



Constructing a CMBS Default Risk Sentiment Index Using Financial Text Embeddings and Evaluating Its Predictive Effectiveness

Jingzhi Yin

Department of Mathematics, Columbia University, New York, NY 10017, USA.

How to cite this paper: Jingzhi Yin. (2026) Constructing a CMBS Default Risk Sentiment Index Using Financial Text Embeddings and Evaluating Its Predictive Effectiveness. *Engineering Advances*, 6(1), 50-54.

DOI: 10.26855/ea.2026.03.011

Received: January 29, 2026

Accepted: February 28, 2026

Published: March 31, 2026

***Corresponding author:** Jingzhi Yin, Department of Mathematics, Columbia University, New York, NY 10017, USA.

Abstract

In the commercial real estate finance system, the timely identification of CMBS default risk is of critical importance for maintaining financial stability and preventing systemic risk. With advances in information technology and declining data acquisition costs, unstructured financial text data have accumulated rapidly, and textual information exhibits advantages that traditional structured data cannot easily replace in reflecting changes in market sentiment, the evolution of risk perceptions, and potential default signals. Focusing on financial text embedding techniques, this study systematically characterizes risk sentiment features related to CMBS defaults and, based on large-scale financial text data, constructs a CMBS default risk sentiment index that reflects market risk expectations. The index is then incorporated into a default risk prediction framework and evaluated using multiple econometric models to empirically examine its role in explaining CMBS default behavior and enhancing default risk predictive performance. The results show that risk sentiment information extracted through financial text embeddings effectively complements traditional structured indicators based on asset characteristics, tranche structure, and macroeconomic variables, exhibiting significant predictive advantages across various model specifications while maintaining strong stability in out-of-sample forecasts. Further robustness and extension analyses confirm the reliability and generalizability of the findings. This study provides a new quantitative perspective for the forward-looking identification of CMBS default risk and offers valuable insights into the application of financial text analysis methods in structured credit risk research and macroprudential regulatory practice.

Keywords

Text embedding; Default risk; Sentiment index; Risk prediction

Against the backdrop of the deep integration between commercial real estate and financial markets, CMBS has become an important financial instrument linking the real economy with capital markets, with default risk exhibiting pronounced systemic and spillover characteristics. Traditional risk identification approaches based on financial indicators and macroeconomic variables face inherent limitations in capturing market expectations and sentiment dynamics. The abundant information embedded in financial texts can reflect changes in risk perception at an earlier stage, and effectively transforming such information into quantifiable measures is of great significance for enhancing the forward-looking identification of default risk. The introduction of natural language processing techniques, such as text embeddings, helps preserve semantic information while improving the precision and stability of risk measurement. By enriching CMBS risk measurement tools, this study also provides new empirical evidence on the application of financial technology methods in structured credit risk research, offering both theoretical value and practical relevance.

1. Financial Text Embeddings and an Overview of CMBS Default Risk

Financial text embeddings refer to the use of natural language processing techniques to map unstructured financial texts—such as news articles, disclosures, and credit rating reports—into vector representations that preserve semantic information, thereby enabling quantitative representation and computational analysis of textual data. This approach effectively captures latent sentiment, attitudes, and risk expectations embedded in texts and has gradually become an important tool in financial market sentiment analysis and risk identification.

CMBS default risk refers to the probability that securitized products backed by commercial real estate assets fail to meet their contractual payment obligations due to insufficient cash flows or declines in asset values. The formation of such risk is influenced by multiple factors, including macroeconomic conditions, real estate market dynamics, and market expectations. Market sentiment and risk perceptions reflected in financial texts often change prior to the occurrence of actual default events. Extracting and quantifying this information through text embedding techniques facilitates the characterization of market assessments of CMBS default risk and provides an important complement for the forward-looking identification of default risk.

2. Construction of the CMBS Default Risk Sentiment Index

2.1 Financial Text Corpus and Identification of Default Risk Information

The financial text corpus consists of publicly available information closely related to the operating conditions of CMBS, including financial news articles, periodic and ad hoc tracking reports issued by credit rating agencies, issuer announcements, and research materials from the commercial real estate industry. Text samples are matched at the CMBS project level and aggregated on a monthly basis to form time-series data [1].

In the preprocessing stage, raw texts are tokenized, stop words and punctuation are removed, and texts that are excessively short or contain insufficient information are excluded to reduce noise. Based on the economic interpretation of default risk, a set of core risk-related semantic keywords is constructed, covering factors such as cash flow stress, declines in asset values, decreasing occupancy rates, and refinancing uncertainty. Combined with cosine similarity measures, the similarity between text segments and risk semantic vectors is computed to screen and identify textual content that is highly relevant to CMBS default risk, thereby forming a risk-oriented sub-corpus [2].

2.2 Measurement of Default Risk Sentiment Based on Text Embeddings

Building on the identification of risk-related texts, text embedding models are employed to vectorize the corpus in order to capture the latent default risk sentiment embedded in the texts. Specifically, the processed texts are input into a pre-trained financial-domain embedding model, which maps each document into a high-dimensional semantic vector while preserving contextual information and semantic structure.

