qualitative information in firm narrative disclosures. Integrating Large-Language-Model (LLM) embeddings of such text with tabular financial features is non-trivial: the 2048-dimensional embeddings overwhelm the dense financial signal in gradient-boosted classifiers, and collapsing heterogeneous narratives into a single vector erases semantic structure. We propose LIFT (LLM-Integrated Feature Transformation), a two-stage framework that splits each Management Discussion and Analysis (MD&A) section into five standardised category corpora—business review, industry, risk, outlook, and core competitiveness—and embeds each independently. A masked-attention multilayer perceptron distils the five 2048-dimensional vectors into three calibrated class probabilities, which are concatenated with 140 financial ratios and fed to a class-weighted LightGBM stacker. On 6082 Chinese A-share firm-year samples spanning 2019–2023, LIFT improves macro-F1 by 8.2 percentage points on five-fold cross-validation and 4.8 points on a temporal 2022–2023 test split. Warning-class recall rises from 84.3% to 94.0% and distressed-class recall from 58.2% to 65.2%. SHAP attribution places all three text-derived features in the top eight of 143 predictors, and a case study identifies approximately 140 distressed firms correctly flagged by LIFT but missed by the financial baseline.

Keywords: financial distress prediction, large language model, text mining, multi-modal fusion, MD&A analysis, gradient boosting, explainable AI, Chinese A-share market

JEL Classification: G17 , G33 , C53 , C55

1 Introduction

Financial distress prediction (FDP) is a central problem in financial regulation and applied analytics, combining high practical value with considerable technical difficulty. Large-scale corporate distress episodes have left long trails of economic and social cost, from the wave of bankruptcies that accompanied the 2008 global financial crisis to more recent defaults across real-estate, manufacturing, and other capital-intensive sectors; reliable early warning of such events is therefore of direct value to regulators seeking to pre-empt systemic risk, to creditors and bondholders pricing and hedging their exposures, and to investors and auditors seeking to avoid abrupt losses. This practical value is reflected in a decades-long research stream that has shaped credit-rating methodology, bank stress-testing frameworks, and listed-firm supervisory regimes across major markets. What makes FDP technically non-trivial, in turn, is the interaction of severe class imbalance (distressed firms are a small minority of listed firms in any given year), temporal non-stationarity (distress waves shift with macroeconomic cycles, industry shocks, and regulatory regimes), and heterogeneous information sources that span accounting statements, regulatory filings, news and analyst coverage, and firms' own narrative disclosures.

Classical FDP methods build their predictors primarily on numerical financial ratios and have achieved meaningful early-warning performance. Typical inputs are on the order of 30 to 150 accounting ratios spanning profitability, solvency, liquidity, turnover, and per-share indicators, often extended with their year-over-year growth rates. Typical models range from Altman's Z-score [1] and Ohlson's O-score [2] to modern tree-based gradient boosting machines such as XGBoost [3] and LightGBM [4]. The reason these models work is direct: firms that ultimately enter distress tend to display persistent accounting signatures—shrinking margins, rising leverage, weakening cash conversion—and these signatures are well captured by structured financial ratios, which is why gradient boosting on tabular features remains the dominant default in the applied FDP literature [5–7].

However, a non-trivial share of the information relevant to distress resides outside the tabular statement space, and classical FDP methods cannot reach it. Industry-cycle shocks, regulatory-policy shifts, litigation exposure, management turnover, supply-chain disruptions, the sudden loss of a major customer, and similar qualitative events are typically discussed in a firm's narrative disclosures well before they propagate fully into the income statement. Forward-looking statements and management commentary in the Management Discussion and Analysis (MD&A) section also express expectations that have no direct numerical counterpart in the current period's ledger. A model that only reads accounting ratios will structurally miss these signals and therefore systematically under-detect the subset of distressed firms whose troubles surface in narrative before they surface in numbers.

A second, related limitation is that classical models cannot exploit cross-modal interactions as discriminative signals. Management has both incentives and discretion to present favourable numbers; when narrative optimism diverges from deteriorating fundamentals—or, conversely, when numerical softness is accompanied by a candid risk disclosure—the divergence itself carries diagnostic information [8]. A purely numerical pipeline cannot see this divergence because it consumes only one side of the pair.

Incorporating the textual modality into FDP has been an active research topic for more than a decade. In the pre-LLM era, textual disclosures were typically mapped to low-dimensional hand-crafted features via sentiment lexicons [8], LDA topic weights [9], or bag-of-words counts. Such pipelines are reproducible and interpretable, but they impose a steep information-loss step: a multi-thousand-word risk section is collapsed to a handful of scalar proxies, discarding much of the dependency structure that makes narrative disclosure informative. The emergence of large language models (LLMs), and in particular the embedding subclass that maps arbitrary-length text directly to a fixed-dimensional dense vector, has fundamentally changed what is feasible here. Such text embeddings preserve enough semantic content to support demanding downstream tasks [10, 11], strongly suggesting that the distilled vectors retain finer-grained signals than hand-crafted sentiment or topic features can recover.

Given this technical substrate, several routes exist for letting textual information assist an FDP classifier: fine-tuning a domain-adapted encoder end-to-end on distress labels, training a deep neural classifier that jointly ingests numerical and textual inputs, or stacking an LLM encoder atop a traditional tabular learner. The route adopted in this paper keeps the LLM encoder frozen as a feature extractor and places a small supervised layer between the encoder and a tree-based tabular classifier. This choice is deliberately lightweight: it inherits the strong predictive power of gradient boosting on structured financial data, it does not require the compute budget or labelled-text volume needed for reliable encoder fine-tuning, and it yields a compact final feature space on which per-prediction SHAP-style interpretation [12] remains feasible—an essential property in a regulatory-facing application. This route faces two intrinsic challenges. First, LLM-produced embeddings are high-dimensional (typically 1024 or 2048 dimensions) and not every dimension carries an interpretable signal; fed directly to a gradient boosting model, these sparse dimensions overwhelm the dense tabular financial features. Second, collapsing a long, heterogeneous narrative document into a single dense vector loses the semantic structure across its sub-sections.

To address both challenges, we propose **LIFT** (LLM-Integrated Feature Transformation). **Contribution 1** is a lightweight masked-attention multilayer perceptron that distils category-specific high-dimensional embeddings into three calibrated class probabilities, producing a compact tree-friendly representation that aligns the density of the text signal with the density of the financial-ratio signal. **Contribution 2** is a category-wise decomposition that splits each MD&A document into five standardised category corpora (business review, industry, risk, outlook, and core competitiveness) prior to embedding, so that the distillation layer operates on a structured set of semantic channels rather than a single averaged vector. To support reproducibility and further research, the full training pipeline, experiment scripts,

figure-generation code, and documentation are released as an open-source repository at https://github.com/jiangleping1985/lift-fdp.

