EUROPEAN CENTRAL BANK

BANKING SUPERVISION

Cassandra

Predicting bank's distress using AI and NLP on financial news

27/06/2025

DG HOL-STE

A. Kouris, J. Arranz Barquero, L. Parafioriti, A. Petropoulos

Overview

1 About Cassandra

2 Cassandra workflow overview

3 Sample and distress events, data sources

4 Large Language Models – Agents, tools and feature creation

- 4.1 Named Entity Recognition – mapping news to banks
- 4.2 Dropping non-relevant and duplicate news – similarity algorithm
- 4.3 Feature creation – agent types

5 Modelling setup and forecasting results

About Cassandra

Cassandra, from the Greek mythology, was given the gift of prophecy

Cassandra is an **AI tool** that uses **Natural Language Processing (NLP)** and **Machine Learning (ML)** to **forecast the probability of distress** of banks using **only text data** coming from financial news.

Value Proposition

- Provide an **early warning system tool** for the SSM using new complex and **advanced technologies** like AI and ML
- **Automatic process** that can be implemented in **a high frequency** manner (daily, weekly, etc.)

Goals

- **Warn supervisors** about distress events giving enough time to act
- Recurrent reporting tool to understand the risk profile of a bank and its evolution over time

Cassandra employs **cutting-edge machine learning methods** to predict bank distress events with an **accuracy rate of 85%**

Cassandra's Workflow

1 Data

Data collected

15M news
2007-2023

REUTERS · benzinga · DOW JONES FACTIVA · Bloomberg · S&P Global Market Intelligence

Bank-level
Central Banks

Banks in Geo
Country-level

+400k
Relevant news
Mapped to the banks

Bank Sample

191
Banks

94
Distress events

2 Sentiment

We use **AI and own-developed neural networks** to know if the news are **positive, negative or neutral** for the bank

Cleaned news

Sentiment analysis

MistralAI · BERT · FINANCIAL BERT · LLaMA by Meta · prosus

News with sentiment

3 Modelling

News with sentiment

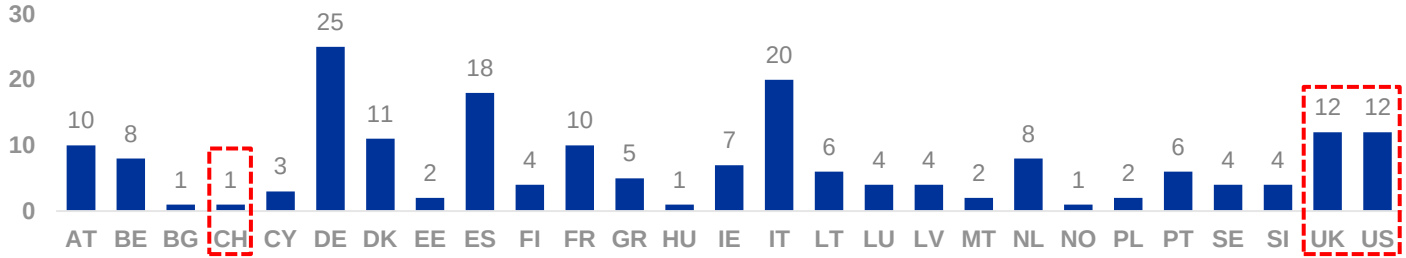
Prediction of probability of distress in the next 2 months

Use of AI *Use of ML*

Banks – sample and distress events

Cassandra

Sample



191 Banks

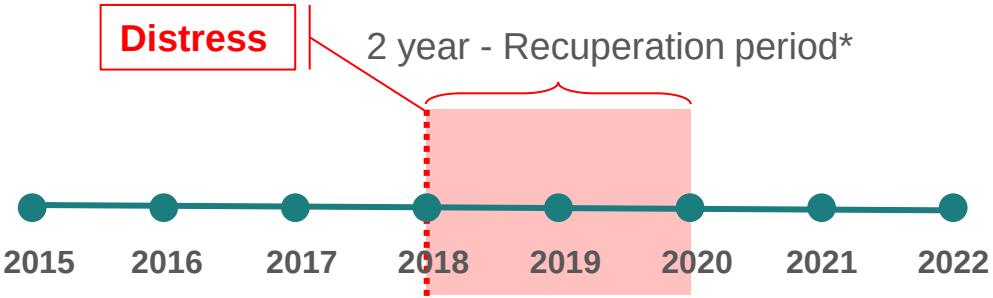
98 Supervised

26 Non-EU

Distress Events

94
Events

- Mergers
- Capital Injection
- Bankruptcy
- State aid / Bailout



** Banks' observations during the recuperation period are dropped from the analysis*

Large Language Models – LLM

Understanding LLM

What?

Artificial Intelligence models

Why?

Understand and **generate** human-like language

How?

Trained on **massive datasets**, allowing to **understand** context, language and topics

TYPES

Agents

BERT

ChatGPT

MistralAI

LLaMA by Meta

XLNET

prosus

Tools

Text classification

Content creation

Image generation

Named Entity Recognition (NER)

Writing assistants

Machine Learning

Data automation

Data extraction

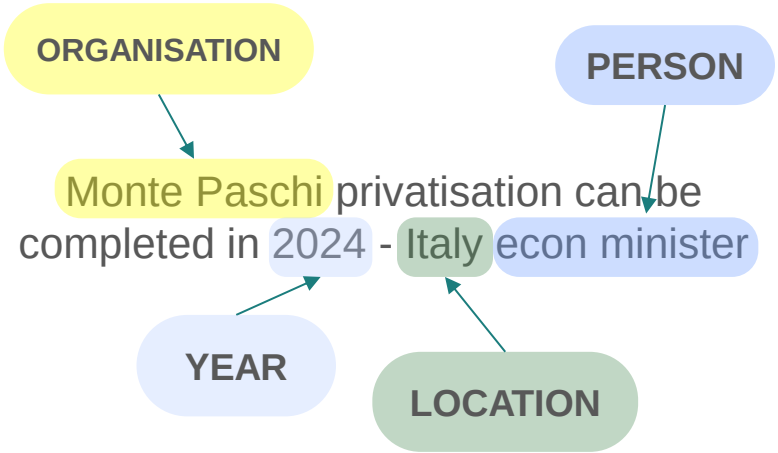
Prompt Template

Mapping news to banks - NER

AI technology to categorise words within a text

Cassandra

Named Entity Recognition (NER)



316,000 Unique News

275,000	Bank-specific
27,000	Country-specific
10,000	Geographical Banking
4,000	Central Banks*

+15M news

+400k news

Dropping non-relevant and duplicate news

Non-relevant news

Some news report the bank modifying ratings or price targets

"BofA downgrades cyrusone to neutral"

The headline does not contain relevant information

"BBVA 2019 q2 - results - earnings call"

Rule-based N-gram analysis for pattern detection

- N-grams are a **sequence of n tokens** from a string
- Help to **identify patterns** in text
- News containing that pattern are removed

1-gram Analysts at UBS raise Microsoft price target

2-gram Analysts at UBS raise Microsoft price target

3-gram Analysts at UBS raise Microsoft price target

News reported by several sources

REUTERS

Bloomberg

DOW JONES FACTIVA

x3

Same information

Overestimation of the issue

UBS

18 Sept 2023

AI-powered similarity algorithm to drop duplicated news

News 1:

Swiss financial regulator could increase team supervising UBS

News 2:

Swiss financial regulator: likely it will increase size of team overseeing UBS in future

T5

Similarity score

5 Same meaning

4.9

1 Not similar

Keep longest

Cassandra

Feature creation workflow

1. Pre-process

Processed Data

Headline

Headline

Headline

Agents for sentiment analysis

2. Sentiment Generation

28 agents

Pure-sentiment agents

- Each agent gives a score to each headline
- The score can be discreet or continuous, but always **bounded between -1 and +1**

1 Dictionary-based
- AFINN
- Loughran
- Vader

2 Self-developed agents – Deep learning
- **CNN** *Convolutional neural networks*
- **LSTM** *Long short-term memory neural network*

3 Pre-trained agents

Prompt-engineering

NVIDIA CUDA

General knowledge: MistralAI, BERT, LLaMA by Meta, prosus

Financial texts, notes and reports: FINANCIAL BERT, FINANCIAL prosus

Weighted-agents

- The sentiment is weighted to create extra **agents with a specific focus**

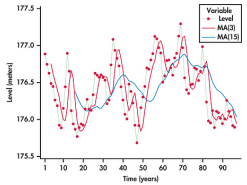
4 Regulatory agents
News' sentiments with high regulatory relevance are given more importance

5 Risk-weighted agents
Zero-shot classification of news into risks categories and weighted accordingly Meta

6 Ensemble agents

3. Features

Moving averages



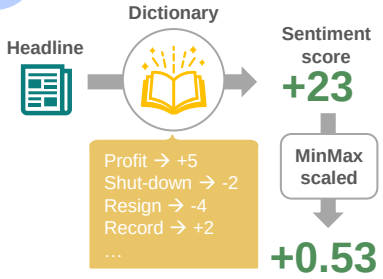
Dynamic Counts

Variable transformations

+500 features

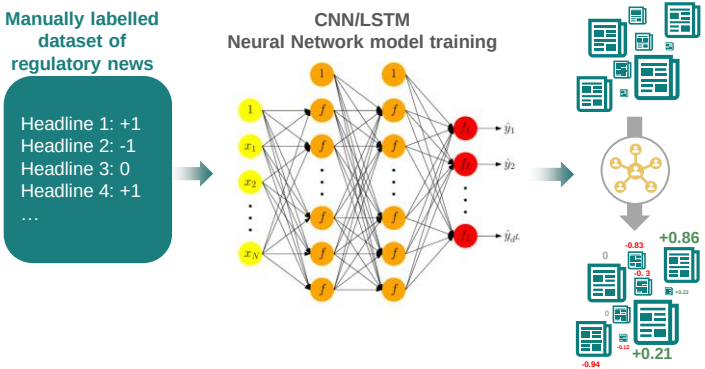
Pure-sentiment agents

1 Dictionary-based



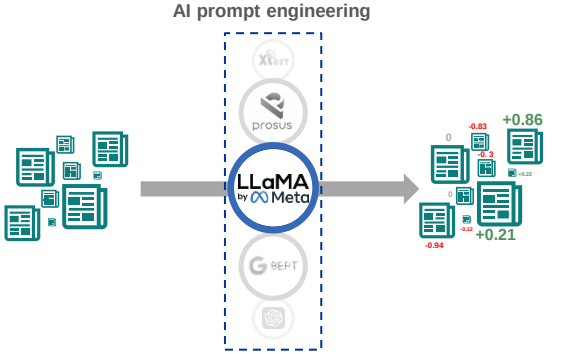
- ✓ No training data is required
- ✓ Less expensive implementation
- ✗ Sentiment is strict and domain-dependent
- ✗ Context is not considered

2 Self-developed agents – Deep learning



- ✓ Training data with a focus on regulatory news
- ✓ Effective and accurate
- ✓ Highly flexible
- ✗ Time-intensive training
- ✗ Complex dynamics

