

Corporate Default Prediction Through Text Mining: Integrating Event, Sentiment, and Network Analyses

Yao Xuan

National University of Singapore
yaoxuan@nus.edu.sg

Tan Tianhui

National University of Singapore
tant@nus.edu.sg

Su Yating

National University of Singapore
yating.su@u.nus.edu

Huang Ke-wei

National University of Singapore
dishkw@nus.edu.sg

Abstract

The importance of textual information in corporate credit risk management is increasingly recognized. While most studies focus on the direct analysis for assessing corporate credit risk, they often overlook the potential impact of inter-company relationships on the likelihood of default. This study, focusing on both intrinsic information about companies themselves and relational information within company networks, explores the potential of advanced text-mining techniques for predicting corporate defaults. We integrate default label tagging, credit sentiment analysis, and relation analysis via co-mention networks using public news on US-listed oil companies between 2014 and 2016. We aim to demonstrate how these advanced text-derived features enhance default prediction during industry upheaval. Our findings reveal that credit sentiment emerges as a crucial predictor of default, alongside network degree and transitivity. High-risk labelled companies are more likely to default than others. Moreover, exposure to media, regardless of being positive or negative, may increase the likelihood of both default and other corporate exits, primarily mergers and acquisitions. This study emphasizes the transformative impact of text analysis on traditional credit risk assessment practices and underscores the value of relational information between companies for default prediction.

Keywords: Text Analysis, Credit Risk, Co-mention Network, Sentiment Analysis, Large Language Model

1. Introduction

Corporate default holds significant importance in economics due to its potential to destabilize

entire industries and economies. Defaults are not isolated events; they often co-occur rather than being independently and identically distributed across firms. In particular, economic shocks can trigger a cascade of defaults, creating a domino effect. This is especially concerning in highly leveraged industries such as oil and real estate, where the failure of a major firm can lead to widespread financial distress. The contagion effect of these defaults can extend beyond the immediate sector, elevating the risk of systemic crises. As defaults accumulate, they may erode investor confidence, tighten credit conditions, and destabilize the broader financial system. Banks frequently underestimate these interconnected risks, exacerbating the problem. Accurately predicting subsequent defaults, therefore, becomes critically important. Our study explores how text-derived features can improve the prediction of defaults during periods of industry upheaval, providing a more nuanced understanding of risk and aiding in the development of more robust financial models.

Securing high-quality, timely, and detailed data to enhance the accuracy of default predictions poses a significant challenge in financial analysis. Most existing studies have focused on 10-K reports, stock prices, or financial accounting data, which, while valuable, have limitations in predicting and quantifying contagion effects. 10-K reports often lag behind real-time conditions, and market prices lack contextual details, while financial accounting data may not capture dynamic market interactions. Alternatively, public news articles offer a more immediate and dynamic perspective, providing intrinsic information such as credit profile analysis, insights from industry experts and financial analysts, and examinations of relational dynamics between companies and their partners or competitors. This information is crucial

for understanding how financial distress in one firm can impact others, highlighting the contagion effects of defaults. Despite the challenges of managing and deriving effective features from the vast amount of news data, our study aims to explore these alternative data sources and develop promising features to improve default prediction, focusing on the contagion effects of defaults.

Notably, a company's credit profile is intrinsically linked to or affected by its counterparties. A major default can propagate through the relationship network, straining connected entities' credit quality. Thereby, overlooking relationship cues is likely to yield an incomplete credit risk view. Despite its potential to reveal intricate financial relationships, this aspect remains underexplored. News authors and analysts can implicitly reveal these exposures by discussing related companies within the same article (Netzer et al., 2012). These relationships can indicate joint financial stressors resulting from market shocks, credit rating downgrades, or negative outlooks. Alternatively, they may reflect contrasting situations where one company thrives while another faces distress. Additionally, such co-mentions may also reflect alliances between companies, where they assist each other through financial difficulties. However, these implicated interconnections are challenging to capture directly through financial ratios or sentiment analysis. Hence, in addition to analyzing features related to the focal company's activities, which could be explored using sentiment analysis, we also aim to investigate the predictive power of relationship features between companies in anticipating defaults. In summary, our study incorporates features that include event analysis (or topic modelling), sentiment analysis, and relation analysis in text mining (Chen et al., 2021).

However, automatically extracting that relation information is a non-trivial task because of the complicated business relations among firms and the variety of news writing styles. While traditional NLP tools, such as NER, are useful for extracting company names from articles, they are not able to detect the connections among companies, especially under a specific scenario, such as a default warning. On the other hand, advancements in Large Language Models (LLMs) have made customized information extraction tasks more feasible. Therefore, our study aims to develop different types of text features from public news, incorporating intrinsic company information and inter-company relationships, to predict corporate defaults over multiple terms.

In this paper, we focus on the U.S. oil sector crisis from 2014 to 2016, when the industry experienced a

significant downturn due to an oil oversupply from the domestic shale revolution, coupled with a sharp decline in global oil prices and economic slowdowns in major oil-consuming regions. Companies struggled to generate sufficient cash flow, as oil prices remained depressed, triggering a wave of bankruptcies and defaults, particularly among smaller, highly leveraged firms unable to withstand the prolonged period of low prices. Thus, this period provides an ideal context for our study.

