The Impact of News Sentiment on Sovereign Credit Default Swap Spreads

Introduction and statement of purpose

In this thesis, I investigate which types of economic news exert the strongest influence on Credit Default Swap (CDS) spreads. Financial market prices are highly sensitive to information, which can take many forms, from expert commentary and news about a country's economic conditions or corporate performance, to traditional macroeconomic indicators such as inflation and unemployment. Although numerous metrics exist to assess a country's economic health, many are released infrequently and are challenging to aggregate. To overcome this limitation, financial markets rely on CDSs, which are priced daily and provide a timely measure of the market's perception of financial stability and credit risk.

In essence, a CDS is a derivate that functions as a form of insurance against the default of a bond[1], whether issued by a corporation or a government. Beyond their role as managing credit risk is used to speculation/interpret macroeconomic conditions. There are three main sorts of CDSs:

- Single-name CDS reflect the perceived default risk of individual firms (can be for example Microsoft and for different time bucket .i.e. one, tow … 30 years).
- Sovereign CDS measure the creditworthiness of entire countries.
- Index CDS (e.g., iTraxx Europe, which tracks 125 of the largest European firms) capture broader market sentiment.

Taken together, CDS markets function as powerful indicators of macroeconomic health. When combined with the flow of economic news, they can provide valuable insights into future economic developments. In this thesis, the focus will be on four broad categories of news macroeconomic, business, political, and financial and standard as these are expected to have the strongest relationships with sovereign CDS spreads.

[1] A Credit Default Swap (CDS) is a financial derivative where the buyer pays a regular premium, known as the CDS spread, to the seller. In return, the seller agrees to compensate the buyer if the underlying reference entity (such as a company or sovereign) defaults or experiences another defined credit event. The payout is intended to cover the buyer's losses, typically based on the difference between the bond's face value and its recovery value after default.

The feature set will consist of conventional financial and macroeconomic variables, including interest rates of various maturities, stock market indices, benchmark treasury yields (the risk free rate), and other key indicators. These traditional variables will form a benchmark model, serving as a reference point for assessing the additional explanatory power contributed by news-based indices. The news indices will be derived from multiple information sources to capture diverse dimensions of economic sentiment and expectations.

Literature review

Avino and Nneji (2014) showed that CDS spreads can be explained or predicted by firm-specific factors, such as assets and liabilities, as well as macroeconomic variables, including the risk-free rate and other yields. Their study employed both linear and nonlinear models to forecast future spreads and demonstrated that these variables can effectively determine CDS spreads. This work provides a foundation for using machine learning and deep learning techniques to identify patterns that may predict future CDS spreads.

More recently, Mao et al. (2023) tested a variety of machine learning and deep learning approaches, with a particular focus on a Merton-LSTM model. They also evaluated alternative models, including standard LSTM networks and Support Vector Machines (SVMs). Their findings indicate that the Merton-LSTM model performs best. This research has paved the way for further development of ML and deep learning methods in the prediction of CDS spreads.

There is already a substantial body of work exploring the use of NLP to enhance the predictive power of economic models or to construct novel economic indices. For example, De Bondt et al. (2025) employed ChatGPT to generate sentiment scores from Purchasing Managers' Index (PMI) commentaries. They found that incorporating these sentiment scores improved GDP prediction models by approximately 20%. Similarly, Silva et al. (2025) leveraged LLMs to convert qualitative central bank statements into quantitative measures of expected economic developments which can be used as signals for the economic development.

Dim et al. (2021) predict CDS spreads for 100 different sovereigns using sentiment analysis of news data. They find that news sentiment has a significant impact on CDS spreads and can also be applied to analyze equity markets and predict credit rating downgrades. Moreover, they show that global news exerts a stronger influence on CDS spreads than local news. The authors also

use topic modeling to categorize news articles into economy, business, security, and two governance topics, although they do not explore how the effects vary across these categories.

Lu et al. (2021) focus on the impact of unexpected macroeconomic news, such as unanticipated changes in GDP, CPI, unemployment, and consumer sentiment. Their findings suggest that market reactions to such news differ in both magnitude and regional patterns across countries.

