



Research on a CLO Secondary Market Spread Volatility Prediction Model Based on RoBERTa Sentiment Factors

Jingzhi Yin

The Department of Mathematics, Columbia University in the City of New York, New York, NY 10027, USA.

How to cite this paper: Jingzhi Yin. (2026) Research on a CLO Secondary Market Spread Volatility Prediction Model Based on RoBERTa Sentiment Factors. *Advances in Computer and Communication*, 7(1), 38-42.
DOI: 10.26855/acc.2026.03.005

Received: January 7, 2026
Accepted: February 5, 2026
Published: March 2, 2026

***Corresponding author:** Jingzhi Yin, The Department of Mathematics, Columbia University in the City of New York, New York, NY 10027, USA.

Abstract

Against the backdrop of a complex macroeconomic environment and intensified fluctuations in credit risk cycles, spreads in the CLO secondary market exhibit increasingly pronounced nonlinear and sentiment-driven characteristics. Traditional forecasting approaches based on fundamentals and market variables face limitations in capturing short-term spread dynamics. This study introduces sentiment factors derived from the RoBERTa pre-trained language model, extracting investor sentiment signals from unstructured textual information and integrating them with CLO secondary market spread volatility to construct a sentiment-driven forecasting model. Through a systematic process of sentiment factor quantification and feature engineering, textual sentiment information is transformed into numerical variables suitable for prediction, followed by model training and optimization. Empirical results show that sentiment factors significantly enhance both the explanatory power and predictive accuracy of spread volatility, with the model demonstrating strong robustness across different testing conditions. These findings provide empirical evidence on the role of sentiment information in the pricing of structured credit products and offer new analytical perspectives for risk management and investment decision-making in the CLO market.

Keywords

RoBERTa; sentiment factor; spread volatility; forecasting model

As structured credit products continue to gain importance in global capital markets, price fluctuations in the CLO secondary market and their risk transmission mechanisms have attracted increasing attention. As a key indicator reflecting credit risk and market expectations, spread volatility is influenced not only by fundamental factors but also increasingly by investor sentiment and changes in the information environment. In a context where information dissemination relies heavily on textual media, effectively capturing and quantifying sentiment factors to improve the understanding and prediction of CLO spread volatility is of considerable theoretical and practical significance. Advances in deep learning-based natural language processing provide new avenues for extracting sentiment information, and incorporating such techniques into related research helps address limitations of traditional approaches while offering useful references for risk management and pricing decisions.

1. Overview of RoBERTa-Based Sentiment Factors and CLO Secondary Market Spread Volatility

The CLO secondary market is a key venue for trading structured credit products, and its spreads reflect market

assessments of underlying asset credit risk, liquidity conditions, and the overall economic environment, serving as a core measure of CLO pricing and risk. Spread volatility typically exhibits strong time-varying and nonlinear characteristics, driven not only by changes in fundamentals but also significantly influenced by fluctuations in market sentiment. RoBERTa-based sentiment factors are quantitative indicators derived from semantic understanding and sentiment recognition of large-scale textual information using pre-trained language models, enabling a relatively accurate depiction of market participants' emotional states during information dissemination. As financial information becomes increasingly text-based, investor sentiment conveyed through news reports, research publications, and market commentary continuously shapes trading expectations and, in turn, affects CLO secondary market spread volatility. Introducing RoBERTa-based sentiment factors into spread analysis frameworks helps reveal the transmission mechanisms through which sentiment information influences CLO spread formation and dynamics.

2. Construction of a Sentiment-Based CLO Spread Volatility Forecasting Model

2.1 Text Data Collection and Sentiment Factor Quantification

Textual data are filtered using keyword matching and named-entity recognition rules, retaining only content related to CLOs, leveraged loans, credit risk, and macro-financial conditions, while removing duplicate and template-based texts. During filtering, information on publication time, source type, and associated entities is preserved for subsequent weighting procedures. All texts are segmented at the sentence level and uniformly fed into the RoBERTa model, which outputs a sentiment score or sentiment probability vector for each text as the raw sentiment representation. To enhance the model's adaptability to financial contexts, RoBERTa can be fine-tuned using publicly available financial sentiment datasets or a small, manually labeled domain-specific corpus. Key technical settings for text processing and RoBERTa-based sentiment factor generation are reported in Table 1.

Table 1. Key Technical Settings for Text Processing and RoBERTa-Based Sentiment Factor Construction

Item	Technical Specification
Text data sources	Bloomberg News and rating agency public announcements
Effective text volume	Approximately 42,000 documents
Text segmentation level	Sentence-level
RoBERTa model version	RoBERTa-base (financial-domain fine-tuned)
Sentiment output format	Continuous sentiment score (-1 to 1)
Daily sentiment factor construction	Weighted aggregation (source weight \times text length)

The sentiment outputs of individual texts are aggregated on a trading-day basis to construct daily sentiment factors. The aggregation process includes calculating indicators such as the daily average sentiment, the intensity of negative sentiment, and sentiment dispersion, in order to capture both the overall level and the cross-sectional variability of market sentiment. During aggregation, sentiment scores are combined using weighted averages based on the authority of the text source and text length, thereby reducing noise introduced by low-quality information [1]. The resulting structured and quantifiable daily sentiment variables can be directly incorporated into subsequent spread volatility forecasting models.