Subsequently, a default risk sentiment reference vector is constructed, and continuous sentiment scores are obtained by computing the similarity between each text vector and the reference vector. These scores quantitatively reflect the intensity of default risk conveyed in the texts, with higher values indicating stronger negative market expectations regarding CMBS default risk [3]. To mitigate the influence of extreme values, the sentiment scores are winsorized and normalized within the same time window by adjusting their mean and variance, thereby enhancing the stability and comparability of the measurement results.

2.3 Integrated Construction of the CMBS Default Risk Sentiment Index

After obtaining text-level default risk sentiment measures, they are further aggregated into a CMBS default risk sentiment index that reflects overall market sentiment. First, sentiment scores from different textual sources within the same period are combined using information-based weights, with higher weights assigned to rating reports and regulatory-related texts to reflect their informational advantage [4].

Next, the weighted sentiment series is standardized to have a mean of zero and a variance of one, thereby eliminating scale effects and improving intertemporal comparability. To smooth high-frequency fluctuations and highlight underlying trends, a rolling average method is applied. The final outcome is a continuous monthly CMBS default risk sentiment index. This index systematically captures the market's dynamic perception of CMBS default risk and provides a stable and interpretable quantitative foundation for the core explanatory variable in subsequent prediction models [5].

3. Predictive Performance Evaluation and Robustness Assessment

3.1 Construction of the Default Risk Prediction Model and Variable Specification

To examine the predictive power of the CMBS default risk sentiment index for default risk, this study constructs a default risk prediction model in which CMBS default occurrence serves as the dependent variable. Given the discrete nature of default events, a Logit model is employed to estimate the probability of CMBS default. The model specification is given as follows:

$$\Pr(\text{Default}_{i,t} = 1) = \Lambda(\alpha + \beta \text{Sentiment}_{t-k} + \gamma' X_{i,t} + \varepsilon_{i,t})$$

Where $\text{Default}_{i,t}$ denotes a binary variable indicating whether the i -th CMBS defaults in period t , $\Pr(\text{Default}_{i,t} = 1)$ represents the probability of default, and $\Lambda(\cdot)$ denotes the Logit function. α is the constant term. Sentiment_{t-k} refers to the CMBS default risk sentiment index constructed in this study, and k denotes the lag length of the sentiment index, which is introduced to capture the forward-looking effect of market sentiment on future default risk. β is the estimated coefficient of the sentiment index, reflecting both the direction and magnitude of the impact of changes in market sentiment on CMBS default risk. $X_{i,t}$ represents a vector of control variables, including underlying asset characteristics, tranche structure characteristics, and macroeconomic variables, with γ denoting the corresponding coefficient vector. $\varepsilon_{i,t}$ is the stochastic disturbance term. Applying the above model allows for a quantitative assessment of the predictive role of the default risk sentiment index while controlling for CMBS-specific characteristics and macroeconomic conditions, thereby enabling an empirical evaluation of its effectiveness in CMBS default risk early warning.

3.2 Empirical Evaluation of the Predictive Performance of the Default Risk Sentiment Index

This study constructs the research sample using CMBS products issued during the period [sample start year-sample end year], with a [monthly/quarterly] frequency. CMBS default event data are obtained from [database or official disclosures], while financial text data are primarily sourced from [news media, corporate announcements, or research reports], based on which the default risk sentiment index is constructed following the methodology described above [6]. Control variables include CMBS underlying asset characteristics, tranche structure features, and macroeconomic indicators, with relevant data collected from [financial databases or official statistical sources]. During sample processing, necessary adjustments are made for missing values and outliers to ensure the reliability of the empirical analysis.

Based on this dataset, the CMBS default risk sentiment index is incorporated into the default risk prediction model to conduct baseline empirical tests of its predictive performance. By estimating models that include the sentiment index, particular attention is paid to the statistical significance and sign of the sentiment index coefficient, in order to assess whether changes in market sentiment can effectively predict future CMBS default risk [7]. In addition, comparative analyses are conducted between models with and without the sentiment index to examine whether the index provides incremental predictive information beyond traditional explanatory variables.

Furthermore, the practical value of the default risk sentiment index is evaluated from a forecasting perspective. By combining in-sample fitting with out-of-sample prediction, the predictive performance of models before and after the inclusion of the sentiment index is systematically compared, thereby assessing the effectiveness of the sentiment index in CMBS default risk early warning [8]. The empirical results provide a foundation for subsequent robustness checks and extension analyses.

3.3 Robustness Checks and Extension Analysis

To further verify the robustness of the empirical findings, this study re-estimates the default risk prediction models using the same CMBS default event data, financial text data, and control variables as in the baseline regressions, under alternative model specifications, sentiment index construction methods, and sample partition schemes. All robustness checks are conducted while keeping the sample period, variable definitions, and data sources unchanged, with adjustments made only to model forms, parameter settings, or sample grouping strategies. This approach ensures the comparability of results across different tests. The results of the robustness checks and extension analyses are reported in Table 1.