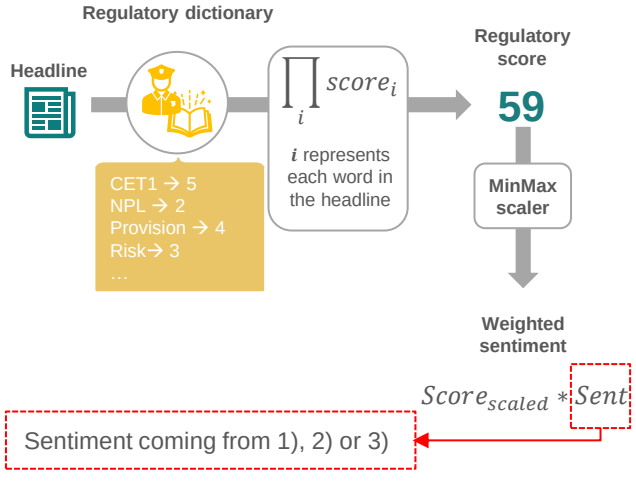
3 Pre-trained agents



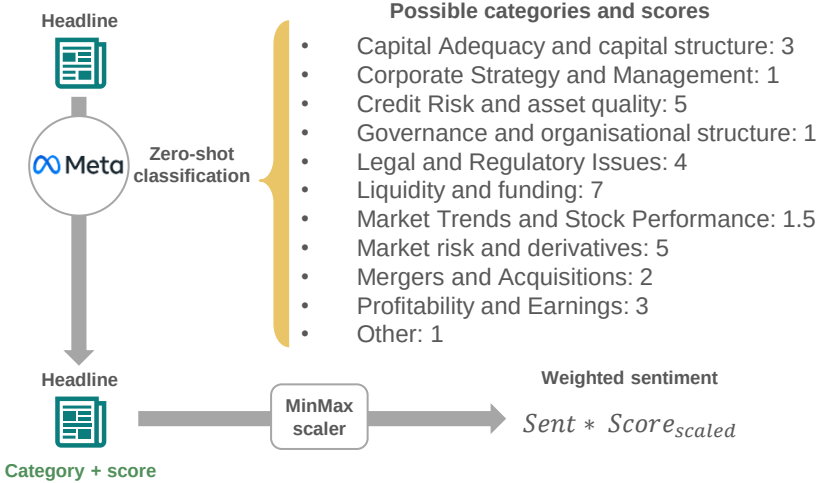
- ✓ Most effective sentiment
- ✓ Agents are trained with billions of data points
- ✗ Computationally-intensive
- ✗ Prompt-dependent

Weighted agents

4 Regulatory agents

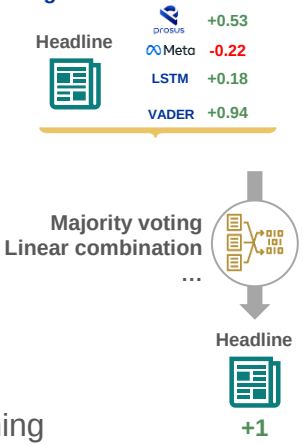


5 Risk weighted agents



6 Ensemble agents

- Agents constructed as **combinations, permutations and transformations of other agents**



- Gives **more importance to news** with topics that directly impact our regulatory interests, either by containing regulatory words in its headline or by being in one specific risk category
- Additionally, it serves as an extra-feature, providing a **regulatory score and a classification**

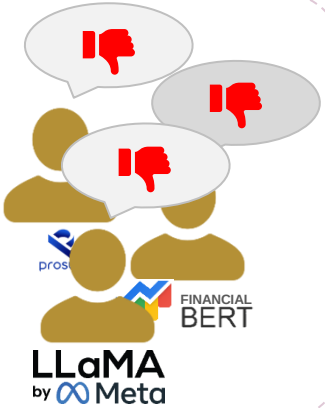
The Importance of own-developed agents

Headline

"Eurobank's NPEs fell by eur 3.7 billion"

Sentiment Generation

Pre-trained and dictionary-based agents



Self-developed and weighted agents

Findings

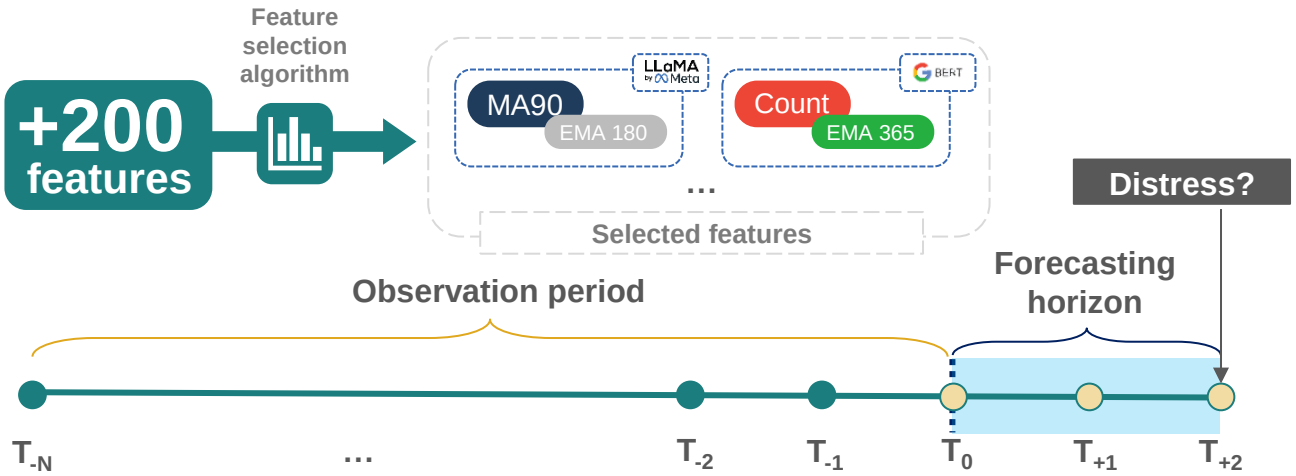
- **Pre-trained and dictionary-based** agents are **prone to error** when presented with **highly specialized news**. The word 'fell' is interpreted as something negative, missing some technical context.
- Our **own-developed agents can capture those technicalities**.
- Weighted agents are able to mitigate that wrong sentiment.

Modelling Setup

Key characteristics

- **2-month** forecasting window
- **Flexible** observation period
- Best features selected using ML

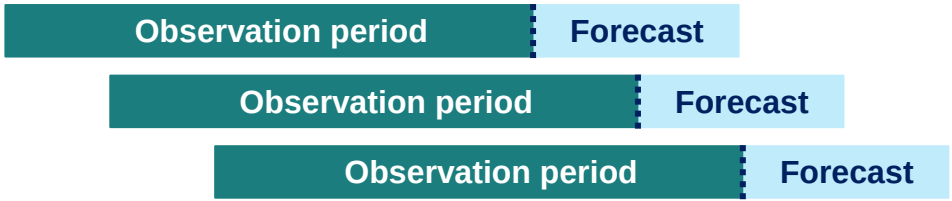
SET UP



EXECUTION

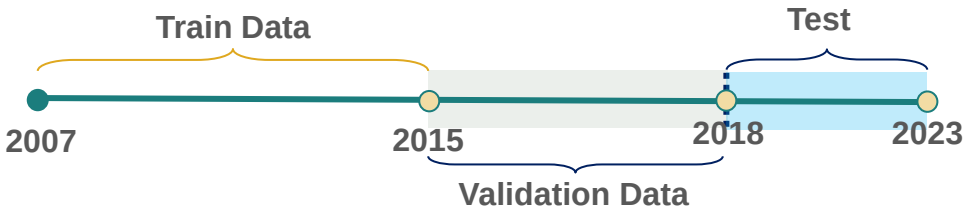
Frequency of the forecast

- Daily
- Weekly
- Monthly
- …



Dataset + SMOTE

Data split and transformations

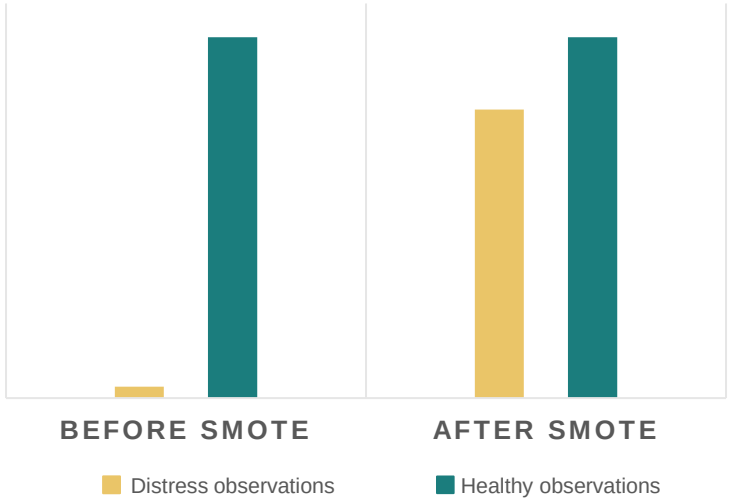


We address potential imbalances in the training data to improve model performance and fairness

- **Scaling** : standardize the features by removing the mean and scaling to unit variance.
- **Oversampling :** techniques to address imbalance in the training dataset specifically we apply Borderline SMOTE to generate a new training set that balances the classes
- **Fill missing values:** ensuring the data is suitable for input into machine learning models without introducing errors from missing data

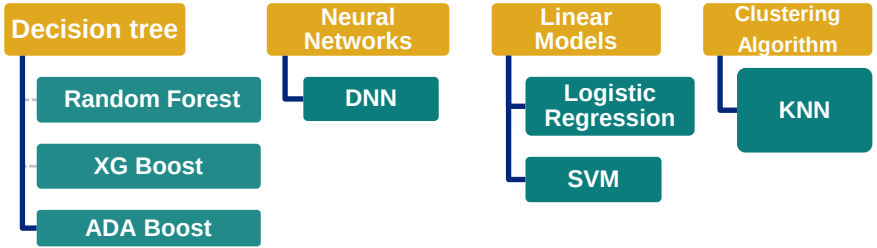
Fixing imbalances - SMOTE

- Generates **synthetic samples** by interpolating between existing minority class samples to **balance the data in the training sample**.



Model Development

Machine Learning Techniques Explored

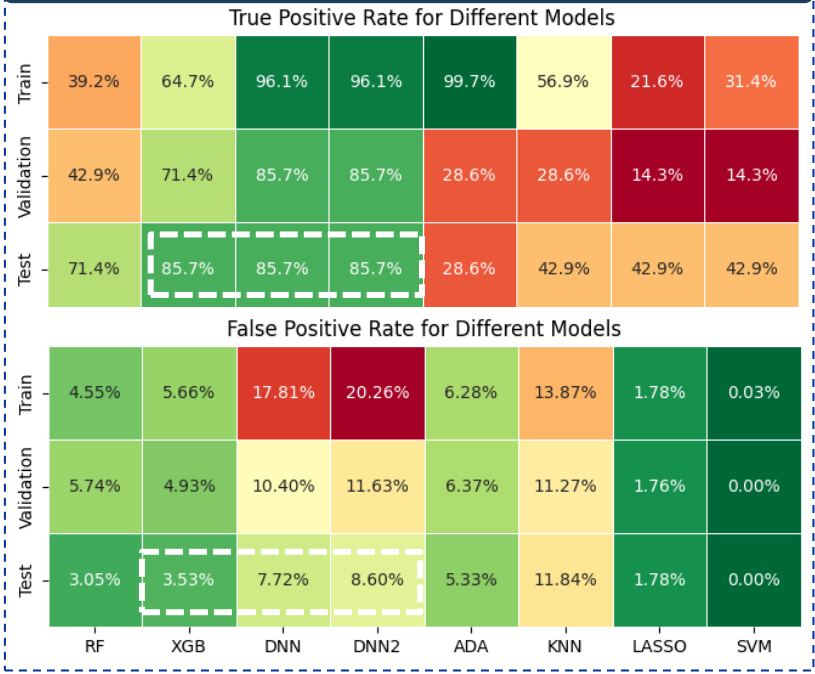


- Conducted a **deep analysis** of various machine learning models
- **XGBoost and DNN** are identified as optimal choices with **higher discriminatory power** indicating **superior predictive capabilities**
- Given the combination of **high TPR and low FPR** both models stand out as **robust models for predicting distress events**