2 Related Work

2.1 Classical FDP on Financial Ratios

Classical FDP research has developed a mature methodology based on numerical financial ratios. Altman's Z-score [1] introduced discriminant analysis over five balance-sheet ratios, and Ohlson's O-score [2] subsequently formulated the task as logistic regression. Subsequent work has expanded both the feature space and the hypothesis class; random forests, support vector machines, and gradient-boosted decision trees such as XGBoost [3] and LightGBM [4] have repeatedly been reported as strong single-modality baselines [5–7]. In the Chinese A-share context specifically, Geng et al. [13] reported competitive FDP performance using data mining on the same tabular feature family, and Bao et al. [14] combined unsupervised and supervised stages to further exploit the tabular modality. The input space of such methods is restricted to structured numerical data, and textual disclosures are not consumed; further gains beyond this paradigm therefore require extending the input modality.

2.2 Text-Augmented FDP in the Pre-LLM Era

A parallel line of research has sought to incorporate textual information into FDP pipelines. Two broad technical approaches can be identified. The first maps textual disclosures into a low-dimensional feature space via lexicon-based sentiment scoring (e.g., the Loughran–McDonald financial word list [8]), topic modelling with Latent Dirichlet Allocation [9], or bag-of-words counts. The second employs domain-adapted neural encoders such as FinBERT [10, 15]. Within FDP specifically, Cecchini et al. [16] linked textual features in corporate filings to fraud prediction; Mai et al. [17] combined deep text encoders with numerical features for bankruptcy prediction; Ding et al. [18] introduced a semantic forecasting approach driven by textual signals; and Wang et al. [19] proposed a regularised sparse Random Subspace ensemble that consumes annual-report text.

2.3 LLM-Era Approaches to Text-Augmented FDP

Work that incorporates modern LLMs into FDP and related financial prediction tasks can be grouped into three technical routes. The first consists of end-to-end deep classifiers that jointly ingest textual and numerical inputs. Che et al. [20] proposed an attentive and regularised deep-learning method on multimodal FDP data; Khatun et al. [21] combined textual narrative features with financial ratios within a multimodal neural network; Zhang et al. [22] applied transformer-based textual analysis to bankruptcy prediction. The second applies LLMs as stand-alone text classifiers or scoring models. Representative examples include Luo et al. [11] on BERT-based multitask Chinese financial text classification and the wider FinBERT family [10]. The third concerns adjacent predictive problems in ML-based finance, including stock-price prediction with sentiment and reinforcement learning [23], text-enhanced stock prediction

[24], GAN–LSTM–attention models [25], graph-neural-network approaches [26], and federated learning for small-enterprise credit risk [27, 28]. Two conceptual components inform the framework developed in this paper: stacked generalisation [29] and tree-based SHAP attribution [12]. Two-stage and multi-stage stacking architectures have recently been explored in the FDP and credit-scoring domain [30–32].

2.4 Positioning of LIFT

Against the background summarised above, LIFT is defined by four design choices. First, the LLM encoder is kept frozen and used as a feature extractor. Second, a small supervised MLP distils high-dimensional text embeddings into a compact, tabular-compatible representation. Third, the final classifier is a LightGBM stacker, which preserves the inductive bias that underpins the reported performance of gradient boosting on structured financial ratios and enables per-prediction SHAP attribution on the resulting 143-dimensional feature space. Fourth, the MD&A document is decomposed into five standardised category corpora prior to embedding. This combination is not jointly realised by any of the three routes reviewed above; together these choices delineate the specific design space occupied by LIFT.

3 Methods

3.1 Problem Formulation

Let $i \in \{1, \ldots, N\}$ index listed firms and let Y denote the fiscal year of the feature frame. Each sample consists of a 140-dimensional financial-ratio vector $x_{\text{fin}}^{(i,Y)} \in \mathbb{R}^{140}$ and an MD&A text block $x_{\text{MD\&A}}^{(i,Y)}$ drawn from the Y-year annual report. The prediction target $y^{(i,Y+1)} \in \{0, 1, 2\}$ encodes the firm's future ST state:

- $y = 0$ (*Healthy*): neither year $Y + 1$ nor year $Y + 2$ is ST;
- $y = 1$ (*Warning*): year $Y + 1$ is not ST but year $Y + 2$ is;
- $y = 2$ (*Distressed*): year $Y + 1$ is ST.

The Warning label captures the hardest early-warning regime, one year ahead of observable ST. The task is to learn $f : (x_{\text{fin}}, x_{\text{MD\&A}}) \mapsto \hat{y}$ with high macro-F1 and high distressed-class recall under a strict temporal train-test split.

Figure 1 overviews the LIFT pipeline from raw financial filings and annual reports through the two-stage architecture to the final prediction.

3.2 Financial Feature Engineering

Financial features are extracted via the AKShare `stock_financial_abstract` interface, which exposes 48 canonical ratios covering profitability, solvency, liquidity, operating efficiency, and per-share indicators. For each ratio, a year-over-year growth rate is additionally computed, and redundant derivations are removed to yield a final 140-dimensional feature vector per firm-year. Preprocessing applies median imputation for missing values, symmetric clipping to five standard deviations around the feature median to bound outliers, and z-score standardization.

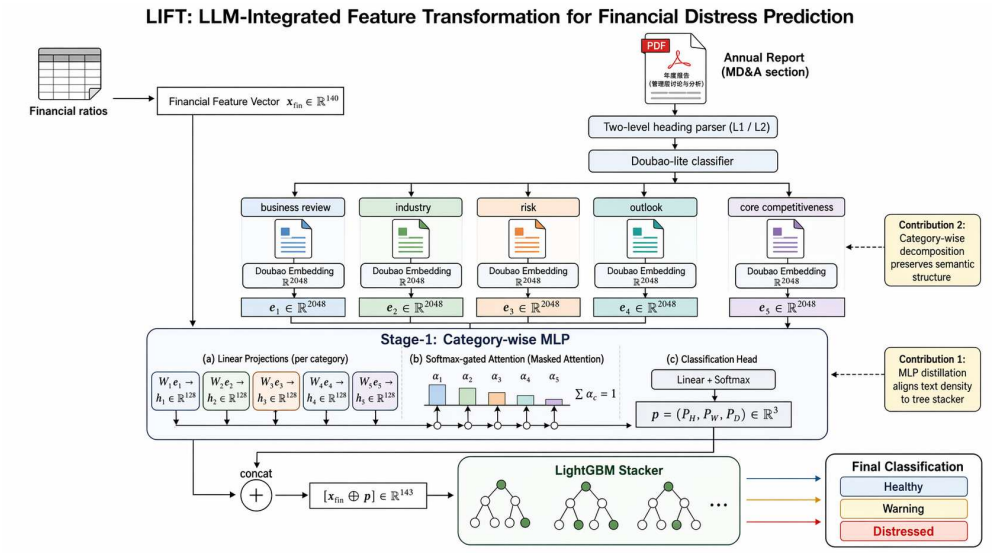


Fig. 1 Overview of the LIFT framework. Financial ratios (left) flow directly into the Stage-2 LightGBM stacker; MD&A text (right) is parsed into five standardised categories by a regex-based heading parser combined with a Doubao-lite classifier, and each category is then independently embedded to 2048 dimensions. A category-wise MLP with masked attention pooling distils the five vectors into three class probabilities that join the 140 financial ratios as Stage-2 inputs

3.3 Choice of Textual Source

Among the textual materials that can be associated with a listed firm—including annual and interim reports, earnings-call transcripts, analyst reports, regulatory announcements, and financial news—the present study uses the Management Discussion and Analysis (MD&A) section of the annual report as its sole textual source. This choice is guided by two considerations. First, MD&A is produced under a disclosure template that is strongly standardised by the China Securities Regulatory Commission [33], and it routinely contains the qualitative signals that motivate the textual modality in the first place: operational review, industry outlook, risk disclosure, forward-looking statements, and competitive positioning. Second, introducing several textual sources simultaneously would confound the contribution of each source with the effectiveness of the integration architecture itself. We therefore restrict the textual modality to a single well-structured source in this study, so as to hold all other factors fixed while the architecture itself is evaluated.