We collect news articles from reputable financial media (such as The Wall Street Journal and Reuters) reporting on publicly listed firms in the US oil industry from 2014 to 2016. We calculate company-specific credit-weighted sentiment scores to capture the focal company's intrinsic information. Additionally, we use GPT-4 to identify co-mentioned companies — those either in default, at high risk of defaulting, or default irrelevant — and construct a comprehensive co-mention network to represent company interconnections. By using traditional financial structured variables as control variables, we aim to assess the incremental predictive power of these text-derived features in forecasting corporate defaults during industry upheaval.

This study contributes in several ways. First, we use three types of text analysis (event label tagging, sentiment analysis, and relation analysis) to provide diverse insights into a company's credit profile from public news. This innovative approach helps us distinguish a company's financial health from its susceptibility to industry-wide contagion. Second, we demonstrate the approach's usefulness across various sectors, from those affected by single factors to those influenced by complex forces. This highlights its adaptability across industries with different levels of complexity. Thirdly, we explore the approach's applicability in developing economies, showing its relevance in diverse economic contexts. Fourthly, we expand the language scope from English to Chinese, making it more cross-culturally applicable and aiding global credit risk assessment insights. Additionally, we uncover how textual information from company news impacts other events, like mergers and acquisitions (M&A), providing insights into corporate strategy implications. These findings offer valuable insights into the broader implications of text analysis in financial markets.

The remainder of this paper is organized as follows. Section 2 reviews text analysis in credit risk assessment. Section 3 outlines data collection. Section 4 describes the methodology. Section 5 presents empirical findings. Section 6 discusses the various scenarios across economies, sectors and languages. Finally,

Section 7 concludes with future research directions.

2. Literature review

Text mining involves extracting valuable insights from vast amounts of unstructured text data, offering the potential to automate news processing and generate market signals, which has garnered significant attention. The approaches employed can be classified into three types: event analysis, sentiment analysis, and relation analysis (Chen et al., 2021). The main task of event analysis is to categorize the events/topics reported in the news. Several studies have explored the task of event tagging and extraction. For example, Cecchini et al. (2010) detected fraud and bankruptcy events from financial statements and annual reports using a lexicon approach. Dong et al. (2018) detected corporate fraud, leaked information, and rumours from social media. Huang et al. (2018) used topical analysis methods and illustrated the incremental information provided by analysts rather than conference calls. J. Duan and Yao (2022) utilise a topic modelling method to recognize credit-related articles. In this study, we would like to use the prompt engineer of LLMs to recognize corporate default events in three statuses: already defaulted, high risk of default, and not related to default.

For the application of text-mining methods in financial market analysis, a more popular approach is sentiment analysis. The impacts of media sentiment on financial analysis for such as stock return, volatility and credit risk have been studied; for example, Groß-Klußmann et al. (2019) and Sun et al. (2020), Xing et al. (2019), Roeder et al. (2020), Tsai et al. (2016), Dunham and Garcia (2021) and Tan and Phan (2018). To conduct sentiment analysis, researchers such as Li (2010) use lexicons in the early stage, and then turn to language models after BERT was launched (J. Duan and Yao, 2022). In this study, we calculate credit-perspective corporate sentiment scores, derived from news data, to capture company-specific information distinct from inter-company relationships.

Different from event and sentiment analysis which focuses on the direct relationship between events and firm performance, relations analysis focuses on the identification of firm connections such as competitors, partners, supply chains, etc. However, the relations analysis in text mining is relatively primitive. A major method is co-mention analysis, in which two firms are considered relevant if they appear in the same news article. The features extracted from the co-mention network such as firm centrality have shown potential use for the prediction of stock return, firm equity value, profit and volatility (Chen et al., 2021; Creamer et al.,

2013; Jin et al., 2012). Studies found that a company is more likely to co-occur with its competitors/partners in news (Netzer et al., 2012). Namely, competitors are more likely in the same sector, which are concurrently exposed to identical macroeconomic shocks. This study focuses on companies' default risk under the same sector-level shock.

3. Data collection

Our study focuses on the U.S. oil sector crisis from 2014 to 2016, which comprises 361 US-listed companies, classified according to the Bloomberg Industry Classification Standard (BICS) 2020.

For news coverage, we extracted 3,357 articles reporting the US-listed oil companies from the Factiva database. These articles primarily originate from Reuters and The Wall Street Journal, two globally recognized financial news sources. After manually verifying the identification of company default events in 150 sample articles, we found that GPT-4o yielded the most accurate results. Consequently, GPT-4o was utilized in this study. For articles mentioning a default event, each was categorized by GPT-4o into three statuses: has defaulted, high risk of defaulting, or not default-related (e.g., this article is not about the default of the focal firm).

Credit events (defaults and other exits) and a comprehensive set of structured financial variables were sourced from Bloomberg and Refinitiv. These variables encompass both macro-financial factors and individual firm attributes. In total, our dataset comprises 10,993 firm-month time series for the U.S. oil sector offering a granular, longitudinal view of each company's financial health.