Discriminatory Power Evaluation

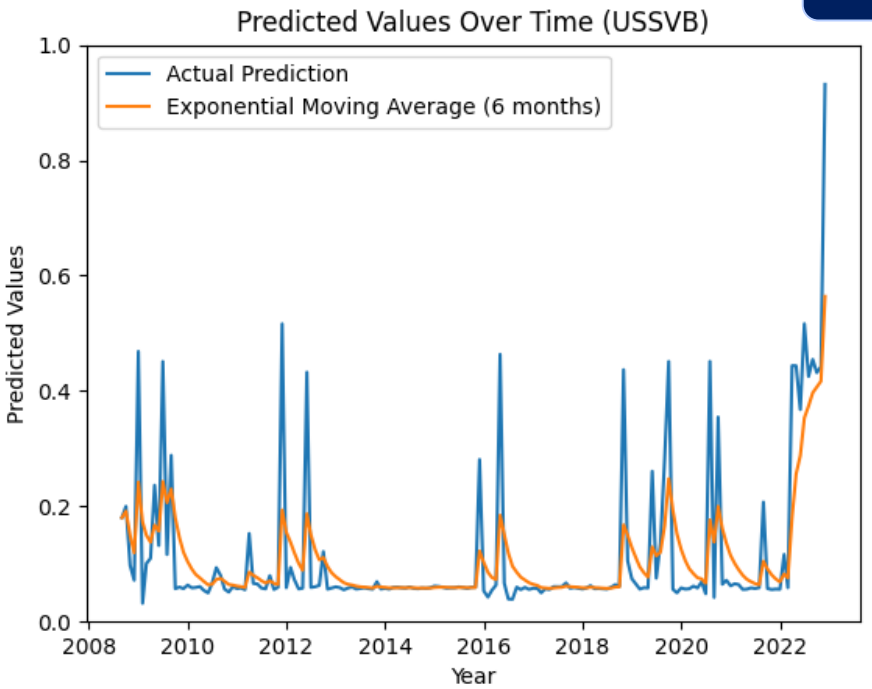
AUC	RF	XGB	DNN	DNN2	ADA	KNN	LASSO	SVM
Train	0.74375	0.90626	0.89213	0.88102	0.99819	0.84972	0.74762	0.75993
Validation	0.80450	0.94232	0.90266	0.94296	0.71422	0.78226	0.63180	0.74998
Test	0.84753	0.90821	0.92045	0.91696	0.71553	0.65988	0.73399	0.73836

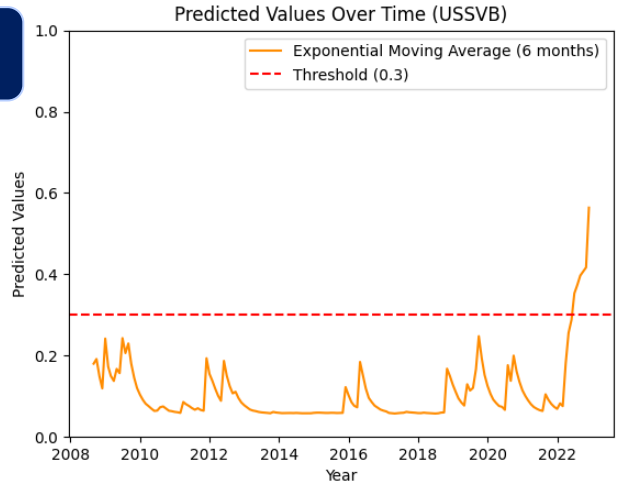
Forecasting Accuracy Evaluation



Case Studies

Silicon Valley Bank

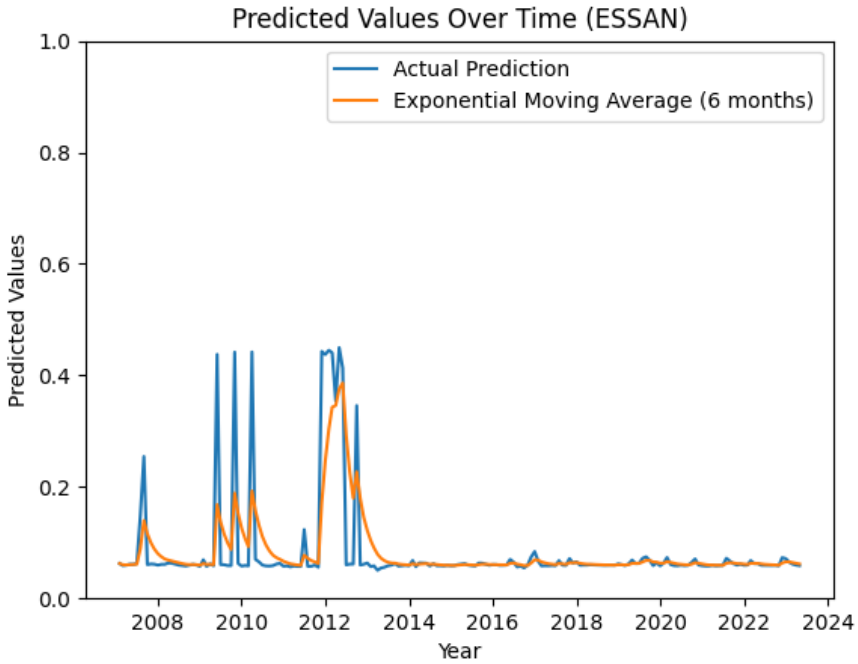


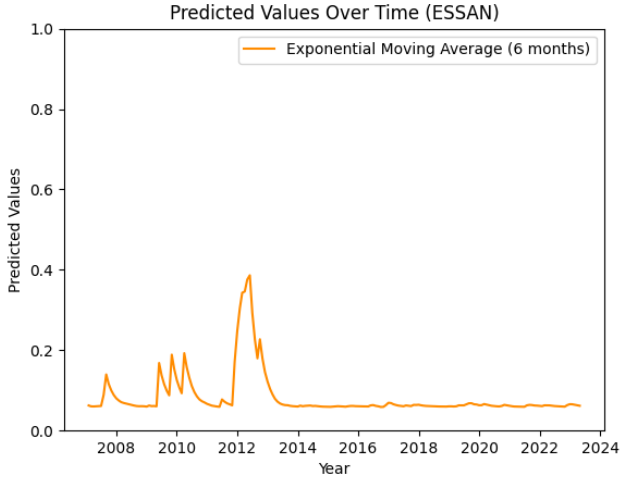


- The probability of distress jumps in specific months as more agent hit a threshold of either strong negative or positive sentiment.
- Both models also perform significant denoising of each agents separately
- **Models capture the collapse of Silicon Valley (USSVB) taking place on March 2023**

Case Studies

Banco Santander





Thank you for your attention.
We encourage to ask any questions.

Cassandra: an automated AI-based surveillance system to forecast bank distress from financial news

Anastasios Petropoulos, Alexandros Kouris, Javier Arranz Barquero, Christoffer Kok, Luca Parafioriti

Anastasios.petropoulos@ecb.europa.eu
Alexandros.kouris@ecb.europa.eu
Javier.arranz_barquero@ecb.europa.eu
Christoffer.kok@ecb.europa.eu
Luca.parafioriti@ecb.europa.eu

Abstract

This article introduces *Cassandra*, an AI-driven automated surveillance system designed to predict bank distress using solely financial news, enabling more timely and proactive supervision by drawing on a broader, real-time data source, never exploited before to this purpose. Unlike traditional approaches to forecast bank distress, *Cassandra* overcomes the endogeneity and time-lags inherent in balance sheet-based indicators by accounting for the financial sector's overall perception of the entity derived exclusively from financial news. To process such qualitative unstructured data, *Cassandra*'s infrastructure incorporates advanced NLP techniques, such as named entity recognition (NER) and sentiment analysis, in a novel application. The system includes a heterogeneous array of general and financial sentiments obtained via large language models (LLMs). Building on this foundation and marking an innovation in this field, we build a set of regulatory-focused sentiment analysis models. This comprehensive architecture forms an automated pipeline for banking supervision applications, serving as a proof-of-concept early-warning system for potential distress events with an out-of-time tested hit rate above 80%.

Keywords: Natural Language Processing, Sentiment Analysis, LLM, Forecasting, Bank Distress, XGBOOST, Deep Neural Network

---

1) Introduction

Early warning tools are key for supervisory institutions in the banking sector. They help regulators and supervisors identify potential risks and vulnerabilities in banks before they materialise and cause financial instability. By detecting sings of potential distress early on, supervisors can act in anticipation, highlighting weaknesses before they manifest in financial statements or regulatory filings.

Historically, banking supervision has been relying on traditional bank income statement analysis and quantitative market data, usually only available with a quarterly frequency. Departing from traditional approaches, we aim to provide a complimentary view to those already existing by leveraging on strictly textual and real-time data by extracting insights from financial news and employing advanced state-of-the-art technologies to predict bank's distress. Our motivation is to help take banking supervision to the next technological level to facilitate more automatic, real-time and proactive forward-looking assessment.

The understanding of increasingly rapid and continuous flows of information, news and market data is key to understand risks facing the banking sector. In today's digital age, real-time access to news has become an essential tool for monitoring market conditions, assessing risks, and identifying emerging vulnerabilities, as they capture market participants' perception and public sentiments. However, processing and understanding these data poses three main challenges. First, the already-existing and newly-created volume of data impedes the correct processing of the information it contains hastily enough to act and prevent the materlization of the risk (also known as the dimensionality problem). Second, disecting the signal from the noise among the vast amounts of data to develop key insights is resource-intensive and prone to error. Third, the ability to predict potential distresses at individual banks is key to allow supervisors to act.

To tackle the first and second challenges, we leverage on the usage of Natural Language Processing (NLP) techniques. Modern NLP models, particularly those powered by deep learning and large language models (LLMs), are now capable of cleaning, labeling, classifying, and extracting intricate patterns and sentiment from vast datasets with remarkable accuracy. Sentiment analysis, a key application of NLP, allows the automated extraction of opinions and emotions expressed in text, making it possible to classify the sentiment of news reports as positive, negative, or neutral. When applied to financial news, these advanced techniques can provide valuable insights by linking the sentiment of news to specific banks, allowing for a nuanced understanding of market sentiment and its potential impact on financial institutions. The ability to process and analyse sentiment across a vast universe of news in real-time has the potential to revolutionize how bank supervisors and regulators monitor the health of financial institutions.

In conjunction with NLP, Machine Learning (ML) techniques provide a powerful toolkit for transforming news sentiment into actionable insights. By leveraging state-of-the-art ML algorithms, it is possible to develop predictive models that associate time-series data on predefined distress events with the sentiment extracted from news reports to warn supervisors with enough time to take action (thus tackling the third challenge). These models can be trained to detect patterns and correlations between shifts in sentiment and subsequent bank distress events, enabling the creation of early-warning systems. Such systems could offer supervisors and regulators an unprecedented ability to forecast potential risks and preemptively address vulnerabilities in the banking system, enhancing the overall stability of the financial sector.

For this purpose, we develop a dataset of daily financial news by gathering, processing and cleaning news from publicly available databases and internet archives. In total, we obtained more than 15 million news articles covering a total of 18 years spanning from 2007 until today. Each piece of news is linked to one (or several) institutions from our sample of 191 banks by using a novel AI tool, Named Entity Recognition (NER).