3.4 Category-Wise MD&A Embedding

Chinese annual reports follow a highly standardised MD&A template that facilitates deterministic splitting. We apply a two-level heading parser:

- *Level 1 (L1):* top-level subsections introduced by the Chinese ordinal numerals (Roman-numeral equivalents: I, II, III, IV, ..., X), each followed by a delimiter character;

- *Level 2 (L2):* sub-subsections introduced by parenthesised Chinese ordinal numerals, which are particularly important because risk disclosures are frequently embedded as L2 children under "Future Development Outlook".

The resulting chunks are labelled by a Doubao-lite classifier (Doubao-Seed-2.0-lite, endpoint `doubao-seed-2-0-lite-260215`) into six categories: `business_review`, `industry`, `risk`, `outlook`, `competitiveness`, and `other`. Chunks sharing a category are concatenated to form a per-firm-year category document $c_k^{(i,Y)}$, a variable-length token sequence, and the `other` chunks are discarded.

Each of the five kept category documents is then passed to a frozen production LLM embedding encoder f_{LLM}, implemented as the ByteDance Doubao embedding model (endpoint `doubao-embedding-vision-251215`, a multimodal model here invoked in text-only mode) accessed through the Volcano Engine production API [34]. Conceptually, the encoder implements the map

$$f_{\text{LLM}} : \mathcal{V}^* \longrightarrow \mathbb{R}^{2048}, \qquad c_c^{(i,Y)} \longmapsto e_c^{(i,Y)} \in \mathbb{R}^{2048}, \tag{1}$$

where $\mathcal{V}^*$ denotes the space of finite token sequences drawn from the model's vocabulary $\mathcal{V}$. Internally, the encoder tokenises the category document, maps tokens to contextualised representations through a stack of transformer blocks, and pools the resulting token states into a fixed-dimensional vector $e_c^{(i,Y)}$. We additionally define a binary mask $m_c^{(i,Y)} \in \{0,1\}$ indicating whether category c was non-empty for sample (i,Y). Empty categories are assigned a zero vector so that downstream batched computation remains well-defined.

3.5 Stage 1: Category-Wise Text Expert

Stage 1 is a lightweight MLP that maps the five category embeddings and their mask into three class probabilities. Per-category encoders produce hidden representations:

$$h_c^{(i,Y)} = \text{LN}\Big(\text{GELU}\Big(W_c e_c^{(i,Y)} + b_c\Big)\Big), \quad c = 1, \ldots, 5, \tag{2}$$

where $W_c \in \mathbb{R}^{128\times 2048}$, LN denotes layer normalisation, and dropout with probability $p = 0.3$ is applied to each category branch. A learned attention query $\mathbf{w}_a \in \mathbb{R}^{128}$ computes masked attention weights:

$$\alpha_c^{(i,Y)} = \frac{\exp\Big(\mathbf{w}_a^\top h_c^{(i,Y)}\Big) \cdot m_c^{(i,Y)}}{\sum_{c'} \exp\Big(\mathbf{w}_a^\top h_{c'}^{(i,Y)}\Big) \cdot m_{c'}^{(i,Y)}}. \tag{3}$$

The pooled representation $\bar{h}^{(i,Y)} = \sum_c \alpha_c^{(i,Y)} h_c^{(i,Y)}$ is passed through a classification head consisting of a 64-dim hidden layer with dropout and GELU followed by a linear projection to three logits. The final probabilities $p^{(i,Y)} = \text{softmax}(z^{(i,Y)}) \in \mathbb{R}^3$ represent the text expert's belief about the firm's class.

Training uses class-weighted cross-entropy with inverse-frequency weights $\{w_H, w_W, w_D\}$ computed from the training labels, the AdamW optimiser with learning rate 10^{-3} and weight decay 10^{-4}, batch size 64, and patience-6 early stopping on validation macro-F1. To generate unbiased training features for Stage 2, Stage 1 is trained with a five-fold cross-validation protocol on the training set and out-of-fold (OOF) predictions are stored. A final Stage-1 model is then trained on the full training set, with a 10% internal validation holdout for early stopping.

3.6 Stage 2: Stacking LightGBM

Stage 2 assembles the compact feature vector:

$$\tilde{x}^{(i,Y)} = \left[x_{\text{fin}}^{(i,Y)} \; ; \; p^{(i,Y)} \right] \in \mathbb{R}^{143}, \tag{4}$$

and trains a class-weighted LightGBM multi-class classifier with learning rate 0.05, 63 leaves, 400 boosting rounds, and early stopping with patience 30 on an internal validation split. Because the three text probabilities summarise the Stage-1 text expert's belief in a form compatible with LightGBM's tree-split inductive bias, Stage 2 can conditionally trust text signals: when financial ratios are ambiguous, the three text-derived features provide decisive evidence; when financial data are decisive, the tree structure naturally down-weights the text features.

3.7 Evaluation Protocol

Two complementary evaluation regimes are employed. First, a strict temporal split uses all firm-years with feature year $Y \leq 2021$ for training (3414 samples) and all firm-years with feature year $Y \in \{2022, 2023\}$ for testing (2668 samples). Second, a five-fold cross-validation with three seeds is applied to the full 6082-sample subset to obtain variance-aware estimates of generalisation performance. Reported metrics include macro-F1, accuracy, Matthews correlation coefficient (MCC), and per-class recall.

4 Results

4.1 Dataset

Financial features are drawn from AKShare covering 5505 Chinese A-share listed firms across 2015 to 2023, yielding 46,462 firm-year records overall. Annual-report PDFs are sourced from cninfo.com.cn (the Juchao Information Network) and parsed locally with PyMuPDF, producing 7445 MD&A extractions with valid heading structure across 1629 target firms (543 firms with ST-related history plus 1086 matched healthy controls) for fiscal years 2019 through 2023. Intersecting non-missing financial features with successfully embedded MD&A texts yields the final LIFT analysis subset of 6082 firm-year samples. The temporal training subset contains 3414 samples and the temporal test subset contains 2668 samples; Table 1 reports the per-year label distribution. ST labels are derived by combining CSMAR/AKShare name-change histories with consecutive-loss inference as prescribed by the exchange's ST rules.