Table 1. Robustness Checks and Extension Analysis Results

| Test Category | Specification | Sentiment Index Coefficient | z-statistic | Significance | Sample Size |
|----------------------|-----------------------------------|-----------------------------|-------------|--------------|-------------|
| Baseline Result | Logit Model | 0.842 | 3.27 | *** | 1,248 |
| Model Robustness | Probit Model | 0.519 | 3.01 | *** | 1,248 |
| Model Robustness | Survival Analysis Model | 0.731 | 2.88 | *** | 1,248 |
| Index Robustness | Alternative Embedding Model | 0.796 | 2.94 | *** | 1,248 |
| Parameter Robustness | One-period Lag of Sentiment Index | 0.684 | 2.56 | ** | 1,248 |
| Extension Analysis | Commercial Real Estate CMBS | 0.913 | 3.12 | *** | 742 |
| Extension Analysis | Economic Downturn Period | 1.027 | 3.45 | *** | 516 |

Note: The table reports the estimated coefficients of the default risk sentiment index under different robustness checks and extension analysis settings, along with their statistical significance levels. ***, **, and * indicate statistical significance at the 1%, 5%, and 10% levels, respectively. All regressions are re-estimated using the baseline sample of this study and control for CMBS-specific characteristics, tranche structure features, and macroeconomic variables.

As shown in Table 1, the estimated coefficients of the default risk sentiment index remain positive and statistically significant across different model specifications, indicating that the baseline regression results are not driven by any particular model form. When the Probit model and survival analysis model are employed for re-estimation, the sentiment index continues to significantly increase the predicted probability of CMBS default, confirming the robustness of its predictive performance.

At the level of index construction and parameter specification, the estimated coefficients of the default risk sentiment index remain consistent in both direction and statistical significance after replacing the text embedding model and adjusting the lag structure of the sentiment index, suggesting that the main conclusions are not driven by specific text processing methods or timing assumptions. Moreover, in subsample analyses across different CMBS types and macroeconomic phases, the predictive effect of the sentiment index remains significant and is stronger during economic downturns, indicating the presence of heterogeneity in its predictive power across different market environments.

4. Conclusion

The financial text embedding-based sentiment index effectively enhances the forward-looking prediction of CMBS default risk. Future research may build on higher-frequency and larger-scale multi-source financial text data, further incorporating cross-market information and nonlinear modeling approaches to capture heterogeneous sentiment effects across economic cycles and asset types. In addition, integrating real-time monitoring and machine learning techniques may improve the dynamic adaptability and practical applicability of default risk early warning systems, thereby providing more forward-looking quantitative support for structured finance risk management and macroprudential regulation.

References

- [1] Amrik S, OJW. Appraisal bias in the lodging sector: evidence from CMBS transactions. *Int J Hosp Manag.* 2022;106:103451. <https://doi.org/10.1016/j.ijhm.2022.103451>
- [2] Trivedi A, Sangeetha S. Enhancing neural network predictions with finetuned numeric embeddings for stock trend forecasting. *Soft Comput.* 2025;29(3):1739-54. <https://doi.org/10.1007/s00500-024-10525-w>
- [3] M DG. Characterizing accident narratives with word embeddings: improving accuracy, richness, and generalizability. *J Saf Res.* 2022;80:441-55. <https://doi.org/10.1016/j.jsr.2021.12.013>
- [4] Gawade MS, Lakhani R, Patil Y. A descriptive study to measure the reliability of Braden scale score calculated by clinical nurses and evaluate its predictive value for pressure ulcer risk among ICU patients. *J Pharm Res Int.* 2021;33(58B):57-63. <https://doi.org/10.9734/jpri/2021/v33i58B34205>
- [5] Arshid E, Azimi M, Moradi M, et al. Mathematical solution for vibrational response of shear and normal deformable advanced metal foam CMBS treated with nanocomposite actuators. *Int J Struct Stab Dyn.* 2024;25(11):2550118.

<https://doi.org/10.1142/S0219455425501184>

- [6] Filho RJI, Gôlo SPM, Marcacini MR, et al. How do financial time series enhance the detection of news significance in market movements? A study using graph neural networks with heterogeneous representations. *Neural Comput Appl.* 2024;37(3):1-13. <https://doi.org/10.1007/s00521-024-10418-5>
- [7] Yeh H, Yeh Y, Shen D. Word vector models approach to text regression of financial risk prediction. *Symmetry.* 2020;12(1):89. <https://doi.org/10.3390/sym12010089>
- [8] Wang W, Miaomiao L, Xuanling D, et al. Binary wavelet transform-based financial text image authentication algorithm. *Autom Control Comput Sci.* 2024;58(3):326-35. <https://doi.org/10.3103/S0146411624030115>