We then generate a set of different and independent sentiment scores for each news leveraging on traditional NLP sentiment analysis and introducing a novel framework by using and developing custom AI agents to provide sentiment classifications specifically tailored to the supervisory and regulatory domains. These agents are based on distinct perspectives and methodologies to ensure we capture the complexities of sentiment generation. The set comprises agents trained on general purpose text, regulatory documentation, or financial text. Additionally, these agents are combined and weighted based on different themes and domain-specific risk categories by leveraging on state-of-the-art AI tools, such as zero-shot-classification. Consequently, we apply advanced machine learning techniques and models to link the sentiment obtained from the agents in the previous step with the probability of distress.

This approach represents a significant innovation in the field of banking supervision in several perspectives. First, we compile a very large and comprehensive news dataset. Second, we introduce the combination of sophisticated methods, tools

and infrastructures to develop a unique signal of bank's distress, which is produced daily (instead of the usual quarterly financial reporting). Third, our approach uses exclusively textual data, departing from the more traditional literature of assessing risks based on financial ratios. Fourth, while there is existing research on bank's distress in the United States (US), we focus on European Union (EU) based institutions, primarily but not limited to those supervised by the European Central Bank (ECB).

2) Literature Review

The work presented here lies at the intersection of three key domains: banking supervision, particularly studies on bank distress events; natural language processing (NLP), focusing on named entity recognition (NER) and sentiment analysis (SA) via large language models (LLMs); and machine learning (ML). This comprehensive body of research forms the basis for ongoing innovation in banking supervision, where the integration of machine learning and NLP techniques holds the potential for developing robust automated early warning systems that can more effectively prevent banking crises.

Forecasting financial distress, including banking crises and insolvencies, has been a critical area of research, motivated by the need to understand and prevent such events, ultimately contributing to the development of early warning systems in banking supervision. In this context, (Demyanyk & Iftekhar, 2010) and (SirElkhatim & Salim, 2015) offer a comprehensive review of empirical studies that aim to explain, predict, or propose solutions for financial crises and banking defaults, as well as an overview of the methodologies employed in these analyses. A significant portion of the literature on bank failure prediction, e.g., (Wheelock & Wilson, 2000), (Cole & White, 2012), (Betz, Oprică, Peltonen, & Sarlin, 2014), centres around the use of supervisory CAMELS indicators – an acronym for Capital, Asset Quality, Management, Earnings, Liquidity, and Sensitivity to market risk indicators. Although these indicators are widely utilized by both investors and regulators to assess the financial stability of banking institutions, there is inconclusive evidence on which variables are important in predicting bank insolvencies. For instance, (Mayes & Stemmel, 2014) claim that a simple leverage ratio (unweighted) is a better predictor than the capital adequacy ratio (risk weighted), while (Poghosyan & Cihak, 2009) show that indicators related to capitalization, asset quality and profitability can effectively identify weak banks. Furthermore, and more importantly, these quantitative indicators are limited to capturing data related to bank balance sheets at fixed intervals. This approach raises two key limitations. First, the endogeneity of the source lacks a broader perspective able to comprehend market insights, network effects and the financial sector's overall perception of the entity. Second, as this data is typically collected only on a monthly or even quarterly basis, the resulting time lag introduces delays in forecasts, hindering timely surveillance and undermining the objective of proactive bank supervision. In addition, as highlighted by (Betz, Oprică, Peltonen, & Sarlin, 2014), much of the literature examining individual bank failures or distress events focuses predominantly on U.S. banks or cross-country panels, with relatively few studies dedicated to European banks. In light of these considerations, this paper diverges from the traditional literature and introduces an innovative approach. By applying cutting-edge machine learning techniques to a dataset focused primarily on European significant institutions (SIs), *Cassandra* goes beyond conventional financial metrics relying exclusively on unstructured textual information. As a matter of fact, financial news constitutes a source of information which is readily available, current and rich in descriptive detail. In contrast to traditional information sources, text data offers a possible route to circumvent the issues of privileged access, lagging publication and low granularity, but most importantly it very directly offers value through the explanatory power of the event-related human language descriptions accompanying the plane signal (Rönnqvist & Sarlin, 2017). In this work, we extract, structure, and perform NLP tasks, such as NER and SA, on news and other textual sources, offering an innovative perspective on bank distress prediction, able to integrate qualitative insights. This represents a significant advancement in the domain of early-warning systems for banking supervision.

As a matter of fact, the use of NLP in financial contexts, particularly NER, SA and the use of LLMs, has seen rapid development due to its ability to extract market sentiment from unstructured textual data. In our approach, we first exploit the use of NLP techniques to include NER and article classification, crucial for filtering relevant information from a vast corpus of financial news. NER models have evolved from rule-based to conditional random field (CRF) models, and more recently to transformer-based models utilizing self-attention mechanisms. Each model type has its strengths and is chosen based on the specific NER task at hand (Pakhale, 2023). This evolution reflects the increasing complexity and capability of NER models in processing and analysing financial texts, as presented by (Wang, Xu, Liu, Gui, & Zhou, 2014) and (Shah, Vithani, Gullapalli, & Chava, 2023). NER has numerous applications in finance, such as sentiment analysis of news to rank financial events, automating stock trading, and creating financial knowledge graphs. These applications demonstrate the broad utility of NER in extracting and utilizing information from financial texts for various analytical purposes. A further application of NLP in *Cassandra* is the SA task. A seminal contribution in this space is the work of (Loughran & McDonald, 2011), who created a specialized financial sentiment dictionary to improve SA in financial texts. General sentiment lexicons, widely used in NLP, were found inadequate for financial contexts, necessitating this domain-specific tool to capture market sentiment more

accurately. Another key work is that of (Nassirtoussi, Aghabozorgi, Wah, & Ngo, 2014), who applied SA techniques to news articles to predict stock market movements. Their research demonstrated that sentiments derived from real-time financial news articles can significantly impact market forecasts, providing a direct link between sentiment analysis and financial decision-making. Motivated by the current literature that underscores the efficacy of SA in financial contexts, e.g., (Jiang & Zeng, 2023), the infrastructure described in this article first performs a NER task which leads to the construction of a bank-specific dataset of news. Then, the system applies LLMs like BERT and its derivatives (e.g., FinBERT, Financial BERT, Financial RoBERTa, DistilRoBERTa), as well as Llama 13b, Mistral, XDistil, Financial SIGMA to perform SA with the most comprehensive approach. In particular, *Cassandra* reads these LLMs as distinct individual agents expressing a sentiment on each news article associated to bank entity. In addition, following a different venue from the existing literature, the architecture comprehends three key innovations:

- self-developed deep-learning agents: these agents are implemented by training Long Short-term Memory (LSTM) and Convolutional Neural Network (CNN) models on a predefined set of supervisory labelled news. The creation of these agents allows the system to refine sentiment classification tailored to financial and regulatory contexts.
- regulatory agents: these agents are derived from the other agents by introducing a weighting lexicon-based scheme created from regulatory documentations. This enables sentiment analysis specific to compliance and supervisory concerns.
- risk-weighted agents: these agents categorize news sentiment through a zero-shot classification approach based on various business thematic sectors. This leads to the incorporation of domain-specific risk profiles.

To our knowledge, such types of agents have not been implemented before for sentiment analysis, yet they constitute a relevant source of information for financial and systemic risk analytics, able to express the financial environment perception from a more tailored perspective, closer to the field of financial regulations and banking supervision.

The sentiments extracted are then harmonised and linked into a general dataset which connects banks, news and sentiments over time. Such a structured dataset allows to derive several features from the sentiments produced, obtaining panel data of time-series at bank-level with a daily frequency. This dataset, paired with a collection of distress events by bank, perfectly sets the ground for the application of ML techniques for the development of a model to forecast the probability of distress of a bank from the features computed from the generated sentiments. A useful analysis related to this specific part of *Cassandra* is described in the article of (De Jesus & Cássio da Nóbrega, 2023) that, in the context of publicly traded banks in Brazil, show the forecasting capabilities of different machine learning models applied to financial ratios and sentiments, which, however, are limited to the lexicon-based type. Although it focuses exclusively on quantitative indicators, the work of (Climent, Momparler, & Carmona, 2019) demonstrates the adaptability and accuracy of the extreme gradient boosting technique applied to predict and prevent bank failure in the Eurozone banking sector. This algorithm is compared to others in the article of (Petropoulos, Siakoulis, Stavroulakis, & Vlachogiannakis, 2020), which offers a comparative analysis of the most applied machine learning models for predicting bank insolvencies, further illustrating the evolution of these techniques in recent years. Inspired by this literature, we did not limit ourselves to specific models, rather we adopted a comprehensive approach, applying a number of different ML techniques such as random forest, extreme gradient boosting, adaptive boosting, deep neural networks, logistic regression, support vector machines and K-nearest neighbours. After implementing a robust testing procedure based on out-of-sample data splitting, we collected the key performance metrics for such a binary classification problem (recall, accuracy, ROC curve and AUC, Gini coefficient, Kolmogorov-Smirnov statistics, balanced accuracy). On the basis of considerations and observations on these statistics, taking also into account the model´s interpretability – by means of SHAP values charts and partial dependence plots – we were able to select an ensembled model, combining both the extreme gradient boosting and the deep neural network approaches, as they showed similar metrics (overperforming the other models) albeit based on different important features.

To conclude, the intersection of forecasting bank distress, NLP (with a focus on LLMs for SA and NER), and machine learning has produced significant advancements in predictive capabilities. While traditional econometric models laid the foundation for financial crisis prediction, the rise of machine learning, deep learning, and NLP, currently driven by LLMs, has allowed for more nuanced and accurate forecasting models. The literature reflects a growing trend toward leveraging textual data through NER and sentiment analysis, while the application of machine learning techniques continues to enhance predictive accuracy in identifying bank distress events. On this ground, *Cassandra* can be seen as a pioneering proof of concept in the domain of early-warning surveillance systems as it presents an automated pipeline constituted of innovative applications in each of these three areas.

3) Data collection and processing

a) Overview and workflow

Our data workflow is shown in Figure 1. We first collect unstructured data and process it in order to have a homogeneous data set of financial news. Then, we classify them into business or non-business-related news, followed by the specification of the bank sample and the linkage between each news and the bank it refers to. After that, we clean further the data to avoid non-relevant and opinionated news, as well as potentially duplications. With the addition of our proprietary dataset of distress events, we arrive to the final data set.