Table 1 Label distribution of the 6082-sample LIFT subset across Healthy (H), Warning (W), and Distressed (D) classes

Split	H	W	D	Total
Train ($Y \leq 2021$)	2833	250	331	3414
Test ($Y \in \{2022, 2023\}$)	2047	83	538	2668
Full subset	4880	333	869	6082

4.2 Single-Modality Baselines

We first compare LIFT against nine widely used single-modality classifiers on the 6082-sample subset with identical preprocessing. Under five-fold cross-validation with five seeds using an `LGBMClassifier` configuration with 200 estimators, maximum depth 6, and learning rate 0.1, LightGBM achieves the highest single-modality macro-F1 of 0.7162, followed by XGBoost (0.6851) and Decision Tree (0.6495); LIFT exceeds even this strongest single-modality configuration by 7.8 CV-F1 points. In the main ablation study that follows (Table 2), we additionally rerun the financial-only LightGBM (V1) under the *same* Stage-2 training protocol used by LIFT (learning rate 0.05, 63 leaves, 400 rounds with early stopping), yielding temporal macro-F1 of 0.6363 for V1 and 0.6847 for V5.

4.3 Main Results: Ablation Matrix

Table 2 reports the main result organised around the two contributions of LIFT. The upper block isolates the two-stage stacking framework by testing alternative fusion strategies. The lower block isolates the category-wise embedding contribution by comparing a whole-MD&A pooled embedding against the category-wise embedding, holding the framework constant. Figure 2 visualises the comparison.

Table 2 Ablation matrix on the 6082-sample subset. Cross-validation (CV) macro-F1 is reported as mean ± standard deviation across five folds and three seeds. W-Rec denotes recall for the Warning class on the temporal test set. Bold indicates the best value in each column. All LightGBM-based variants share identical hyperparameters

Variant	Description	Dim	Temp. F1	CV F1	D-Rec	W-Rec
Contribution 1: two-stage stacking vs alternative frameworks						
V-MLP	End-to-end MLP on all feats.	—	0.5843	0.7087 ± 0.018	59.3%	83.1%
V3	LGB + raw category concat	10,380	0.6258	0.7084 ± 0.017	53.9%	—
V4	LGB + category + PCA-50	190	0.6337	0.7156 ± 0.015	57.8%	—
V1	Fin.-only LGB (baseline)	140	0.6363	0.7124 ± 0.018	58.2%	84.3%
Contribution 2: category-wise vs whole-pool (both use LGB + MLP distillation)						
V2b	LGB + MLP(whole-MD&A)	143	0.6640	0.7560 ± 0.016	62.1%	86.8%
V5	**LIFT = LGB + MLP(category)**	**143**	**0.6847**	$\mathbf{0.7939 \pm 0.016}$	**65.2%**	**94.0%**

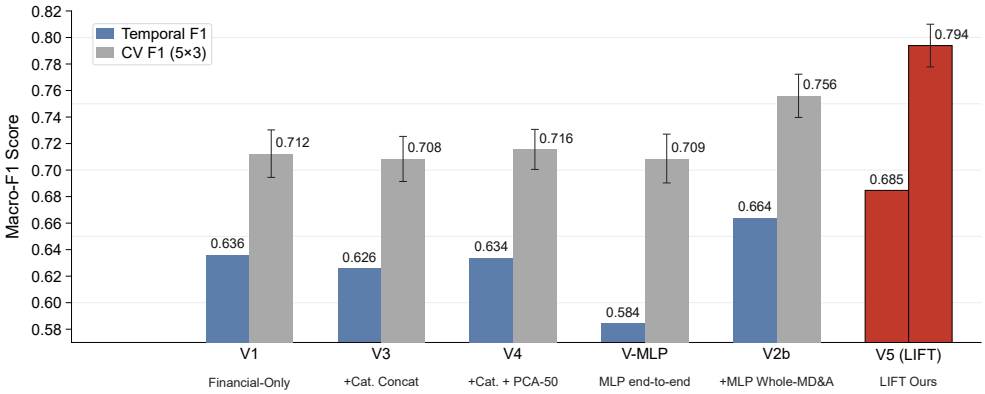


Fig. 2 Ablation comparison across six variants on the 6082-sample subset. LIFT (V5) achieves the highest temporal and cross-validation macro-F1 while using only 143 features, matching the dimensionality of the financial-only baseline. Error bars on the CV bars denote one standard deviation over the fifteen folds (five folds times three seeds)

Three observations stand out. First, replacing LightGBM with an end-to-end neural classifier (V-MLP) under-performs both the financial-only LightGBM baseline (V1) and the full LIFT framework (V5) by 0.4 and 8.5 CV-F1 points, respectively, confirming that LightGBM's tree inductive bias on the 140-dimensional financial feature space cannot be cheaply replaced by a neural network of comparable capacity. Second, naively concatenating the 10,380-dimensional raw category embedding to the financial features (V3) actually *reduces* temporal macro-F1 relative to the financial-only baseline by 1.1 points, demonstrating that high-dimensional sparse text inputs overwhelm LightGBM without first passing through supervised compression. Third, unsupervised compression via PCA (V4) recovers a small portion of the value, but remains 7.8 CV-F1 points below LIFT because PCA retains variance rather than discriminativeness. Holding the LightGBM + MLP framework constant, moving from whole-MD&A pooling (V2b) to category-wise embedding (V5) further adds 3.8 CV-F1 points and 7.2 percentage points of warning-class recall, isolating the contribution of semantic structure preservation.

4.4 Stage-1 Text Expert Analysis

We next examine the Stage-1 text expert in isolation. Evaluated on the five-fold OOF training predictions, Stage 1 achieves macro-F1 of 0.5650; on the temporal test set, Stage 1 alone achieves macro-F1 of 0.5100. These are both below the financial-only baseline, which is expected because text without financial context is a weaker signal for distress detection. The value of Stage 1 lies not in standalone prediction but in providing complementary signal that Stage 2 can combine with financial ratios.

Figure 3 visualises the mean attention weight each category receives, broken down by the true class of the firm. Three patterns emerge. The `competitiveness` category dominates attention across all classes, receiving between 0.58 and 0.80 of the total

attention mass, suggesting that discussions of core competitive advantages carry the strongest diagnostic signal. The `risk` category receives high attention for Healthy firms (0.40) but markedly lower attention for Warning and Distressed firms (0.10 and 0.11 respectively); we interpret this as an effect of boilerplate risk disclosures, where firms in better shape use more distinctive risk language while distressed firms rely on standard template text. The `outlook` category's attention increases monotonically with class severity, from 0.18 for Healthy firms to 0.24 for Distressed firms, consistent with the intuition that forward-looking narrative carries additional signal near distress.