Data

The dataset will be divided into two main categories: (1) standard variables commonly used to predict CDS spreads, and (2) novel sources of information. Sovereign CDS data will be collected for a selection of countries with available business-day-level observations (hereafter referred to as on a daily level). For most countries the CDS spread will be on a five year tenor but some countries can have multiple tenors (increasing datapoints) and their inclusions will be analyzed.

The dataset will be complemented by other financial features reported on a daily basis, including interest rates of different maturities (e.g., Euribor 3M, Euribor 6M, Libor 3M, Libor 6M), benchmark treasury yields, swap spreads, equity indices (for example the S&P 500), and foreign exchange rates (against USD). These data will be compiled from a combination of free and subscription-based sources, covering a period of approximately 10 to 15 years.

All data will be aggregated on a sovereign basis that is, the above-mentioned variables will be linked to individual countries. The analysis will focus on identifying the average effects across countries rather than producing country-specific CDS spread forecasts in order to answer the research question, i.e. the data will be in the form of panel data.

The daily dataset will be further enhanced with various news- and information-based variables, transformed to a daily frequency. These will include preprocessed news datasets obtained from platforms such as Kaggle, Hugging Face, and GitHub, as well as official communications from central banks (for example, statements from the European Central Bank). In addition, standard macroeconomic indicators such as inflation, unemployment, and GDP growth will be incorporated to provide a broader economic context (so called macro indicators).

Data storage solutions will depend on the overall dataset size. Smaller datasets may be stored locally on a laptop or external hard drive, while larger datasets will be hosted on external storage

platforms such as CSC's computing servers. The data sources can be found on the last page of this paper.

Methods and Metrics

The modeling of CDS spread dynamics will involve experimenting with various models and methodological approaches to identify the best-performing framework or, at a minimum, to establish a robust performance baseline. The process will begin with linear machine learning models, followed by tree-based algorithms and Support Vector Machines (SVMs), and finally progress to deep learning approaches such as Long Short-Term Memory (LSTM) networks. LSTMs are expected to be particularly effective, as the relationships within CDS data are unlikely to be linear due to the complexity of financial market behavior.

The objective of the ML/DL modeling is to develop a robust benchmark model for predicting CDS spreads. This benchmark will provide a reference point for subsequent experiments incorporating the news-based indices. In the benchmark setup, the target variable will be the CDS spread, while the feature set will include lagged values of the CDS itself as well as other relevant daily financial indicators such as interest rates across different maturities, foreign exchange volatility, and risk-free rates.

The target variable in all models will be the CDS spread, while the feature set will include a wide range of variables described previously. A key component will be the inclusion of lagged versions of daily numeric variables, based on the intuition that past values—such as prior CDS spreads—carry predictive information for future movements. Similarly, variables like interest rates may exert delayed effects, as changes on one day may take several days to be reflected in CDS spreads.

Model performance will be evaluated using standard metrics, including Mean Squared Error (MSE), the coefficient of determination (R^2), loss, test accuracy, and other relevant measures. Multiple models will be compared, with hyperparameter tuning conducted through grid search or similar optimization techniques. The data will be divided into training, validation, and test sets to ensure robust and unbiased evaluation of result

A central component of this research is the development of news-based indices that can be incorporated into the predictive models. These indices will draw upon both textual news data (e.g., news articles, research papers, and central bank communications) and traditional

economic indicators (e.g., inflation, unemployment). Each piece of information will be represented in sentiment form positive, negative, or neutral to capture how markets perceive new developments.

Processing the textual data involves two main tasks: categorization and sentiment analysis. Certain data sources, such as central bank statements, will require only sentiment analysis since their category is already known. In contrast, general news datasets, compiled from multiple platforms such as Kaggle, Hugging Face, and GitHub, will first need to be categorized into relevant topics. This categorization will be performed using topic modeling and clustering methods, such as BERT-based sentence embeddings combined with clustering algorithms like DBSCAN or HDBSCAN.

Once clustered, the news items will be assigned to predetermined categories (e.g., macroeconomic, business, financial, political). Subsequently, a pre-trained sentiment analysis model will classify each news item as positive, neutral, or negative. For days with multiple news items within the same category, an average sentiment score will be computed. Days with no relevant news will be assigned a neutral sentiment value. This approach is designed to capture the informational impact of "events" that may correlate with movements in CDS spreads.