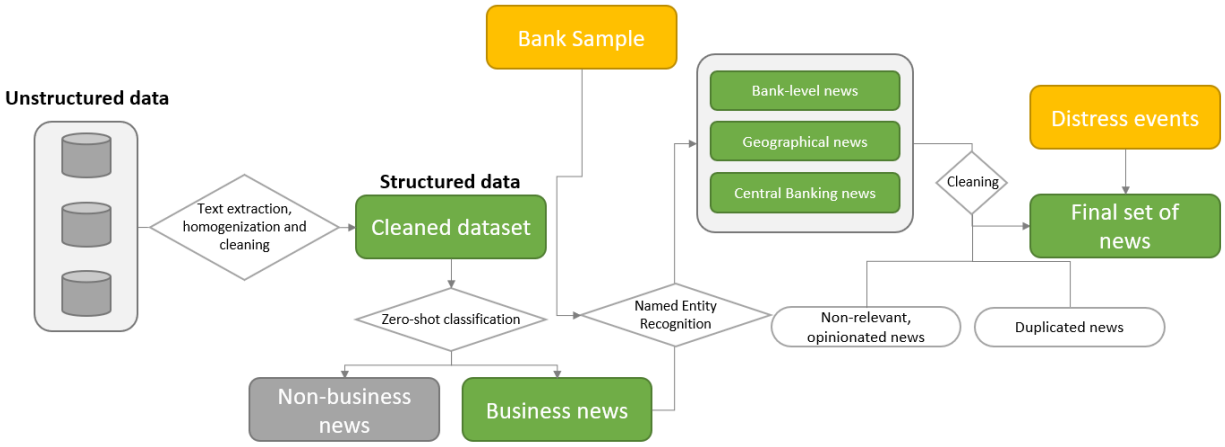


Figure 1: General workflow - Data collection and processing

b) Data collection, sample and distress events

For the purpose of this analysis, it was crucial to concentrate on a representative sample of banks to ensure both (1) ample news coverage and (2) a sufficient number of distress events as per a predefined criterion. To achieve this, we expanded our sample to include banks under the supervision of the European Central Bank (ECB), as well as banks of comparable significance closely financially-linked to the European area. As a result of this approach, as explained in the following paragraphs, we obtained a sample of 191 banks with over 15 million news articles covering 94 distress events. This, however, resulted in a still highly unbalanced dataset with a relatively small amount of news covering the distress periods and a majority of news linked to 'healthy' banks. This, in turn, posed a significant challenge in the modelling of the distress probabilities.

The dataset consists of an initial coverage of 191 banks, as shown in Figure 2. Within this set, 98 banks fall under the supervision of the European Central Bank (ECB) and are consequently located within the European Union (EU). In order to enhance the sample's comprehensiveness, we included an additional 26 banks from non-EU regions, specifically from the United Kingdom (UK), the United States (US), and also Switzerland (CH) and Norway (NO).

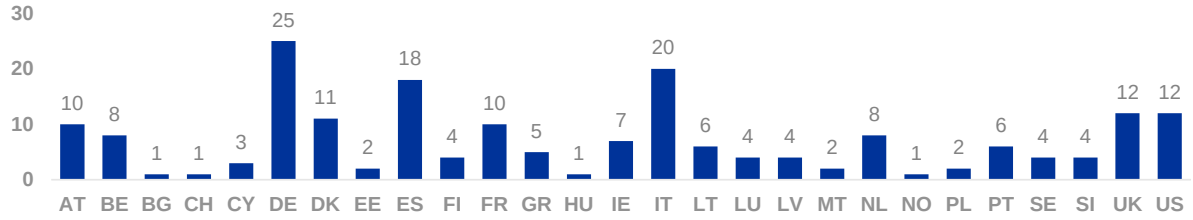


Figure 2: Number of banks by geography in the sample

To identify non-European Union (EU) banks for inclusion in our study, we employed data retrieved from financial databases (S&P Capital IQ and Moody´s Analytics) and recent banking distress events (the collapse of Switzerland´s Credit Suisse or the failure of Silicon Valley Bank in early 2023). A bank was deemed relevant if its total assets on the balance sheet aligned with

those overseen by the European Central Bank (ECB) – therefore, usually more than €30 bn, as per the size criteria for significant banks.

Instances of bank distress, characterized by defaults, are infrequent occurrences as financial institutions are typically either bailed in/out by governmental authorities or acquired by other institutions before reaching such a state. Moreover, the scholarly literature lacks consensus on the definition of distress. While some researchers quantitatively define distress based on the evaluation of financial ratios derived from the bank's balance sheet, others opt for a qualitative definition.

Our conceptualization of distress encompasses the notion of negative and significant events within a bank, aligned in scope with that of the supervisory perspective.

These events may be financial in nature, such as :

- bail in/out,
- receipt of state aid,
- capital or liquidity injections,
- nationalizations,
- tail-like operating losses, or
- outright defaults.

Or qualitative:

- severe legal issues, such as money laundering
- mergers and acquisitions

Based on these criteria, we successfully identified a total of 94 instances of distress events. A breakdown of the identified reasons of distress and their corresponding occurences can be found in Table 1. The predominant category among these events is bailouts, occurring in 28 instances. Following this, instances of nationalization and state aid transpired 10 times over the observed time span, closely trailed by occurrences of capital injections.

Reason of distress	Occurrences
Bailout	28
Nationalised	10
State aid	10
Acquired	10
Capital Injection	9
Merger	7
Bankrupt	7
PnL Losses	4
Liquidity	3
Default	3
Other	1

Table 1: Amount of distress events by reason of distress

Aligned with the objective of our analysis, the initial step involved aggregating the largest possible set of news articles from reliable financial sources over an extended time horizon. This approach was designed to capture a sufficient number of distress events while maintaining uniformity in the raw data.

Due to the reliability of the sources and the volume of news collected, we focused on extracting headlines rather than full articles. This decision was based on the significant difference in length between the headlines and the body of accredited financial news, relative to the actual information conveyed. Given the size of the dataset, analysing the full text of news articles would have been computationally prohibitive. Future work could include the analysis of the article text.

c) *Data processing*

Upon completing the news collection process, we collected 15 million articles spanning the period 2007 to 2023, deemed insigthful for the number of distress events related to the selected sample of banks, as well as covering a wide variety of financial conditions. The articles originate from a diverse array of sources, primarily compilations of financial news such as Reuters, Bloomberg, Dow Jones, the Financial Times (FT), and S&P Global Market Intelligence. To further enrich the dataset, we incorporated ECB's internal news compilations as a complementary resource. Since the ECB's internal coverage commenced in 2014, leveraging external sources was essential to ensure comprehensive temporal and contextual coverage.

Among the totality of the news sourced, we also encountered miscellaneous and non-business-related news, which were dropped from the analysis. The initial step involved classifying the news articles into business-related and non-business-related categories. For this purpose, we employed a fine-tuned model based on DistilBERT-base-uncased , which is a distilled version of the Bidirectional Encoder Representations from Transformers (BERT) model uncased (i.e., without differentiating between lower and upper cases). This classification was vital in filtering out articles not relevant to the business domain, thereby streamlining the dataset to focus primarily on business news.

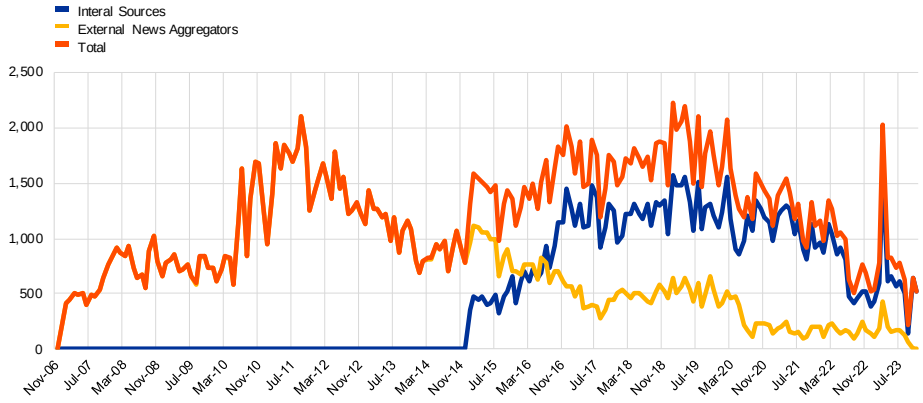


Figure 3: Evolution of amount of business-related news per source and year

Figure 3 shows the distribution of number of business-related news per year and source that were kept for the analysis, covering a broad period from 2007 to 2023. After the removal of news, however, the data remained extensive in its scope, encompassing a wide variety of business-related news.

One challenge was the linkage of news to a specific bank. Traditional string-matching algorithms compute pair-wise distances between strings (words or sentences). With the extension of our data this would turn out computationally costly and prone to errors, as the length of the headlines are heterogeneous and number of news is high. To overcome this challenge, we implemented a state-of-the-art NLP tool that allows to identify and classify entities (like names of people, organizations, locations, etc) within a headline: Named Entity Recognition (NER).

After collecting the news and cleaning the news, we applied a Named Entity Recognition (NER) model to categorize the text more effectively and link each news to a bank (or set of banks).

In this phase, we employed *bert-large-NER* model (Devlin, Chang, Lee, & Toutanova, 2018), a fine-tuned *BERT* variant optimized for this specific task. It recognizes and categorizes specific entities such as bank names, locations, and financial terminologies. A general example can be seen in Figure 4.

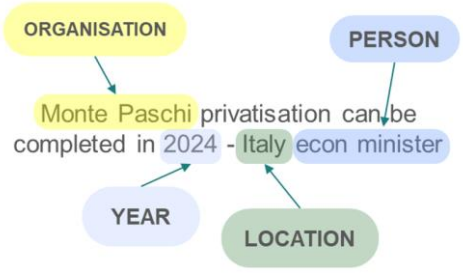


Figure 4: Named Entity Recognition Process

We created three distinct datasets, as described in Table 2. First, a bank-specific dataset comprised of around 175,000 unique news related exclusively to one of the banks in the sample. Second, a general-purpose dataset aiming at capturing the overall financial perception within a geography, so that news referring to the Spanish territory were mapped to all the Spanish banks. Third, a strictly central-banking-related dataset, which includes news potentially capturing spotted issues by national central banks at a national level. This third dataset is mapped following the same logic as the second one.

Category	Number of News
Bank	174,180
Bank Special Cases	10,888
Country Banks	118,970
Grand Total	**304,038**

Table 2: Number of distinct news titles by mapped category

After that, we encountered two challenges. First, we spotted news in which the object of the headline was not the bank itself, but rather was referring to other companies. Secondly, due to the variety of sources used, the same piece of news might have been reported on two or more sources in that same day, so there could be a duplication of news.

On the first challenge, we found that one example of those news is: "Analysts at Switzerland's UBS raise Apple's price target to $215". The object of the news is Apple, but still our matching algorithm would map this headline to UBS. We identified that, often, the structure of these news was rather homogeneous "Analysts at … raise/lower price target of …", allowing us to identify the pattern and drop them. However, as the dataset was large and not visually-explorable, we had to ensure that patterns like these did not happen in other cases. To identify potential patterns, we employed an N-gram algorithm. N-grams are a graphical model for capturing patterns in textual data with a sequential form, proved invaluable. An N-gram is a tuple-like data structure containing an *N* number of items, where *N* is a positive integer. On top of that, to ensure better efficacy of our pattern-recognition algorithm, we employed traditional NLP techniques, such as removing of stop-words, stemming and lemmatizing to only evaluate the core words. This means that in our previous example, the headline will look like: "Analy Switz UBS rise Apple target $215". All the relevant patterns were then visually explored and, upon expert judgement, news following non-relevant patterns were dropped from the analysis.

On the second challenge, we used an AI-powered similarity algorithm. The reason to leap over traditional approaches like the Levenshtein distance was that they miss the general meaning of the text itself. We suspected that, even though the information of the news was equal, two sources might report it differently yielding a high grammatical-distance between them. Therefore, we employed context-understanding tools, like *Google's T5* (Raffel, y otros, 2023). *T5* is a Text-To-Text AI tool that is specifically trained on this task (among others). It not only considers differences in length, wording and order (considering the Levenshtein distance or other similar metrics) but also the intrinsic meaning behind the text. This AI tool distinguishes itself from other NLP tools by providing both input and output as text strings, allowing for a more nuanced analysis of news content.

Since the T5 similarity algorithm provides a similarity score between two texts, we applied it to the each of the three datasets of news. For each pair of similar news headlines, if the similarity score exceeded a predefined threshold, we removed the shorter headline, retaining only the longer one.