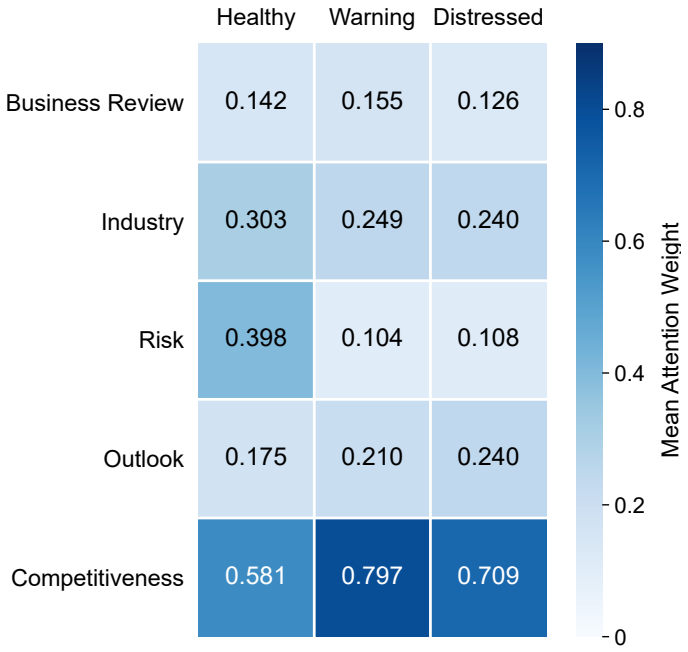


Fig. 3 Mean Stage-1 MLP attention weights across five MD&A categories, stratified by true class. Darker cells indicate higher attention. Competitiveness dominates attention across all classes, while the risk category shows a strong Healthy-versus-others split. Cell values are conditional on the corresponding category being non-empty for a given firm-year; therefore columns are not expected to sum to one across a single sample

Figure 4 further visualises the effect of MLP distillation on embedding separability. The left panel displays a two-dimensional t-SNE projection of raw whole-MD&A mean-pooled embeddings, where classes overlap heavily. The right panel displays a t-SNE projection of the Stage-1 MLP pooled hidden representation, where visibly separated class clusters emerge, illustrating how supervised distillation reshapes an initially overlapping embedding space into one where class structure is more apparent.

We caution that t-SNE is a visualisation technique that can exaggerate local cluster structure, so this figure should be read as qualitative support rather than quantitative proof; the numeric ablation in Table 2 is the primary evidence.

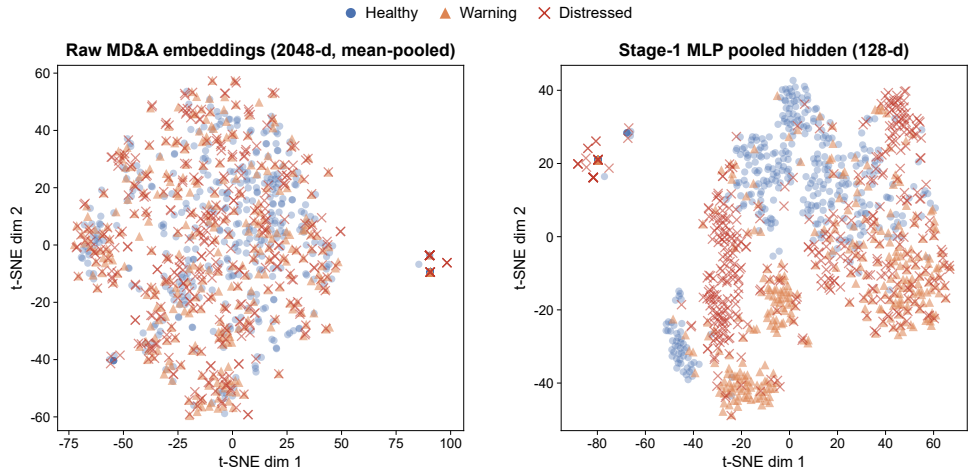


Fig. 4 Effect of Stage-1 MLP distillation on t-SNE class separability. Left panel: raw whole-MD&A mean-pooled 2048-dimensional embeddings show heavy class overlap. Right panel: Stage-1 pooled hidden 128-dimensional representations show visibly separated class clusters. Both panels use a balanced subsample of 1133 firm-years (400 Healthy, 333 Warning, 400 Distressed). t-SNE is a non-linear dimensionality reduction that can exaggerate local cluster structure; the figure should be read as qualitative illustration, not as quantitative proof of separability

4.5 Feature Importance

To quantify how LIFT's Stage-2 classifier weights the three text-probability features against the 140 financial ratios, we compute SHAP values [12] via the TreeExplainer variant on 500 randomly sampled test firms. Figure 5 reports the top 15 features by mean absolute SHAP value. All three text-derived features rank in the top eight: `text_P_H` is the third most important feature overall, while `text_P_D` and `text_P_W` rank sixth and eighth respectively. The largest-impact text feature `text_P_H` has its strongest per-class impact on the Healthy class (0.185), indicating that LIFT relies on the text expert primarily to *confirm* genuinely healthy firms when financial ratios alone are ambiguous.

4.6 Confusion-Matrix Comparison

Figure 6 compares the 3-by-3 confusion matrices of the financial-only LightGBM baseline and the full LIFT model on the 2668-sample temporal test set. LIFT increases Warning-class recall from $70/83 = 84.3\%$ to $78/83 = 94.0\%$ (eight more correctly identified Warning firms) and Distressed-class recall from $313/538 = 58.2\%$ to $351/538 =$

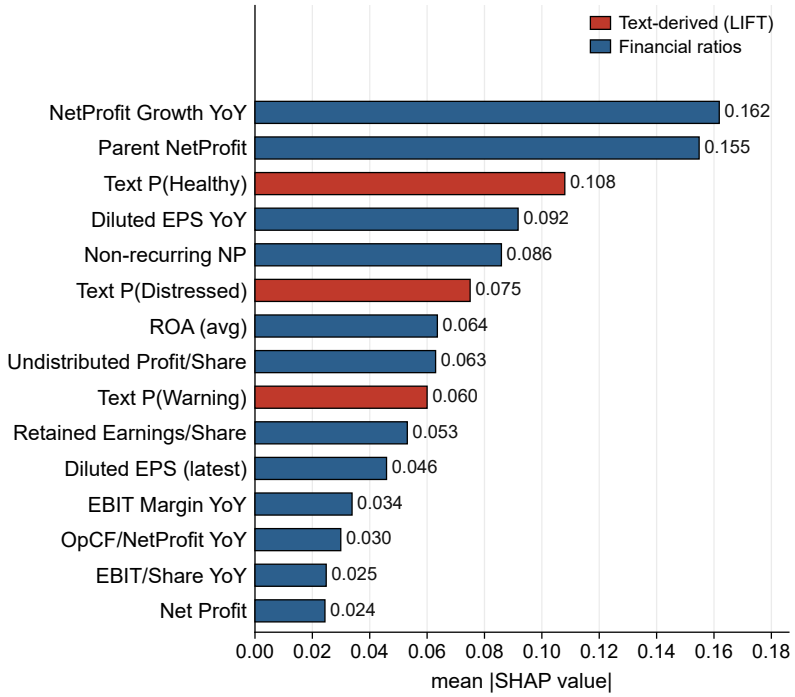


Fig. 5 Top 15 features by mean absolute SHAP value on 500 test firms. Text-derived features (red) are `text_P_H`, `text_P_D`, and `text_P_W`, ranking 3rd, 6th, and 8th respectively out of 143 features. Financial ratios are shown in blue

65.2% (38 more correctly identified Distressed firms), while the Healthy class improves marginally from 1794/2047 to 1800/2047. The combined improvement in minority-class recall without degradation of Healthy precision is the source of LIFT's overall MCC gain from 0.5580 to 0.6380 on the temporal split.