2.2 Alignment of Sentiment Factors with Spread Series, Feature Engineering, and Sample Construction

The alignment between sentiment factors and spread series follows strict time-mapping rules to ensure the absence of look-ahead bias. Texts released during trading hours are assigned to the same trading day's sentiment factor, while texts published after market close are aggregated into the next trading day. To further avoid timing mismatches, both sentiment factors and spread data are aligned using a unified trading calendar, with observations corresponding to non-trading days removed. All sentiment factors enter the prediction framework in lagged form, with commonly used lags of 1, 5, and 20 trading days to capture the persistence and decay of sentiment effects across different time horizons.

The spread volatility label is defined as the absolute change or squared change in spreads over the future h -period, with extreme values mitigated through percentile-based winsorization. In addition to sentiment factors, the sample includes lagged spread terms as well as benchmark credit spreads and interest rate variables as controls, with all features standardized prior to estimation [2]. The final dataset uses the “trading day” as the basic observational unit, constructing a feature matrix that combines sentiment and market characteristics with corresponding target vectors, and is partitioned into training and testing sets in chronological order to provide a stable foundation for subsequent model estimation.

2.3 Model Training, Optimization, and Forecast Output Design

Model training adopts a parallel multi-model strategy. Linear regression and GARCH-type models are used as benchmarks to compare predictive performance before and after the inclusion of sentiment factors, followed by the application of XGBoost or LSTM models to capture nonlinear relationships. Training is conducted using rolling time windows to ensure that model parameters are updated solely based on historical information. Hyperparameters are selected by minimizing validation-set errors, with mean absolute error (MAE) and root mean squared error (RMSE) serving as the primary evaluation metrics for regression tasks [3].

To enhance the interpretability of forecasting results, feature importance decomposition methods are employed to quantify the contribution of sentiment factors across different forecasting horizons. Forecast outputs include both point estimates and volatility intervals derived from quantile regression, providing directly actionable quantitative inputs for risk monitoring and position management.

3. Validation of the CLO Secondary Market Spread Volatility Forecasting Model

3.1 Research Methodology and Sample Data Sources

This study employs a time-series-based supervised learning framework to empirically examine the role of sentiment factors in forecasting CLO secondary market spread volatility. The research focuses on short-term spread dynamics, with samples constructed at a daily frequency to capture the immediate impact of sentiment information on market pricing [4]. Spread data are drawn from tradable CLO tranches or corresponding CLO indices, with secondary market quoted spreads obtained from Bloomberg and Intex, covering the period from 2016 to 2024 and exhibiting good trading continuity.

Textual data are sourced from information channels closely related to credit markets, including financial news, rating agency announcements, and research report summaries, primarily obtained from Bloomberg News, Refinitiv, and rating agency websites. All texts retain precise publication timestamps for subsequent temporal alignment. To ensure sample relevance, only texts related to CLOs, leveraged loans, or credit risk within the study period are retained [5]. All datasets are organized along a unified timeline to construct daily observations, providing the basis for model specification and empirical analysis.

3.2 Model Specification, Variable Definitions, and Empirical Analysis Procedure

The empirical model uses future spread volatility as the dependent variable, with sentiment factors serving as the core explanatory variables alongside a set of control variables to form the forecasting equation. Spread volatility is measured as the absolute change in spreads over the future h -period, calculated as follows:

$$V_0 I_{t+h} = |Spread_{t+h} - Spread_t|$$

In the equation, $V_0 I_{t+h}$ denotes the magnitude of CLO secondary market spread volatility over the future h - period starting from day t ; $Spread_t$ represents the level of the CLO secondary market spread on day t ; $Spread_{t+h}$ denotes the corresponding spread level on day $t+h$; and h indicates the forecast horizon. In this study, the forecast horizon is primarily set to ($h = 1$), ($h = 5$), and ($h = 20$) trading days to capture short- and medium-term spread volatility characteristics.

The core explanatory variable is the daily sentiment factor constructed using the RoBERTa model, denoted by $Sent_t$, which reflects the overall sentiment embedded in market-related textual information on day t . To avoid look-

ahead bias, only lagged values of the sentiment factor are included in the forecasting model. Control variables include lagged spread terms, changes in benchmark credit spreads, and interest rate levels [6].