Non-textual macroeconomic variables—such as inflation and unemployment—will be transformed into sentiment-based signals to make them conceptually comparable with the textual news variables (macro indicators). This transformation also addresses the challenge of their low reporting frequency. For instance, a decline in inflation would represent a positive sentiment (indicating reduced inflation risk), while an increase in unemployment would represent a negative sentiment (reflecting weaker economic conditions). These transformed values will be combined into a single "macro sentiment" indicator that reflects whether recent macroeconomic data suggest improvement, deterioration, or stability ala Lu et al. (2021) .

Between data release dates, the most recent sentiment value will be carried forward (with carry forward/ exponential Decay) or set to neutral if no new data are available. This creates a continuous daily sentiment series that aligns with the frequency of CDS data. The approach captures how financial markets primarily react to changes in expectations rather than static data points, offering a more behaviorally realistic representation of market sentiment toward evolving economic conditions.

To answer the central research question: which types of news have the strongest influence on CDS spreads the study will employ various feature importance techniques. These include assessing the contribution of news-based features to model accuracy, as well as conducting ablation tests to observe performance changes when specific variables are added or removed, feature importance algorithms, and Shapley Additive exPlanations (SHPA) values. The selection of the methods is key since it will determine the outcome of the study.

Both the machine learning and deep learning components, as well as the text processing pipeline, will require substantial computational capacity. Given the limitations of local hardware, external computational resources will be essential. Access to CSC's Puhti supercomputing environment is expected to play a critical role in enabling large-scale data processing, model training, and optimization.

Timetable

The plan is as follows:

- September – October: Develop a well-defined and focused thesis, finish the research plan, obtain supervisor, gather the necessary data, and determine the models and methods to be applied.
- November – December: Produce initial results in the form of a preliminary model and begin drafting the report.
- January – March: Write the majority of the thesis, aiming to have a near-complete draft by the end of this period.
- April: Focus primarily on refining, polishing, and finalizing the thesis.

Deliverables

The primary deliverable of this thesis will be a written report presenting the key findings and analyses. In addition, all code used for data collection, preprocessing, and modeling will be made available, along with the final predictive model for CDS spreads and the constructed macroeconomic and news-based indices.

The success criteria will differ across the two main objectives of the thesis. The first criterion concerns the predictive performance of the final model specifically, how accurately it can estimate CDS spreads for individual sovereigns. The second, and more important, criterion focuses on interpretability: identifying which types of news, information, or events exhibit the strongest relationship with a country's CDS spreads. Insights from this analysis will not only enhance understanding of market behavior but also guide future researchers and investors in prioritizing the most relevant informational inputs for similar predictive models.

References

Avino, D., & Nneji, O. (2014). Are CDS spreads predictable? An analysis of linear and non linear forecasting models. International Review of Financial Analysis, 34, 262-274.

de Bondt, G. J., & Sun, Y. (2025). *Enhancing GDP nowcasts with ChatGPT: A novel application of PMI news releases* (No. 3063). ECB Working Paper.

Dim, C., Koerner, K., Wolski, M., & Zwart, S. (2021). Hot off the press: News-implied sovereign default risk. *Available at SSRN 3955052*.

Mao, W., Zhu, H., Wu, H., Lu, Y., & Wang, H. (2023). Forecasting and trading credit default swap indices using a deep learning model integrating Merton and LSTMs. Expert Systems with Applications, 213, 119012.

Kirtac, K., & Germano, G. (2024). Sentiment trading with large language models. Finance Research Letters, 62, 105227.

Lu, M., Passariello, M., & Wang, X. (2021). Sovereign CDS Premiums' Reaction to Macroeconomic News: An Empirical Investigation. Complexity, 2021(1), 5568698.

Silva, T. C., Moriya, K., & Veyrune, R. M. (2025). From Text to Quantified Insights: A Large-Scale LLM Analysis of Central Bank Communication (No. 2025/109). International Monetary Fund.

Data sources:

ECB speeches: https://www.kaggle.com/datasets/robertolofaro/ecb-speeches-1997-to-20191122-frequencies-dm

Financial news data: https://github.com/Webhose/financial-news-dataset

Global central bank speeches: https://github.com/DRomelli/cbspeeches

CDS spread data: https://www.investing.com/rates-bonds/germany-cds-10-year-usd