This final processing step allowed us to merge all three datasets, clearly identifying the news coming from each dataset, and adding the distress events information into a consolidated dataset.

4) Sentiment analysis: feature creation and LLM agents

Following the curation and processing of news, we employ a set of agents on each news headline to generate a sentiment score. These agents have very diverse origins and perspectives and can be grouped in different categories and subcategories, as described below.

- Dictionary-based agents: Loughran-McDonald, AFINN, VADER.
- Pre-trained agents: FinBERT, LLAMA, Mistral, XDistil, Financial BERT, Financial RoBERTa, DistilRoBERTa, Financial SIGMA.
- Self-developed deep learning agents: CNN and LSTM.

- Weighted agents: built from any of the previous agents, weighted by a categorization of each risk.

The pure-sentiment agents constitute a set of models or rules that generate sentiment as output from a given news headline input. In contrast, refined agents take the sentiments produced by the pure agents as input and introduce an additional level of computation – they combine the sentiments from the pure-sentiment agents and/or weight them according to a classification scheme based on a relevance scale. All these agents express a sentiment in very different ways (by adjectives, with a discrete range of values or via a continuous one). The sentiment is then transformed into a score which represents the original sentiment on a scale from -1 to 1, with 1 signifying extreme positivity and -1 signifying extreme negativity. Each headline is evaluated by all the agents, having a set of scores per observation.

Figure 5 represents the workflow followed to arrive to a numerical sentiment score from a headline.

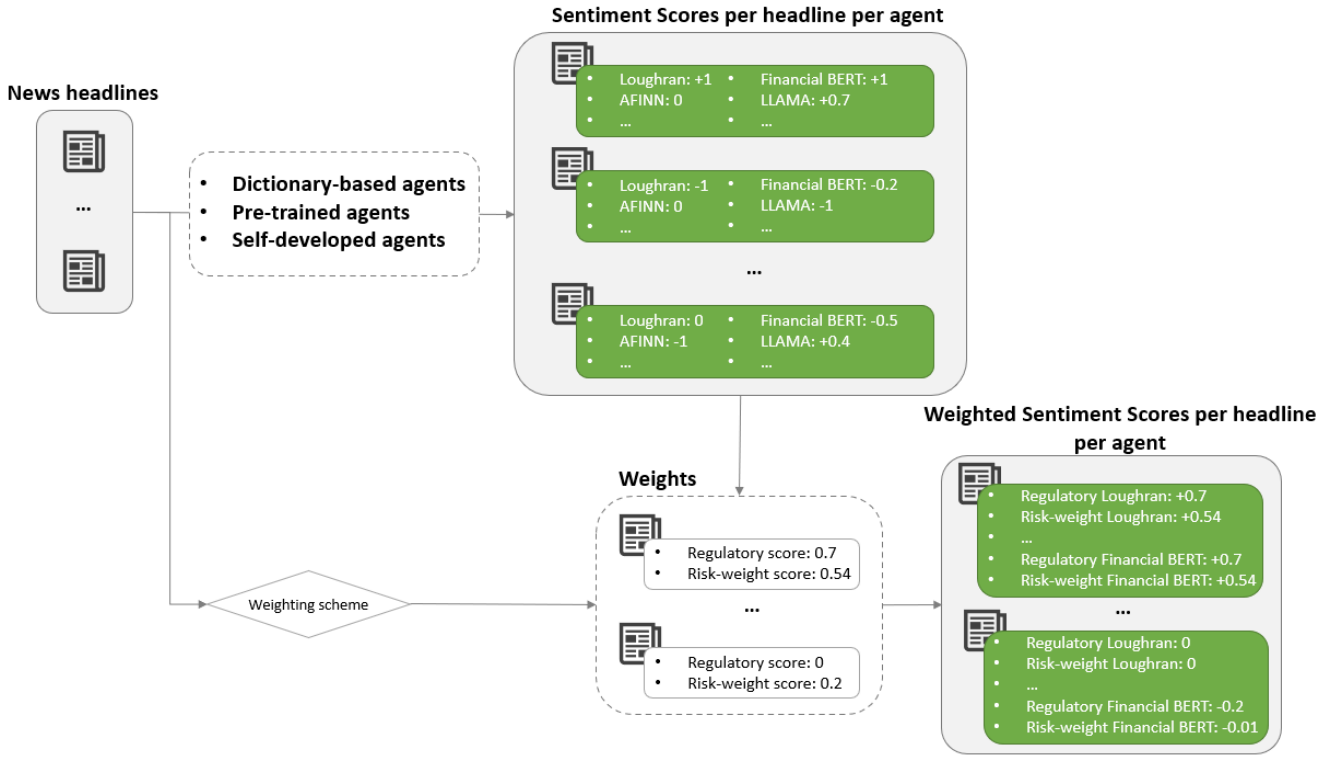


Figure 5: Workflow to obtain the sentiment analysis process

The decision to employ a range of widely different models instead of a single model as agents is driven by the diverse perspectives offered by each agent. Each of them is uncorrelated with the others and independent, due to the different philosophy each of them follows. This approach enables a more nuanced understanding of news dynamics based on the specific viewpoint employed. Consequently, if the primary goal is to gauge financial relevance, utilizing, for instance, an LLM trained on financial data is likely to yield a more precise score. Conversely, if the objective is to assess the regulatory implications of a news article, the inclusion of other agents (e.g., the regulatory weighted agents or our self-developed agents) becomes essential. In this context, agents can be likened to independent consultants, each offering specialized insights.

a) Dictionary-based agents

The initial set of agents we present are the more traditional and basic ones, constructed using lexicons (dictionaries) enhanced with rule-based logics. These dictionaries, sourced from previous research and other authors, serve the purpose of evaluating the valence of a text, measuring the perception of the goodness or badness of an event.

These dictionaries consist of word-score pairs. Each text is tokenized into individual words, and if a word is present in the dictionary, its score is aggregated. The cumulative score of all the words within a headline is then considered the headline score, which is subsequently normalized between –1 and 1. The main advantage of using these agents is that they do not require any training data and are computationally inexpensive. However, due to the simplicity of their approach they are more prone to offer a rigid sentiment based on their specific domain, failing to capture the overall context.

We use a total of 3 different dictionary-based agents:

- Loughran-McDonald (Loughran & McDonald, 2011): tailored for financial texts with categories particularly relevant to capture financial sentiments (it was specifically created to analyse the tone of financial reports, earning calls, and other financial documents).
- AFINN (Nielsen, 2011): general-purpose lexicon with sentiment scores, applicable to a wide range of texts but without domain-specific nuances.
- VADER (Hutto & Gilbert, 2014) (Valence Aware Dictionary and sEntiment Reasoner): sentiment lexicon that is especially attuned to microblog-like contexts (e.g., social media) thanks to the incorporation of various heuristics like punctuation (e.g., use of exclamation points), capitalization, degree modifiers, and slang/colloquial terms.

b) Pre-trained agents

Pre-trained agents are a set of LLM models that have already been trained on a specific and distinct dataset. These datasets are extensive, often reaching billions of observations. The main advantage of using these agents is the bigger precision on the assessment of sentiment and the lack of training requirements. However, using them is computationally intensive. We use a total of 8 different agents, as shown in Table 3.

Agent	Aim of the agent	Trained on
FinBERT[1]	Financial view	PhraseBank
LLAMA[2]	General view	Generalised text
Mistral[3]	General view	Generalised text
XDistil[4]	Fine-tuned to classify sentiment of news headlines	Annotated news using GPT 3.5
Financial BERT[5]	Fine-tuned for sentiment analysis	Large corpora of financial texts + Financial PhraseBank
Financial RoBERTa[6]	Fine-tuned for financial sentiment analysis	Financial statements, Earnings announcements, ESG News, Financial news
DistilRoBERTa[7]	General view	OpenWebTextCorpus (generalised text)
Financial SIGMA[8]	Fine-tuned for financial analysis	PhraseBank

Table 3: Description of agents used

[1] (ProsusAI, 2021)

[2] (Meta, 2023)

[3] (Mistralai, 2023)

[4] (Hægland, 2023)

[5] (Rachid, 2023)

[6] (Soleimanian, 2023)

[7] (Romero, 2020)

[8] (Sigma, 2023)

When using LLM models, it is imperative for users to formulate well-defined prompts to produce homogeneous and consistent outputs. Prompts are command-like requests given to each agent that, essentially, ask for the sentiment of a sentence. We experimented with multiple prompts until we reached an optimal level of accuracy and efficiency.

These agents are not limited to a ternary classification (i.e., a positive, negative or neutral sentiment) but rather they can provide sentiment in a continuous scale [-1, +1], allowing for a more precise analysis than dictionary-based sentiments.

c) Self-developed deep learning agents

In contrast to the agents presented above that rely on external models, we developed two different agents in-house. This allowed us to train the model on domain-specific data, tailored to our supervisory needs. Hence, we picked and labelled a set of 10,000 curated news that we deemed especially relevant from a supervisory perspective into positive, negative or neutral sentiment and used them to train the models. First, we trained a Convolutional Neural Network (CNN) and then a Long Short-Term Memory (LSTM) neural network. Both belong to the family of Recurrent Neural Network (RNN) architectures. These models are proficient in sentiment analysis and are able to capture long-term dependencies.

To assess the accuracy of the model, we compared the model output with the actual sentiment of each of the headlines. We consider a prediction to be correct if the predicted sentiment (positive/negative/neutral) is in line with the actual one Both models show strong performance in predicting sentiments. In the test sample, the CNN model accurately classifies 77% of positive, 76% of neutral and 86% negative headlines. Similarly, the LSTM model in the correctly classifies 77% of positive, 84% of neutral, and 91% of negative sentiments in the test sample, as shown in Table 4. Both models are quite effective with an overall accuracy of over 80%. However, the LSTM model excels in neutral and negative regulatory news sentiment classification.

	CNN	LSTM
Sentiment	**Accuracy**	**Accuracy**
Positive	0.77	0.77
Neutral	0.76	0.84
Negative	0.86	0.91
All Classes	**0.81**	**0.85**

Table 4: CNN and LSTM model performance statistics

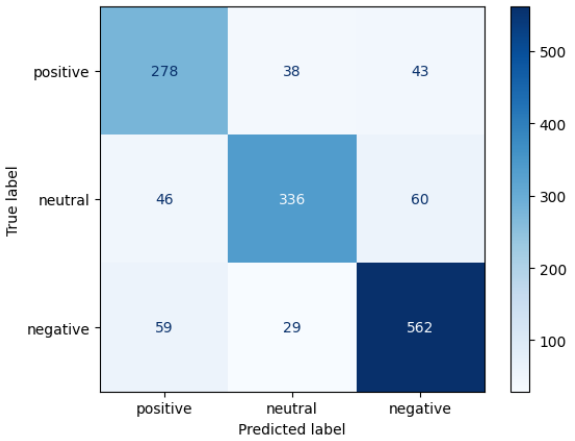


Figure 6: CNN accuracy statistics in the test sample

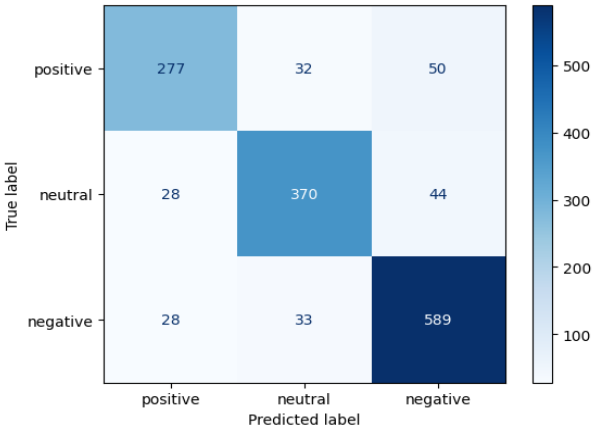


Figure 7: LSTM accuracy statistics in the test sample

d) Weighted agents

With the aim of further introducing supervisory-specific views on the sentiment scores, we developed two different weighting schemes: regulatory lexicon and risk-specific weights. Both are aimed at giving each headline a score from which to weight the already-defined sentiment. These, by no means, were meant to replace the existing sentiment but rather to complement it with a tailored and domain-specific view.