4.7 Case Studies

To provide qualitative evidence for the hypothesis that text leads numbers, we examine three firms where LIFT correctly predicted Distressed but the financial-only baseline predicted Healthy. Figure 7 displays the SHAP attribution breakdowns for the top ten features contributing to each firm's Distressed prediction.

Case A (firm 002872, fiscal year 2022, pharmaceutical sector) illustrates the canonical pattern. Several financial ratios, most prominently year-over-year net-profit growth, pushed the prediction toward Healthy, and the financial-only baseline accordingly predicted Healthy. However, the MD&A risk section emphasised accelerating regulatory pressure from volume-based procurement and medical-insurance cost controls, language that Stage 1 interpreted as highly discriminative for distress. LIFT's text-derived features contributed a positive net effect on the Distressed class, overriding the misleading financial signal and recovering the correct prediction. Case B

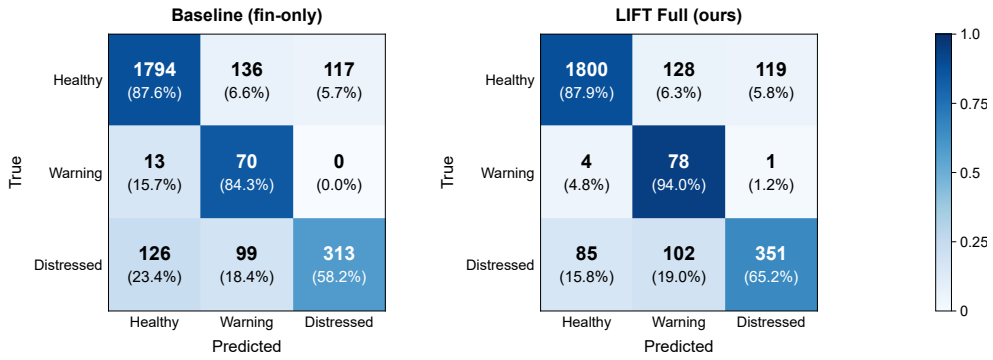


Fig. 6 Row-normalised confusion matrices on the 2668-sample temporal test set. Left: financial-only LightGBM baseline. Right: LIFT Full. LIFT catches 38 more Distressed firms and 8 more Warning firms while marginally improving Healthy-class performance. Row-normalised percentages are shown under the raw counts in each cell

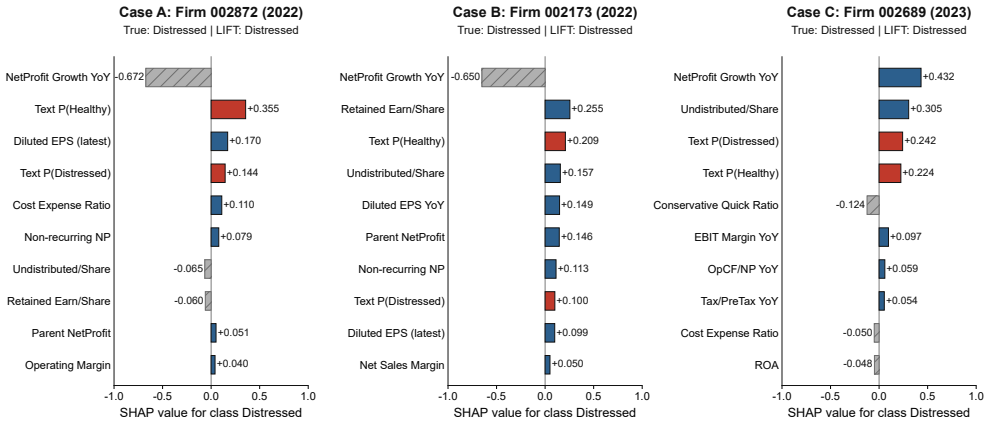


Fig. 7 Per-firm SHAP attribution for three case-study firms. Bars with positive values push the prediction toward Distressed; negative values push away. Red bars denote LIFT's text-derived features (`text_P_H`, `text_P_D`, `text_P_W`); blue bars denote positively-contributing financial features; gray hatched bars denote negatively-contributing features (which in these cases push the prediction toward Healthy but were overridden by the aggregate)

(firm 002173, fiscal year 2022, private hospital sector) exhibits a similar pattern; the industry-section narrative quantified a 5.2% year-over-year decline in national hospital consultations due to pandemic aftermath, a signal that materialised in revenue lines only with a lag. Case C (firm 002689, fiscal year 2023, elevator manufacturing) flags downstream real-estate exposure that the financial ratios had yet to reflect. Across the entire temporal test set, approximately 140 firms exhibit this text-leads-numbers

pattern, compared with 88 firms where the text introduces a false positive, yielding a net gain of 52 correct predictions concentrated in the high-value Distressed class.

5 Discussion

5.1 Why Category Structure Helps

The competitiveness, risk, and outlook categories are conceptually orthogonal dimensions of firm narrative. A whole-MD&A mean pool averages them into a generic "tone" vector in which distress-specific language is diluted by boilerplate business-description content. Category-wise embedding preserves these channels, and the Stage-1 MLP learns per-category attention weights that allocate representational capacity to the diagnostic categories.

5.2 Why Supervised Distillation Beats PCA

PCA maximises variance preservation regardless of label relevance, allowing dimensions dominated by industry-specific vocabulary or reporting style to survive while distress-relevant semantic directions are attenuated. In contrast, a supervised MLP trained with class-weighted cross-entropy explicitly optimises for discriminativeness, yielding three highly informative probability features that LightGBM can treat as calibrated signals from a text expert.

5.3 Why Two-Stage Stacking Beats Naive Concatenation

LightGBM's feature selection implicitly favours features with high information gain. With 10,380 raw text dimensions competing against 140 financial ratios, the effective per-feature signal is diluted, and the booster may overfit noisy text dimensions before converging on the relevant ones. Three-probability distillation gives LightGBM exactly three highly informative text features to evaluate against the financial baselines, aligning the feature density between modalities.