4. Methodology and variables

We develop a three-step approach: first, tagging firm status from news articles using LLM prompt engineering; second, conducting sentiment analysis and constructing co-mention networks; and third, developing default prediction models.

Default Event Label Tagging. We employ the GPT-4o language model to extract and categorize the corporate entities mentioned in each article. Specifically, the model was prompted to identify company entities and assign one of the following credit labels to each: 'High risk of defaulting', 'Already defaulted', or 'Not related to default'. Next, we standardize the corporate names (e.g., "Apple Inc." vs. "Apple") by referring to a database of 26,845 publicly listed firms in the United States and China to ensure

consistency and accuracy in entity resolution. After mapping, there are 361 US-listed oil companies in total and 268 of them have been covered by media news. An example prompt is shown in Table 1.

Table 1. Prompt for Default Event Analysis

<p>#BACKGROUND# In financial markets, defaults are significant events that can indicate financial distress or instability. The definition of default event in this study includes:</p> <ol style="list-style-type: none"> 1. Bankruptcy filing, receivership, administration, liquidation, or any legal impasse affecting timely interest and/or principal payments. 2. Missed or delayed payment of interest and/or principal, excluding those within a grace period. 3. Debt restructuring or distressed exchange resulting in a reduced financial obligation (e.g., debt-to-equity conversion, lower coupon, lower principal, lower seniority, or longer maturity). <p>#TASK# Your task is to analyse a news article titled {title} and identify companies mentioned within the article. For each company, strictly follow the provided format and provide only one tag: 'Indications of potential default.', 'Has already defaulted.', or 'Not related to default'. Do not provide additional information beyond this format.</p> <p>#RESPONSE FORMAT# Here are the companies mentioned in the article with indications of potential default: - Company A: High risk of defaulting. - Company B: Already defaulted. - Company C: Not related to default.</p> <p>#NEWS ARTICLE# {content}</p>
--

Sentiment Calculation. Following the methodology of J. Duan and Yao (2022), we construct company-specific, credit-oriented sentiment time series. We first apply source-LDA Wood, 2016 to identify credit-related topics and compute each article’s credit weight (0–1). In parallel, we fine-tune a RoBERTa model using the hyper-parameters in J. Duan and Yao (2022) to classify the article’s sentiment toward the focal company on a scale from –2 (extremely negative) to +2 (extremely positive). We then weight each sentiment score by its credit relevance and aggregate at the daily level, applying a 7-day moving average to capture persistence. The month-end value represents the firm’s credit sentiment for that month. Figure 1 summarizes the overall process.

Negative sentiment associated with companies is generally regarded as indicative of underlying financial distress or adverse market perceptions. Therefore, articles expressing concerns about a company’s credit profile signal an elevated risk of default for the focal

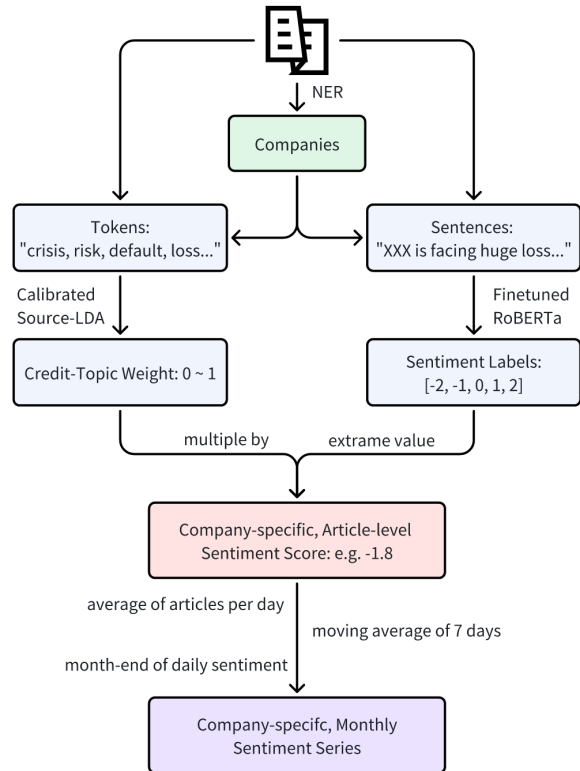


Figure 1. Company-specific, Topic-oriented Sentiment Score Construction Process

company, making the tracking and analysis of credit sentiment crucial for assessing default risk accurately.

Co-mention Network Construction. We build a monthly co-mention network by linking two companies whenever they are co-mentioned in the same article, using only news published within the prediction month to ensure temporal alignment and avoid look-ahead bias. From this network, we compute two standard measures: *degree* (the number of co-mentioned firms) and *transitivity* (the likelihood that two co-mentioned firms are also connected), calculated as:

$$k_i = \sum_j A_{ij}, \quad t_i = \frac{2e_i}{k_i(k_i - 1)} \quad (1)$$

where $A_{ij} = 1$ if an edge exists between i and j and e_i is the number of edges among neighbors of node i .