The regulatory lexicon functions in the same way as the dictionary-based agents: a lexicon is built, and a score is given to each word. Each headline's score is then aggregated by the sum of the individual word's score. The lexicon, however, was built from the N-gram inspection of official regulations and the addition of words based on our expert judgement. It comprises 63 unique terms and associated weights.

The risk-specific agents offer a more complex classification. For that, we pre-defined a set of 11 different categories and their corresponding weights (see Table 5). The weights were derived from expert judgment, informed by the relative significance of each risk factor in the supervisory context. Notably, the weights do not sum to 100% because they are not probabilities but rather multipliers used to emphasize each category's importance in the scoring process. We employed Facebook's *BART-Large* model (Lewis, y otros, 2019) for a zero-shot-classification of each of the headlines (Alhoshan, 2023). Zero-shot classification is a machine-learning approach that allows models to classify text into categories without prior training specifically on those categories. Instead, the model uses its pre-existing understanding of language to evaluate the probability of a headline belonging to a given category. In this case, we inputted the headline and the predefined risk categories as "prompts" to assign probabilities across all categories.

Category	Weight
Capital Adequacy and Capital Structure	3
Corporate Strategy and Management	1
Credit Risk and Asset Quality	5
Governance and Organizational Structure	1
Legal and Regulatory Issues	4
Liquidity and Funding	2
Market Trends and Stock Performance	1.5
Market Risk and Derivatives	5
Mergers and Acquisitions	2
Profitability and Earnings	3
Other	1

Table 5: Categories and weights chosen to build risk-specific agents.

This classification provided a probability of belonging to each of those categories per headline, adding up to one. In order to select the main category among those presented, we employed a rule-based approach depending on the variability of the probabilities, as measured by the standard deviation. We defined a bucketing criterion based on thresholds. In cases where the standard deviation was higher than the upper threshold, we assigned the category with the highest probability to that headline. In cases where it was in-between the upper and lower threshold, we took the top two categories (the one with the highest probability was assigned as the main category and the following one as the second-best). In cases where it was below the lower threshold, it was assigned to *Other*.

After this process, we were left with two scores per headline: one regulatory (based on the lexicon explained above) and one risk-specific (based on the categorization technique). Therefore, each of the agents defined in the previous sections can be weighted by these scores to achieve new, complimentary agents with a domain-specific perspective.

5) Modelling bank´s distress

After successfully completing the process described above, we have compiled a dataset consisting of news articles at the bank level, with each news item assessed by all agents. However, our goal is to model the future probability of distress. In order to do that, we follow these steps:

- Daily dataset: aggregation per bank and day of the sentiment obtained by each agent.
- Feature creation: arithmetic transformation of existing sentiments per agents into quantitative indicators.
- Monthly dataset: shifting from a daily towards a monthly dataset and adjoining the corresponding distress events.
- Modelling: application of different machine learning techniques to predict the probability of distress per bank in the future. All the different sentiment features are aggregated into a single and unique signal.

A more in-detail workflow is outlined in Figure 8:

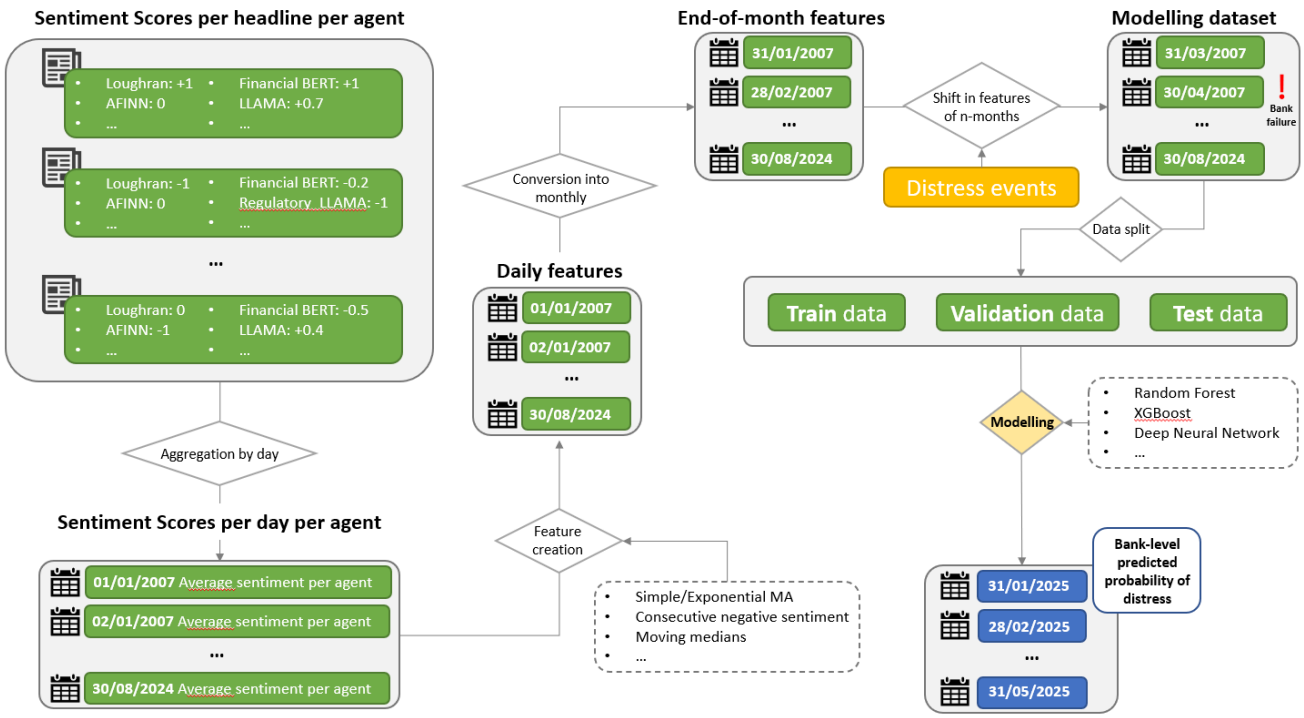


Figure 8: Modelling workflow

a) Feature creation

To obtain the daily dataset, we aggregate per bank and day all the sentiments obtained in the previous section. This acts as the foundation layer from which the features are computed. The features used in this analysis are derived from two primary variable sources: the volume of news articles and the sentiment scores assigned by each agent to the news. The derived feature set comprises first-level calculations executed using various methods – such as counts, moving averages, exponential moving averages, moving medians, and moving modes – applied across multiple temporal windows (i.e., 30, 60, 90, 180, and 365 days on a rolling basis) to capture dynamic sentiment variations over time according to different perspectives. In addition, to introduce more pronounced variations while preserving the interpretability of the original features, second-level transformations were applied. These transformations included discretization techniques based on fixed or percentile-defined thresholds and scaling functions incorporating caps and floors, which resulted in bucketed and binary variables derived from the initial feature set. Following both levels of feature transformations, the initial dataset – comprised of news counts and sentiment scores – was expanded into a comprehensive set of over 500 distinct features. Table 6 represents the features used:

Feature	Feature type	Temporal Window (days)	Applicable on
Moving average	Sentiment	30, 60, 90, 180, 365	All agents
Exponential moving average	Sentiment	30, 60, 90, 180, 365	All agents
Moving median	Sentiment	30, 60, 90, 180, 365	All agents
Consecutive negative days (rolling)	Sentiment	-	All simple and exponential moving averages for all agents
Dynamic counts (amount of news in the past selected temporal window)	News volume	30, 60, 90, 180, 365	-

Table 6: Feature types and applicable windows and agents

In our analysis, each feature follows a common naming structure. Features´ names start by indicating the kind of feature they are: moving average (MA), exponential moving average (EMA), moving median (MM), consecutive negative days (Consect_neg_days) and dynamic counts (Dyn_count); followed, if applicable, by an integer indicating the time window in which the feature is computed. Lastly, the name of the agent from which the sentiment is computed is indicated.

In addition to that, we leveraged on the two supplementary datasets presented in the Data collection and processing section to obtain predictors that capture more general trends that can be associated with a specific group of banks. A group is defined as all the banks operating under the same Central Bank and in the same country. Finally, two additional features emanate:

- Geographical sentiment: it gives an overall temperature of the country and is not necessarily exclusively related to banks. All news mentioning the country in question are captured. However, due to the nature of our data sources, a strong bias exists towards financial news.
- Central Bank sentiment: captures the aggregated perception of the banking sector in that country.

Then, we transformed the frequency of our dataset from daily to monthly. This procedure follows a three-step process. First, we kept the end-of-month observations. Second, we adjoined the compiled distress events to the end-of-month dataset, with a shift of two months. This way, the associated sentiment of the time whereby a bank distress event happened is linked to the features two months back. The rationale is that, at the modelling step, we want to predict the probability of a potential bank default in the following two months in order to give supervisors enough time to act.

b) Sample selection

To arrive to the final dataset used for the modelling section, we apply some conditions to ensure a higher quality and reliability of our data. First, banks with less than 10 news in the 60 days leading to the distress event are eliminated from the analysis, as the lack of news (and therefore reliable features) can deem the results spurious. Secondly, after a distress event occurs, the following two years of observations for the impacted bank are dropped from the analysis to avoid double-counting and biasing the features towards a more negative sentiment. We consider this two-year window is enough for a bank to recover after a distress event, regaining its financial health and an independent news coverage once again.

After these considerations, we are left with monthly observations of 168 banks and a total of 76 distress events, spanning from January 2007 to May 2023. We are left with a classification task in which our dependent variable is binary, indicating if a bank is suffering a distress or not at that period of time. By nature, however, our dataset is highly imbalanced, as the number of "healthy" observations (18,098) greatly exceeds those of distress (65), which is a challenge from a modelling perspective.

Following standard machine learning practices, we split the dataset into train, validation, and test sets. While a simpler train/test split is often sufficient for basic evaluations, incorporating a validation set provides a more robust framework for model development. The validation set acts as an intermediary, allowing us to fine-tune model hyperparameters, monitor overfitting, and compare the performance of different model variants without directly exposing the test set. This ensures that the test set remains a truly unseen dataset, reserved exclusively for the final evaluation of the model's generalization performance. By doing so, we mitigate the risk of overfitting hyperparameters to the test data, which can lead to overly optimistic assessments of model performance. Consequently, this three-way split setup improves the stability and generalization capacity of the models, as it enables iterative improvement using the validation set while preserving the integrity of the test set. We end up splitting the dataset by year. The training set comprises the years 2007-2015, the validation set 2016-2018 and the test set 2019-2023. The final setup is outlined in Table 7.

We choose to perform an out-of-time (OOT) test as it constitutes an even more robust approach than the out-of-sample (OOS) one due to its ability to test the model's robustness against temporal changes in the underlying data (Blümke, 2017). Studies indicate that models validated with OOT data often achieve superior predictability and lower misclassification costs compared to those using OOS validation. This advantage arises because OOT validation more effectively captures temporal dynamics and potential shifts in data distributions (Xia, 2020).