5.4 Prediction Horizon Sensitivity

To probe whether the text modality's value extends to longer prediction windows, we re-evaluated both the financial-only baseline and LIFT under a T-2 horizon where same-year features are used to predict firm status two years ahead rather than one. Under this harder horizon, baseline CV macro-F1 drops from 0.7124 to 0.6437 and LIFT from 0.7939 to 0.6900. LIFT retains an advantage of approximately 4.6 macro-F1 points, but the gain is substantially smaller than at T-1 (see Figure 8). We interpret this as a reflection of the regulatory mandate that MD&A disclosures concentrate on the reporting period rather than on multi-year forecasting; the text modality is thus most useful for near-term early warning rather than long-horizon forecasting.

5.5 Limitations

Three limitations are worth acknowledging. First, text coverage is limited to fiscal years 2019 to 2023 because of PDF availability constraints; extending to earlier

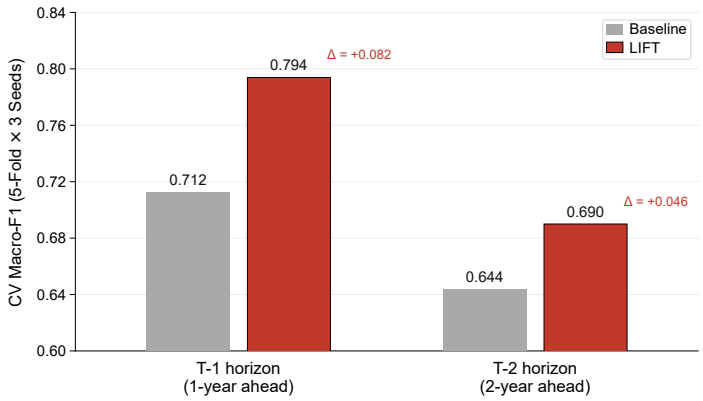


Fig. 8 Cross-validation macro-F1 for baseline and LIFT under T-1 and T-2 prediction horizons. LIFT retains a positive gain at T-2 (+0.046) but substantially smaller than at T-1 (+0.082), consistent with MD&A disclosures emphasising the current reporting period

periods would require access to older annual-report filings and may involve lower text-extraction quality. Second, the LLM embedding endpoint used in this study is a production multimodal model configured to accept text-only input. Dedicated text embedding models such as BGE-zh may yield different trade-offs and are a natural avenue for extension. Third, the category taxonomy and heading-based splitter are tailored to the Chinese A-share MD&A template; transferring LIFT to US 10-K filings or other regulatory regimes would require adjusting the category schema.

6 Conclusions

This paper introduced LIFT, a two-stage framework for financial distress prediction that integrates category-wise embeddings of MD&A narratives with financial ratios. Category-wise embedding preserves the semantic structure of standardised MD&A disclosures; a small supervised MLP compresses the high-dimensional text vectors into three class probabilities; and LightGBM stacking produces the final prediction on a compact 143-dimensional feature vector. On 6082 Chinese A-share firm-year samples, LIFT improves cross-validation macro-F1 by 8.2 percentage points over a financial-only LightGBM baseline and lifts warning-class recall from 84.3% to 94.0% together with distressed-class recall from 58.2% to 65.2%. SHAP attribution and a case-study analysis confirm that text-derived features contribute complementary, non-redundant signal. Directions for future work include enriching the text modality with real-time news and analyst commentary, and investigating hybrid alternatives that can partially recover the benefit of the text expert at longer prediction horizons.

Acknowledgements. The authors thank the open-source maintainers of AKShare and PyMuPDF, whose tools enabled the reproducible data pipeline described in this paper.

Abbreviations. AUC: Area Under the Receiver Operating Characteristic Curve; CSRC: China Securities Regulatory Commission; CV: Cross-Validation; FDP: Financial Distress Prediction; LGB: LightGBM; LLM: Large Language Model; MCC: Matthews Correlation Coefficient; MD&A: Management Discussion and Analysis; MLP: Multilayer Perceptron; OOF: Out-of-Fold; PCA: Principal Component Analysis; SHAP: SHapley Additive exPlanations; ST: Special Treatment.

Declarations

- **Funding.** This research received no external funding.

- **Conflict of interest.** The authors declare no competing interests.
- **Ethics approval and consent to participate.** Not applicable. This study uses only publicly available corporate disclosures and aggregated financial data; it does not involve human subjects or animal studies.
- **Consent for publication.** Not applicable.
- **Data availability.** All datasets used in this study are derived from publicly available open sources and require no special access. Financial-ratio features were retrieved through the AKShare open-source Python interface; annual-report narratives in PDF form were obtained from the Juchao Information Network, the official disclosure platform of the China Securities Regulatory Commission for A-share listed firms, and parsed locally with PyMuPDF; ST (Special Treatment) labels and listed-firm name-change histories were cross-referenced with CSMAR and AKShare. The complete reproduction-ready analysis subset (6,082 Chinese A-share firm-year samples spanning 2019–2023) can be reconstructed end-to-end by running the open-source pipeline released at https://github.com/jiangleping1985/lift-fdp under an MIT license, which contains all data-collection scripts, preprocessing code, the training pipeline, and figure-generation code in English. Trained model weights and processed dataset splits will be deposited in the same repository upon acceptance. No datasets generated or analysed in this study involve human subjects, personal data, or restricted-access materials.
- **Materials availability.** Not applicable.
- **Code availability.** The complete source code, training pipeline, figure-generation scripts, and documentation are openly available at https://github.com/jiangleping1985/lift-fdp under an MIT license.
- **Author contribution.** Conceptualisation, L.J.; methodology, L.J.; software, L.J. and H.H.; validation, L.J., H.H. and Y.L.; formal analysis, L.J. and Y.L.; investigation, L.J.; resources, L.J. and H.H.; data curation, H.H.; writing—original draft preparation, L.J.; writing—review and editing, L.J., H.H. and Y.L.; visualisation, L.J. and Y.L.; supervision, L.J.; project administration, L.J. All authors have read and agreed to the published version of the manuscript.

References

[1] Altman, E.I.: Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. The Journal of Finance **23**(4), 589–609 (1968) https://doi.org/10.1111/j.1540-6261.1968.tb00843.x

[2] Ohlson, J.A.: Financial ratios and the probabilistic prediction of bankruptcy. Journal of Accounting Research **18**(1), 109–131 (1980) https://doi.org/10.2307/2490395

[3] Chen, T., Guestrin, C.: XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), San Francisco, CA, USA, pp. 785–794 (2016). https://doi.org/10.1145/2939672.2939785

[4] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.-Y.: LightGBM: A highly efficient gradient boosting decision tree. In: Advances in Neural Information Processing Systems 30 (NeurIPS 2017), Long Beach, CA, USA, pp. 3146–3154 (2017)

[5] Barboza, F., Kimura, H., Altman, E.: Machine learning models and bankruptcy prediction. Expert Systems with Applications **83**, 405–417 (2017) https://doi.org/10.1016/j.eswa.2017.04.006

[6] Le, T.T., Gerstner, P., Bose, R.: Deep learning in financial distress prediction: A comprehensive review. Engineering Applications of Artificial Intelligence **114**, 105112 (2022) https://doi.org/10.1016/j.engappai.2022.105112