To incorporate event information, we construct a monthly binary indicator that equals 1 if the company is most frequently labeled as high-risk by the LLM in that month’s news, and 0 otherwise. This yields a high-risk sub-network from which we re-compute degree and transitivity, termed *Degree High Risk* and *Transitivity*

High Risk. These measures form the Labeled Text Model (LM).

Finally, we introduce a *Network Dummy* to distinguish between firms with zero degree because of no media coverage and those mentioned but without co-mentions, allowing us to examine the effect of media presence separately from network connectivity.

Logistic Regression Development. Since Altman (1968)'s seminal Z-score model, the corporate default prediction field has seen significant advancements, incorporating sophisticated econometric, statistical, and machine learning approaches. In this study, we deploy logistic regression for credit risk analysis. We also factor in an often-overlooked fact that corporate exits may occur for reasons other than default, such as mergers and acquisitions. As demonstrated by Duffie et al. (2007) and J.-C. Duan et al. (2012), these "other exits" significantly impact a firm's survival, typically occurring at a rate an order of magnitude higher than defaults. Thus, our categorical outcome variable Y takes values 0, 1, or 2 to denote survival, default, or other exit, respectively, at time t over horizon τ .

$$\ell_j = \ln\left(\frac{P_j}{1 - P_j}\right) = \beta_0^j + \beta_1^j x_1 + \beta_2^j x_2 + \dots + \beta_n^j x_n \quad (2)$$

The formula 2 shows how to calculate the logit (ℓ) of a category j . The β_0^j indicates the intercept of the regression curve for the class j , the β_i^j are the coefficients of each predictor x_i for the respective class j . Thus, $\beta^j = \{\beta_0^j, \beta_1^j, \dots, \beta_n^j\}$; n is the number of features. Since we set $event = 0$ as the reference category. Thus, we will get two sets of parameters for default ($Y = 1$) and other exit category ($Y = 2$).

In our sample, there are a total of 361 companies. Of them, 54 companies have defaulted, while 75 companies exited for reasons unrelated to default. Together, these companies generate 10,993 firm-month observations. As the horizon τ varies, there will be different amounts of 1 or 2 events in the monthly time series. Table 3 shows the event counts and default rate (precision) across horizons.

Because corporate defaults are rare events, our panel exhibits a highly imbalanced distribution between default and non-default observations. Rather than applying oversampling or synthetic rebalancing (e.g., SMOTE), we retain the natural event rate so that the estimated logistic regression intercept and predicted probabilities reflect the true unconditional default likelihood during the 2014–2016 oil crisis. This preserves the economic interpretability of our results

and avoids potential distortions introduced by synthetic data. Model robustness is ensured through the use of prevalence-insensitive metrics such as PR-AUC.

Summary of Variables. Our baseline set of structured financial variables is based on the corporate default prediction framework in NUS Credit Research Initiative (2023), which has been widely adopted and validated in global markets for more than a decade. It combines monthly macroeconomic indicators, firm-specific financial measures extracted from audited statements, and the distance-to-default (DTD) metric derived from the Bharath and Shumway (2004)'s KMV-Merton Model, thereby capturing both systematic and idiosyncratic drivers of default risk. This provides a robust and economically meaningful benchmark before incorporating textual sentiment and network features. The variables used in our study are summarized below, with descriptive statistics reported in Table 2.

Table 2. Descriptive Statistics of Text Features for the US Oil Companies

	Count	Mean	Std	Min	Max
Credit_Sentiment	10993	(0.006)	0.128	-1.838	1.700
Degree	10993	25.064	87.949	0	730
Transitivity	10993	0.170	0.353	0	1
Degree_High_Risk	10993	1.454	4.688	0	36
Transitivity_High_Risk	10993	0.114	0.306	0	1
Network_Dummy	10993	0.217	0.412	0	1

Table 3. Default Counts and Rates for US Oil Companies Across Different Horizons

Event Label	One Month	Three Months	Six Months	One Year
0 (Survival)	10882	10649	10301	9643
1 (Default)	62	189	374	674
2 (Other Exits)	49	155	318	676
Total	10993	10993	10993	10993
Default Rate	0.56%	1.72%	3.40%	6.13%

• **Focal Variables**

- Credit Sentiment: credit-focused sentiment scores of the focal company on each month-end
- Degree: the number of other companies that the focal company is directly co-mentioned with
- Transitivity: the likelihood that two companies connected to the focal company are also connected
- Network Dummy: the dummy variable if the focal company was reported by article news within a given month

For degree and transitivity, we include two types: unlabeled in the text model and labeled as high-risk of defaulting in the labeled text model.

- **Macro Variables** We include a comprehensive set of macroeconomic indicators: trailing one-year stock index return, 3-month interest rate, aggregate distance-to-default (DTD), real GDP growth, change in unemployment rate, CPI and PPI growth, NEER change, 3-month interbank rate change, S&P GSCI commodity index return, VIX change, house price index growth, current account balance change, and Cushing OK WTI oil price as a sector-specific indicator. Quarterly variables are held constant within the quarter, while higher-frequency predictors use month-end values.