In the empirical implementation, all continuous variables are first standardized to eliminate the impact of scale differences on parameter estimation, and extreme observations are mitigated through percentile-based winsorization to reduce the influence of abnormal fluctuations. The sample is then divided into training and testing sets in chronological order to avoid information leakage caused by random splitting. During model estimation, parameters are fitted using the training sample and updated continuously through a rolling-window approach to simulate a realistic forecasting environment. Finally, period-by-period forecasts are generated for the testing sample, and prediction error metrics are computed to evaluate model stability and forecasting performance under different specifications [7].

3.3 Empirical Results Analysis and Robustness Checks

After model estimation, the predictive performance of the models incorporating sentiment factors is systematically examined. The baseline model includes only lagged spread terms and market control variables to capture the time-dependent characteristics of CLO secondary market spread volatility [8]. On this basis, sentiment factors constructed using the RoBERTa model are added, and the explanatory power and predictive accuracy of the two model specifications are compared. Table 2 reports the main regression results.

Table 2. Predictive Results of Sentiment Factors on CLO Secondary Market Spread Volatility

Variable	Baseline Model	Sentiment Factor Model
Sentiment Factor (Sent ₋₁)	—	0.121 (0.039)
Lagged Spread Volatility (Volt)	0.362 (0.065)	0.334 (0.061)
Benchmark Credit Spread	0.097 (0.041)	0.083 (0.038)
Interest Rate Level	-0.069 (0.036)	-0.062 (0.034)
Constant	0.013 (0.009)	0.011 (0.008)
Observations	1,980	1,980
R2	0.214	0.262
MAE	0.084	0.071
RMSE	0.113	0.097

Note: Robust standard errors are reported in parentheses; all variables are standardized.

As shown in Table 2, the explanatory power of the model improves markedly after the inclusion of sentiment factors, with the R2 increasing from 0.214 to 0.262. The estimated coefficient on the sentiment factor is positive and statistically significant, indicating that textual sentiment information has a significant predictive effect on future spread volatility. At the same time, the coefficient on the lagged spread term declines, suggesting that sentiment factors absorb part of the historical volatility information that is not captured by traditional variables.

From the perspective of forecast error metrics, the inclusion of sentiment factors leads to notable reductions in both the mean absolute error (MAE) and the root mean squared error (RMSE) in the test sample, indicating higher predictive accuracy in practical forecasting applications. Robustness checks further involve alternative measures of spread volatility, different forecast horizons, and repeated estimations using alternative model specifications. The sign and statistical significance of the core variables remain consistent across these settings, indicating that the empirical results are not driven by specific model choices and exhibit strong robustness.

4. Conclusion

Sentiment information provides a stable and statistically significant incremental contribution to forecasting CLO secondary market spread volatility. Future research may extend the construction of sentiment factors by incorporating longer time spans, higher-frequency data, multilingual text sources, and cross-market sentiment signals to further elucidate sentiment transmission mechanisms. In addition, integrating the heterogeneous features of CLO tranches

and liquidity differences with forecasting outputs and practical trading or risk-warning frameworks could enhance the model's applicability and decision-support value in complex market environments.

References

- [1] Kumar P, Ganapathy SM. An Analysis of Marketing Channels and Price Spread of Chrysanthemum in Chikkaballapura District of Karnataka. *Asian J Agric Ext Econ Soc.* 2023;41(5):93-98.
- [2] Luo M, Liu S, Zhu L, et al. Analysis of a super-transmission of SARS-CoV-2 omicron variant BA.5.2 in the outdoor night market. *Front Public Health.* 2023;11:1153303.
- [3] Jiang C, Zhang Y, Han Y. Do Bond Investors Care about Margin Trading and Securities lending?:—An Empirical Study Based on Secondary Market of Medium Term Notes. *Procedia Computer Science.* 2022 Jan 1;199:613-20.
- [4] Wang S, Liu Q, Hu Y, et al. Public Opinion Evolution Based on the Two-Dimensional Theory of Emotion and Top2Vec-RoBERTa. *Symmetry.* 2025;17(2):190.
- [5] Ramalingaswamy C, Khaja H, Ilaiah K, et al. Sentiment classification with modified RoBERTa and recurrent neural networks. *Multimed Tools Appl.* 2023;83(10):29399-417.
- [6] Redouane E, Yoshio N. Fire-sale risk in the leveraged loan market. *J Financ Econ.* 2022;146(3):1120-47.
- [7] Maranon Issues Eighth Middle Market CLO. *Manuf Close-Up.* 2021.
- [8] Dennis V, Mike N, Vivian BV. Security design and credit rating risk in the CLO market. *J Int Financ Mark Inst Money.* 2021;72:101290.
- [9] Lak JA, Boostani R, Alenizi AF, et al. RoBERTa, ResNeXt and BiLSTM with self-attention: The ultimate trio for customer sentiment analysis. *Appl Soft Comput.* 2024;164:112018.