[7] Tsai, C.-F., Hsu, Y.-F., Yen, D.C.: A comparative study of classifier ensembles for bankruptcy prediction. Applied Soft Computing **24**, 977–984 (2014) https://doi.org/10.1016/j.asoc.2014.08.047

[8] Loughran, T., McDonald, B.: When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. The Journal of Finance **66**(1), 35–65 (2011) https://doi.org/10.1111/j.1540-6261.2010.01625.x

[9] Dyer, T., Lang, M., Stice-Lawrence, L.: The evolution of 10-K textual disclosure: evidence from Latent Dirichlet Allocation. Journal of Accounting and Economics **64**(2–3), 221–245 (2017) https://doi.org/10.1016/j.jacceco.2017.07.002

[10] Yang, Y., Uy, M.C.S., Huang, A.: FinBERT: A pretrained language model for financial communications. arXiv preprint **arXiv:2006.08097** (2020) https://doi.org/10.48550/arXiv.2006.08097

[11] Luo, R., Zhang, L., Zhou, D.: BERT-based multi-task learning for Chinese financial text classification. Neurocomputing **520**, 276–287 (2023) https://doi.org/10.1016/j.neucom.2022.11.085

[12] Lundberg, S.M., Lee, S.-I.: A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems 30 (NeurIPS 2017), Long Beach, CA, USA, pp. 4765–4774 (2017)

[13] Geng, R., Bose, I., Chen, X.: Prediction of financial distress: An empirical study

of listed Chinese companies using data mining. European Journal of Operational Research **241**(1), 236–247 (2015) https://doi.org/10.1016/j.ejor.2014.08.016

[14] Bao, W., Lianju, N., Yue, K.: Integration of unsupervised and supervised machine learning algorithms for credit risk assessment. Expert Systems with Applications **128**, 301–315 (2019) https://doi.org/10.1016/j.eswa.2019.02.033

[15] Araci, D.: FinBERT: Financial sentiment analysis with pre-trained language models. arXiv preprint **arXiv:1908.10063** (2019) https://doi.org/10.48550/arXiv.1908.10063

[16] Cecchini, M., Aytug, H., Koehler, G.J., Pathak, P.: Making words work: Using financial text as a predictor of financial events. Decision Support Systems **50**(1), 164–175 (2010) https://doi.org/10.1016/j.dss.2010.07.012

[17] Mai, F., Tian, S., Lee, C., Ma, L.: Deep learning models for bankruptcy prediction using textual disclosures. European Journal of Operational Research **274**(2), 743–758 (2019) https://doi.org/10.1016/j.ejor.2018.10.024

[18] Ding, A.S., Tian, S., Wu, Y., Hou, W.: A new semantic forecasting approach to financial distress prediction using textual information. International Review of Economics and Finance **65**, 170–179 (2020) https://doi.org/10.1016/j.iref.2019.10.010

[19] Wang, G., Ma, J., Chen, G., Yang, Y.: Financial distress prediction: Regularized sparse-based Random Subspace with ER aggregation rule incorporating textual disclosures from annual reports. Information Sciences **536**, 1–25 (2020) https://doi.org/10.1016/j.ins.2020.05.040

[20] Che, W., Wang, Z., Jiang, C., Abedin, M.Z.: Predicting financial distress using multimodal data: An attentive and regularized deep learning method. Information Processing and Management **61**(4), 103703 (2024) https://doi.org/10.1016/j.ipm.2024.103703

[21] Khatun, M.A., Zoha, A., Yazdan Pour, A.: Predicting financial distress with textual narrative features using multimodal neural networks. International Journal of Accounting Information Systems **48**, 100603 (2023) https://doi.org/10.1016/j.accinf.2022.100603

[22] Zhang, Y., Liu, J., Wang, H., Chen, Q.: Corporate bankruptcy prediction with transformer-based textual analysis. Expert Systems with Applications **229**, 120623 (2023) https://doi.org/10.1016/j.eswa.2023.120623

[23] Suhail, K.M.A., Sankar, S., Kumar, A.S., Nestor, T., Soliman, N.F., Algarni, A.D., El-Shafai, W., Abd El-Samie, F.E.: Stock market trading based on market sentiments and reinforcement learning. Computers, Materials & Continua **70**(1), 935–950 (2022) https://doi.org/10.32604/cmc.2022.017069

[24] Huang, Y., Wang, Y., Zhang, Y., Wu, Q., Sun, L.: Text-enhanced stock prediction by combining financial news with stock price movement. IEEE Access **8**, 215037–215050 (2020) https://doi.org/10.1109/ACCESS.2020.3041113

[25] Li, P., Wei, Y., Yin, L.: Research on stock price prediction method based on the GAN-LSTM-attention model. Computers, Materials & Continua **82**(1), 609–625 (2025) https://doi.org/10.32604/cmc.2024.056651

[26] Chen, J., Wen, Y., Nanehkaran, Y.A., Suzauddola, M., Chen, W., Zhang, D.: Machine learning techniques for stock price prediction and graph neural network based models: A comprehensive review. Neural Computing and Applications **35**(1), 463–487 (2023) https://doi.org/10.1007/s00521-022-07707-2

[27] Xu, Z., Cheng, J., Cheng, L., Xu, X., Bilal, M.: MSEs credit risk assessment model based on federated learning and feature selection. Computers, Materials & Continua **75**(3), 5573–5595 (2023) https://doi.org/10.32604/cmc.2023.036937

[28] Gao, Y., Li, Y., Wang, Y.: Machine learning-based credit risk assessment with federated learning: A survey. ACM Computing Surveys **55**(10), 201 (2022)

[29] Wolpert, D.H.: Stacked generalization. Neural Networks **5**(2), 241–259 (1992) https://doi.org/10.1016/S0893-6080(05)80023-1

[30] Li, P., Wang, S., Zhao, J.: Two-stage stacking heterogeneous ensemble learning for bankruptcy prediction. Journal of the Operational Research Society **72**(11), 2528–2539 (2021) https://doi.org/10.1080/01605682.2020.1796535

[31] Kumar, A., Narayanan, S., Subramanian, K.: Two-stage hybrid machine learning for financial distress prediction with attention mechanism. Expert Systems with Applications **202**, 117284 (2022) https://doi.org/10.1016/j.eswa.2022.117284

[32] Shen, F., Zhao, X., Kou, G.: Three-stage reject inference learning framework for credit scoring using unsupervised transfer learning and three-way decision theory. Decision Support Systems **159**, 113815 (2022) https://doi.org/10.1016/j.dss.2022.113815

[33] China Securities Regulatory Commission: Guidelines no. 2 for the content and format of information disclosure by listed companies: Contents and formats of annual reports. Technical report, China Securities Regulatory Commission, Beijing, China (2021)

[34] ByteDance Volcano Engine: Doubao Embedding API Documentation. https://www.volcengine.com/docs/82379. Accessed: 20 April 2026 (2025)