- **Firm-specific Variables**

- DTD: The estimated distance-to-default is derived from the Merton structural model, as a measure of volatility-adjusted leverage
- Liquidity: Log ratio of cash and short-term investments to total assets for Chinese real estate; Log ratio of current assets to current liabilities for US oil
- Probability: A ratio of net income to total assets
- Relative Size: A log ratio of market capitalization to the economy’s median market capitalization
- Growth Opportunity: A ratio of market capitalization and total liabilities to total assets
- Sigma: The standard deviation of the residuals of a regression of the daily returns of the firm’s market capitalization on the daily returns of the economy’s stock index as a measure of idiosyncratic volatility

Time Alignment. All features are aligned based on public data availability. Firm-specific variables use the latest quarterly financial statements available at each prediction month-end with reporting lags applied; macroeconomic variables use the most recent published values; and news-based features (credit sentiment and network measures) are built strictly from articles published within the prediction month. This ensures no forward-looking information enters feature construction, preventing temporal leakage.

5. Empirical Findings

Regression Results. We show the results of text models in Table 4. The text model (TM) incorporates text-derived features, both sentiment and network features, such as credit sentiment, degree and transitivity within the co-mention network and network dummy in the logistic regression. The numbers 1, 3, 6 and 12 after TM indicate different months of horizons. Due to the space limit, we only show the statistics of text features.

Table 4. Text model for US Oil Companies

	TM1	TM3	TM6	TM12
Event = 1				
Control_Variables	Yes	Yes	Yes	Yes
Credit_Sentiment	0.451	-2.282***	-0.977*	-1.234***
Degree	0.008**	0.003*	0.002*	0.001
Transitivity	-0.666	-1.475***	-1.158***	-0.716*
Network_Dummy	2.322***	2.367***	1.921***	1.645***
Event = 2				
Control_Variables	Yes	Yes	Yes	Yes
Credit_Sentiment	0.079	0.323	0.655	0.767*
Degree	-0.029	-0.014**	-0.007***	-0.003***
Transitivity	-0.278	-0.669	-0.904**	-0.475
Network_Dummy	1.956***	1.139**	1.410***	0.625**

Note: Statistical significance: *p < 0.10, **p < 0.05, ***p < 0.01.

Our analysis reveals distinct patterns in the impact of credit sentiment on the likelihood of default and other exit events, within the oil industry. In the default analysis (Event = 1), the credit sentiment of oil companies is significant across most time horizons, with negative coefficients. This indicates that negative credit sentiment about a company is associated with a higher probability of default. However, the coefficient for the 1-month horizon is positive but not significant. This could be because when the default is imminent, financial market information, such as stock prices, likely reflects all available information, thereby diminishing the significance of sentiment (The control variables DTD level and DTD trend which are calculated by stock price have shown high significance at 0.01 level for 1 month horizon). When it comes to other exit which most cases are M&A (Event = 2), credit sentiment is only significant at the 12-month default prediction with a positive coefficient. This is in line with J. Duan and Yao (2022)’s empirical findings which claim that a financially distressed firm often becomes a cheap acquisition target but a well-performing company may be even more attractive to potential suitors.

Regarding the network features, we find that the degree is significant for short-term default predictions (less than one year) with a positive coefficient, while transitivity is most significant for terms longer than one month with a negative coefficient. In a company network, a high degree combined with low transitivity suggests that a company is connected to many others that are not strongly interconnected. While being frequently mentioned (high degree) may indicate market attention, have exposure to various sectors or industries, the lack of tight connections (low transitivity) suggests that these mentions do not represent strong relational ties or dependencies, such as the mutual support or collaborative ties that could help mitigate default risks. Together, these elements indicate that such companies

face greater unmanaged risks, leading to a higher probability of default.

When it comes to the other exit, both degree and transitivity coefficients become negative. In a company network, high degree and high transitivity indicate that a company is well-connected within tightly-knit clusters. This structure provides strong support networks, enhances stability, fosters collaborative synergy, and increases strategic value, making the company less vulnerable to financial difficulties and less attractive as an acquisition target. Additionally, such companies may benefit from mutual defence mechanisms against hostile takeovers. Consequently, these factors collectively reduce the likelihood of the company being acquired.

An interesting finding worth noting is the significant impact of media exposure (network dummy) on both default and other exit events. This suggests that a company being reported is more likely to undergo events such as default or acquisition compared to those not being reported. Media acts as a catalyst or watchdog, exposing financially distressed companies to a wider investor audience. This increased visibility prompts investors to exercise caution, leading to stock or bond sell-offs and accelerating the default process. Conversely, media attention enhances the visibility and market value of high-quality companies, potentially positioning them as acquisition targets. Media coverage often emphasizes a company's potential and value, attracting potential buyers and facilitating deal negotiations.

Table 5. Labeled Text Model for US Oil Companies

	LM1	LM3	LM6	LM12
Event = 1				
Control_Variables	Yes	Yes	Yes	Yes
Credit_Sentiment	0.297	-2.437***	-1.043*	-1.287***
Degree_High_Risk	0.018	0.073**	0.049**	0.024
Transitivity_High_Risk	-0.138	-0.524	-0.490	-0.228
Network_Dummy	2.228***	1.249***	1.099***	1.082***
Event = 2				
Control_Variables	Yes	Yes	Yes	Yes
Credit_Sentiment	0.133	0.421	0.761	0.819*
Degree_High_Risk	-0.003	-0.052	-0.041	-0.027
Transitivity_High_Risk	-2.241*	-1.902**	-1.572***	-0.859**
Network_Dummy	1.735***	0.887***	1.079***	0.533***

Note: Statistical significance: * p < 0.10, ** p < 0.05, *** p < 0.01.

Next, we examine the network effect by incorporating high-risk label information, which indicates companies facing elevated default risk (see Table 5). The labeled text model results show that for default events, the degree associated with high-risk labels becomes significantly more important at the 0.05 level compared to the text model's significance level of 0.1. However, the significance of transitivity with high-risk labels diminishes. These findings

suggest that refining the network to focus on high-risk companies, by removing irrelevant or already defaulted companies, improves the precision of default risk assessment. The high-risk subnetwork is sparser, with fewer connections than the full network. The predictive relevance of transitivity decreases, likely because the clustering among high-risk nodes contributes less incremental signal for default prediction, with overall connectivity already captured by degree centrality. This nuanced behavior highlights the importance of carefully selecting companies when constructing the network.

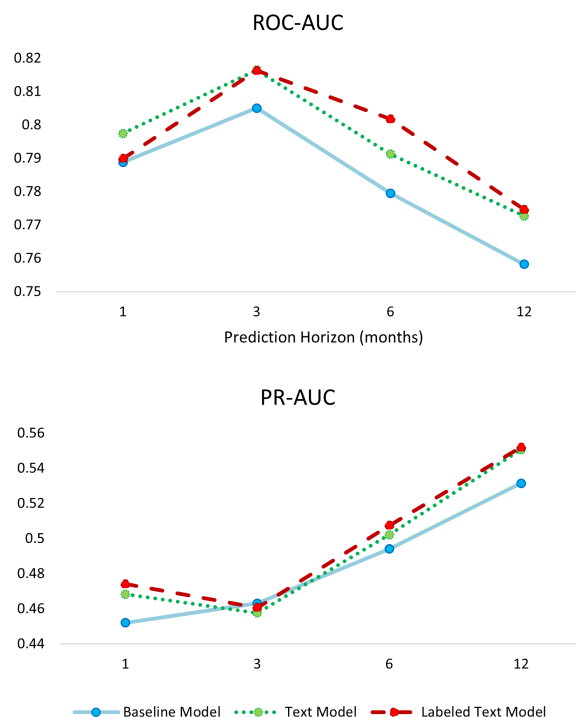


Figure 2. Model Performance for US Oil Companies

Prediction Performance. We use the Area Under the Curve (AUC) metric, ROC-AUC and PR-AUC, to evaluate how much the inclusion of text features enhances the predictive power of default events and to compare the performance of different models. While ROC-AUC reflects the ranking ability and is insensitive to the event rate, PR-AUC is more sensitive to the increasing prevalence of defaults at longer horizons. To ensure comparability across horizons, we also report the baseline precision (i.e., default rate). The results are presented in Figure 2. It is evident that both models incorporating text features outperform those relying solely on traditional structured financial variables. Specifically, the ROC-AUC of the Text Model and the Labeled Text Model are 1% to 2% higher

than that of the model without text features across all time horizons. Notably, the Text Model excels in short-term predictions (one month), while the Labeled Text Model shows superior performance in predicting defaults at the 6-month and 1-year horizons. The superior performance of the label model over longer horizons indicates that label-specific information may capture more persistent risk factors. On the other hand, performance diminishes over longer horizons. This is consistent with the nature of time-series prediction, where uncertainty accumulates as the forecast horizon extends. It also highlights the importance of regularly updating models with timely information.

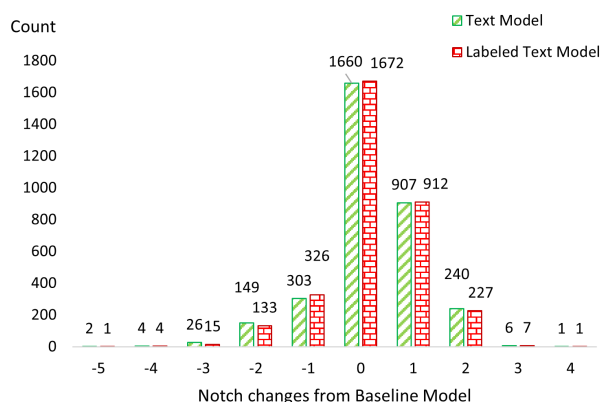


Figure 3. Distribution of PD-implied rating changes (in notches).

Economic Impact. To provide an economically intuitive interpretation, we map the model-predicted one-year PDs into PD-implied credit ratings using the standard S&P’s rating-PD correspondence table. We then compare the implied ratings obtained from the baseline model and the model augmented with textual sentiment and network features. The difference is reported in terms of rating notches, where a positive change indicates an upgrade (lower implied default risk) and a negative change indicates a downgrade (see Figure 3). Overall, we observe a right-shifted distribution, with more firms experiencing upgrades (lower PD-implied ratings) than downgrades, suggesting that textual information generally improves credit risk assessments. However, we also find that the left tail is fatter—the number of extreme downgrades (≥ 3 notches) exceeds that of extreme upgrades. This asymmetric tail behavior is consistent with the stylized fact that negative information shocks in credit markets tend to be more abrupt and severe, highlighting the model’s ability to capture downside risk more accurately.

6. Discussion

Applying text analysis to public news for predicting corporate default events has significant theoretical and practical implications. Theoretically, our study contributes to the literature by employing three types of text analysis for credit risk assessment and prediction, offering valuable insights into investor behavior. By examining the relationship between credit sentiments and corporate default events, we can better understand how opinions expressed in news articles influence the default process of financially distressed companies. Additionally, we highlight the significant role that company relationships within a co-mention network play in shaping a company’s vulnerability.

Although other exits, mostly including mergers and acquisitions (M&A), are not the primary focus of this study, it is noteworthy that the company network also provides valuable information about these events. For investors seeking acquisition targets, our findings offer useful insights into how company networks can indicate potential exits.

The practical implications of applying text analysis to default prediction are equally noteworthy. Our primary results have demonstrated the incremental predictive power of text features for default events in the US oil sector across different time horizons. However, factors impacting the oil industry, such as oil prices, are relatively straightforward. Although various economic or geopolitical factors can cause fluctuations in oil prices, the ultimate shock to the sector is due to changes in oil prices. This simplicity makes default prediction in the oil sector easier to analyze using text features.

A Supplementary Case Study of China Real Estate. To validate and generalize our approach across languages, economies, and sectors, we also considered the Chinese real estate upheaval from 2020 to 2023.

Starting in 2020, the Chinese government’s “three red lines” policy, aimed at curbing excessive borrowing, forced many property developers to deleverage. China Evergrande Group, one of the largest developers, epitomized the crisis, with its over-reliance on debt financing leading to severe liquidity issues and default. Subsequently, several other major developers faced defaults, sparking widespread concerns about systemic risks in China’s financial system. This regulatory tightening, coinciding with a slowdown in property sales exacerbated by the COVID-19 pandemic, has further strained developers’ finances and increased their default risks, signalling profound shifts in the future dynamics of China’s real estate market.

We collect 35,386 news pieces from Sina Finance, which were processed using native Chinese NLP

pipelines and prompts, preserving semantic nuances. Following the procedures outlined in the Session 4, we constructed variables of interest, including macro variables and firm-specific variables, consistent with those derived for the US Oil industry. There are altogether 298 companies giving rise to 12,247 firm-month observations in our sample. Among them, 60 companies have defaulted while 32 companies exited due to other reasons. And 257 of the 298 companies have been covered by media. The descriptive statistics of text features are shown in Table 6.

Table 6. Descriptive Statistics of Text Features for Chinese Real Estate Companies

	Count	Mean	Std	Min	Max
Credit_Sentiment	12247	(0.012)	0.367	-2	1.948
Degree	12247	8.341	23.095	0	370
Transitivity	12247	0.215	0.362	0	1
Degree_High_Risk	12247	4.021	9.857	0	128
Transitivity_High_Risk	12247	0.190	0.347	0	1
Network_Dummy	12247	0.318	0.466	0	1

Table 7 reports the default regression results for the text and labeled text models (TM and LM). Results for the 1-month horizon are omitted due to model non-convergence caused by too few default events. Across all horizons, credit sentiment is consistently negative and highly significant ($p < 0.01$), indicating that negative news reliably signals heightened default risk and may reveal information not fully priced by the market.

Relational variables show weaker and less stable effects. Degree and transitivity become predictive only when focusing on high-risk co-mentions (LM3), suggesting that network connections matter mainly when distress signals are concentrated among peer firms. The network dummy is generally insignificant, likely because most firms receive at least some media coverage, reducing its cross-sectional variation.

Table 7. Default Regression Results for Chinese Real Estates

	3 Months	6 Months	1 Year
Text Models			
Control Variables	Yes	Yes	Yes
Credit_Sentiment	-0.649***	-0.551***	-0.574***
Degree	0.004	-0.004	-0.003
Transitivity	-0.768	-0.674*	0.089
Network_Dummy	0.959**	0.830***	0.414
Labeled Text Models			
Control Variables	Yes	Yes	Yes
Credit_Sentiment	-0.650***	-0.555***	-0.585***
Degree_High_Risk	0.028***	0.007	0.004
Transitivity_High_Risk	0.226	-0.138	0.568**
Network_Dummy	-0.015	0.313	-0.018

Note: Statistical significance: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 8. Model Performance for Chinese Real Estate Companies

	3 Months	6 Months	1 Year
ROC-AUC			
Baseline Model	0.820	0.832	0.841
Text Model	0.830	0.833	0.842
Labeled Text Model	0.830	0.833	0.843
PR-AUC			
Baseline Model	0.390	0.460	0.498
Text Model	0.401	0.475	0.518
Labeled Text Model	0.399	0.475	0.519

Table 8 shows that incorporating textual features leads to consistent but modest gains in ROC-AUC and PR-AUC across 3-, 6-, and 12-month horizons, highlighting that text information provides incremental predictive power beyond structured financial variables.

7. Future research

The study in its current form has a few limitations that point to directions for future research. First, other alternative data sources such as earnings conference calls, financial forum discussions, and YouTube videos could also be mined to predict corporate defaults. However, due to data constraints, our analysis primarily derives predictors from news articles. Moving forward, we can explore these additional data sources to enrich our models. This would allow us to capture a broader spectrum of financial sentiment and operational nuances that are not fully represented in news articles alone.

Second, the current study utilizes a binary co-mention network, which, while effective, simplifies the relationships between co-mentioned companies by categorizing them as either present or absent. We plan to incorporate weights for co-mentions to capture the intensity and significance of co-mentions across articles. By providing deeper insights into the dynamics within the network, we can build a more nuanced understanding of the interconnections and influence among companies.

Third, while the current study focuses primarily on exploring new text features, we acknowledge that a deeper discussion of potential mechanisms and a stronger connection to management or organizational theory is an avenue for future research, to better understand the theoretical rationale behind different features (e.g., co-mention and sentiment features).

Last, we acknowledge the limitation that the U.S. dataset ends in 2016, which constrains the real-time applicability of our findings. Nonetheless, it provides a solid foundation for developing and validating our modeling approach. To further validate and generalize

our approach across languages, economies, and sectors, we also examined the Chinese real estate upheaval from 2020 to 2023.

References

- Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *Journal of Finance*, 23(4), 589–609.
- Bharath, S. T., & Shumway, T. (2004). Forecasting default with the kmv-merton model [Available at SSRN: <https://ssrn.com/abstract=637342>]. *AFA 2006 Boston Meetings Paper*.
- Cecchini, M., Aytug, H., Koehler, G. J., & Pathak, P. (2010). Making words work: Using financial text as a predictor of financial events. *Decision support systems*, 50(1), 164–175.
- Chen, K., Li, X., Luo, P., & Zhao, J. L. (2021). News-induced dynamic networks for market signaling: Understanding the impact of news on firm equity value. *Information Systems Research*, 32(2), 356–377.
- Creamer, G. G., Ren, Y., & Nickerson, J. V. (2013). Impact of dynamic corporate news networks on asset return and volatility. *2013 international conference on social computing*, 809–814.
- Credit Research Initiative. (2023). *NUS Credit Research Initiative Technical Report* (tech. rep.). National University of Singapore. <https://nuscri.org/en/technical-report.html>
- Dong, W., Liao, S., & Zhang, Z. (2018). Leveraging financial social media data for corporate fraud detection. *Journal of Management Information Systems*, 35(2), 461–487.
- Duan, J., & Yao, X. (2022, August). *Media sentiments for enhanced credit risk assessment* (tech. rep.) (Working Paper).
- Duan, J.-C., Sun, J., & Wang, T. (2012). Multiperiod corporate default prediction – a forward intensity approach. *Journal of Econometrics*, 170, 191–209.
- Duffie, D., Saita, L., & Wang, K. (2007). Multi-period corporate default prediction with stochastic covariates. *Journal of Financial Economics*, 83, 635–665.
- Dunham, L. M., & Garcia, J. (2021). Measuring the effect of investor sentiment on financial distress. *Managerial Finance*.
- Groß-Klußmann, A., König, S., & Ebner, M. (2019). Buzzwords build momentum: Global financial twitter sentiment and the aggregate stock market. *Expert Systems with Applications*, 136, 171–186.
- Huang, A. H., Lehavy, R., Zang, A. Y., & Zheng, R. (2018). Analyst information discovery and interpretation roles: A topic modeling approach. *Management science*, 64(6), 2833–2855.
- Jin, Y., Lin, C.-Y., Matsuo, Y., & Ishizuka, M. (2012). Mining dynamic social networks from public news articles for company value prediction. *Social Network Analysis and Mining*, 2, 217–228.
- Li, F. (2010). The information content of forward-looking statements in corporate filings—a naive bayesian machine learning approach. *Journal of Accounting Research*, 48(5), 1049–1102.
- Netzer, O., Feldman, R., Goldenberg, J., & Fresko, M. (2012). Mine your own business: Market-structure surveillance through text mining. *Marketing Science*, 31(3), 521–543.
- Roeder, J., Palmer, M., & Muntermann, J. (2020). Utilizing news topics for credit risk management: The explanation of bank cds spreads. *Journal of Decision Systems*, 29(sup1), 32–44.
- Sun, Y., Liu, X., Chen, G., Hao, Y., & Zhang, Z. (2020). How mood affects the stock market: Empirical evidence from microblogs. *Information & Management*, 57(5), 103181.
- Tan, T., & Phan, T. Q. (2018). Social media-driven credit scoring: The predictive value of social structures. *Available at SSRN 3217885*.
- 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 and Finance*, 68, 100–116.
- Wood, J. (2016). Source-lda: Enhancing probabilistic topic models using prior knowledge sources. *CoRR, abs/1606.00577*. <http://arxiv.org/abs/1606.00577>
- Xing, F. Z., Cambria, E., & Zhang, Y. (2019). Sentiment-aware volatility forecasting. *Knowledge-Based Systems*, 176, 68–76.