As previously mentioned, from a modelling perspective, the main challenge is the highly imbalanced dataset. Indeed, the variable that we target to model, expressing in a binary outcome the presence of distress events, is highly imbalanced by nature, as these events occurs rarely. To overcome this limitation without losing information from the sentiment analysis and NLP work previously conducted, we adopted an oversampling technique, implementing the Borderline-SMOTE (Han, Wen-Yuan, & Bing-Huan, 2005), an improved version of the Synthetic Minority Over-sampling Technique (SMOTE), which oversamples, via linear combinations of the variables, the datapoints close to the classifier borderline. Using this method, we were able to enrich the original sample with 6,480 synthetic datapoints linked to distress events, reaching a rather balance 65/35 split in the final dataset. This oversampling technique is key to help the classification algorithms learn from the distress events, allowing to better differentiate between healthy and unhealthy observations in the validation and test sets. Hence, after oversampling, we are left with the following setup:

Set	Before oversampling			After oversampling		
	Healthy observations	Distress events	Total observations	Healthy observations	Distress events	Total observations
Train (2007-2015)	9801	51	9852	9801	6532	16333
Validation (2016-2018)	3346	7	3353			
Test (2019-2023)	4951	7	4958			

Total number of features: 503

Table 7: Number of observations in the train, validation and test samples before and after oversampling

c) Model development

In our effort to predict banking distress events from sentiments of news articles, we include a wide range of explanatory variables, produced as described in the previous section. We then apply most of the available and applicable machine learning statistical methods. Namely, the following models are included in our analysis: random forest (RF), extreme gradient boosting (XGB), deep neural networks (DNN), logistic regression (LogR), support vector machines (SVM), K-nearest neighbours (KNN). Each of these models is used in the context of a binary classification problem for bank distress, given the whole set of features produced from the agents. All models in this study underwent a systematic hyperparameter tuning process, wherein the models were trained on the training dataset to iteratively adjust hyperparameters with the objective of optimizing performance on the validation set. While decreasing the chances of overfitting the training dataset, this process ensured that the models were fine-tuned to achieve the best possible balance between model complexity and generalization capabilities. Subsequently, the models were assessed for their performance in the unseen test dataset.

As the primary objective of this analysis is to develop an early-warning system for bank distress, minimizing Type II errors (false negatives) holds greater importance than minimizing Type I errors (false positives). This prioritization stems from the observation that failing to flag distressed banks (Type II error) would prevent the activation of necessary controls to mitigate potential risks. In contrast, over-flagging banks (Type I error) may lead to further investigations, which can ultimately resolve or dismiss erroneous flags. Consequently, in addition to maintaining an acceptable level of accuracy, our validation focuses primarily on sensitivity metrics or hit rate, while precision is deprioritized in the evaluation of model performance.

d) Out-of-time model validation

The confusion matrices for each model's application on the out-of-time test dataset are presented in Table 8, along with their respective hit rates and false alarm rates. It is evident that the LogR and SVM models deliver the poorest results in terms of hit rate. This may stem from their relatively simple linear architectures, which may struggle to capture complex, non-linear relationships between news sentiment and bank distress events, leading to subpar performance.

Similarly, the KNN and RF models exhibit limited effectiveness in capturing the impact of news on bank distress. For KNN, this could be due to its reliance on local neighbourhood-based learning, which might overlook broader patterns critical for predicting rare events like bank distress. RF, while robust in many cases, can be limited when dealing with subtle temporal dynamics and feature interactions, which are better handled by more advanced models.

In contrast, XGB and DNN outperform the other models by a significant margin, showcasing the highest sensitivities and strong predictive capabilities. XGB benefits from its iterative nature, where errors from previous iterations are progressively corrected, allowing it to identify complex interactions within the data effectively. Similarly, DNN excels in processing high-dimensional and non-linear data due to its ability to learn hierarchical feature representations. Indeed, as previously mentioned, we are looking for the best models in terms of hit rate with reasonably good performance in terms of false alarm rate. The results confirm comparatively good performances of the selected models, XGB and DNN, as their false alarm rates are aligned with the ones related to the other models.

RF	Prediction			
Actual	**0**	**1**	**Signal**	**Rate**
0	4698	253	False alarm	5.11%
1	2	5	Hit rate	71.43%
XGB	**Prediction**			
Actual	**0**	**1**	**Signal**	**Rate**
0	4464	487	False alarm	9.84%
1	1	6	Hit rate	85.71%
DNN	**Prediction**			
Actual	**0**	**1**	**Signal**	**Rate**
0	4623	328	False alarm	6.62%
1	1	6	Hit rate	85.71%
KNN	**Prediction**			
Actual	**0**	**1**	**Signal**	**Rate**
0	4527	424	False alarm	8.56%
1	4	3	Hit rate	42.86%
LogR	**Prediction**			
Actual	**0**	**1**	**Signal**	**Rate**
0	4627	324	False alarm	6.54%
1	4	3	Hit rate	42.86%
SVM	**Prediction**			
Actual	**0**	**1**	**Signal**	**Rate**
0	4951	0	False alarm	0.00%
1	4	3	Hit rate	42.86%

Table 8: Prediction hit and false alarm rates for each model

To assess the models from additional perspectives, we calculated a series of metrics as proposed by (Bekkar, Djemaa, & Alitouche, 2013) and present them in Table 9. These metrics are used to derive a comprehensive evaluation of each model's classification capability in relation to the others. While some correlation among metrics may exist, all are incorporated to accurately assess the predictive performance of the methods used, even in the cases where performance differences are minimal. Specifically, we examined the Receiver Operating Characteristic (ROC) curves for each, with particular emphasis on the Area Under the Curve (AUC) and the associated Gini coefficient (GC=2·AUC-1) as key performance indicators. The AUC metric and the GC provide valuable insights into the sensitivity of each classification model to varying threshold values, offering a more comprehensive evaluation of model performance in terms of discriminatory power. The Kolmogorov-Smirnov (KS) statistics is computed as the maximum difference between the empirical cumulative distribution functions of the predicted scores for the positive and negative classes. As such, it evaluates the model's ability to distinguish between the positive and negative classes based on the predicted probability scores. Finally, the Balanced Accuracy (BA) is computed as the average between the sensitivity (hit rate, in our analysis) and the specificity. Therefore, the BA metric conveys a similar meaning as the overall accuracy, but it overcomes the imbalanced nature of the two classes, providing a more targeted result for our aim. As illustrated in the table below, despite a more uniform distribution of results, the final statistics reaffirm our earlier findings. Specifically, again the XGB and the DNN models demonstrate superior performances, exhibiting consistently high values across the different metrics, regarding the discriminatory power, the class separation capability and accuracy of the models tested. This consistency further solidifies their position as the best-performing models in the analysis.

Metric	AUC	GC	KS	BA
RF	74.4%	48.8%	45.8%	72.7%
XGB	90.6%	81.3%	69.2%	84.4%
DNN	89.2%	78.4%	80.7%	90.4%
KNN	85.0%	69.9%	56.9%	73.2%
LogR	74.8%	49.5%	42.2%	68.9%
SVM	76.0%	52.0%	41.1%	70.6%

Table 9: Model performance metrics

e) Interpretability of best candidate models

Focusing on the two best-performing models, we now address model transparency by investigating their interpretability — specifically, whether the models' use of features aligns with the objective of predicting bank distress events. To assess feature contributions, we employ the Shapley Value approach (Fryer, Strümke, & Nguyen, 2021), which measures a feature's impact by calculating its average marginal contribution to model predictions. This is done by comparing model outputs with and without the targeted feature for the possible subsets of features. This approach fits our analysis as it is model-agnostic, additive and offers both global and local interpretability. The SHAP package in Python provides a useful Beeswarm plot for visualization, where the Shapley Value determines the horizontal position, vertical spread indicates feature density, and colour represents the feature's original value. This is presented in Figure 9 and Figure 10, which report the 20 most important features for the XGB and DNN models, respectively. For instance, we can see that for the XGB model, high values of the 180-day moving average of the sentiment derived from Llama (MA180_LLAMA2) are decreasing the probability of distress when included in the model, while the opposite can be said for the lower values. The same applies in the case of the DNN model for the yearly moving mode based on the mixed self-developed LSTM agent (MM365_Mix_LSTM).

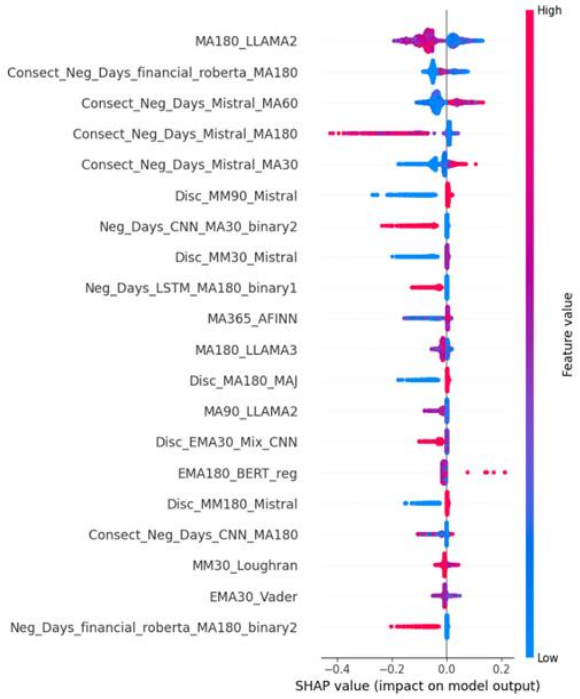


Figure 9: Shapley values of the 20 most important features for the XGB model

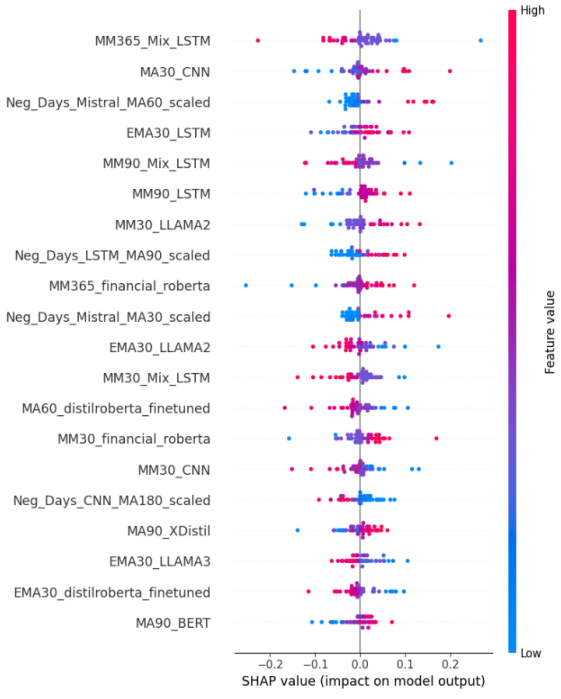


Figure 10: Shapley values of the 20 most important features for the DNN model

These observations are also reflected in the related partial dependence plots (PDP), illustrated in Figure 11 and Figure 12, each displaying the relationship between a value of the selected feature and the conditional expectation of the model predictions

if the feature is kept constant at that value, while the other features are varying (Goldstein, Kapelner, Bleich, & Pitkin, 2015). Specifically, the PDP of the feature MA180_LLAMA2 in the XGB model shows a stepwise constant trend that, aligned with the intuition, presents a jump around 0, associating negative values of this feature to a higher probability of distress and positive values to a lower one. The intuitive nature of this interpretation also holds in the case of the PDP of the feature MM365_Mix_LSTM for the DNN model, where the trend appears almost linear, with negative inclination.

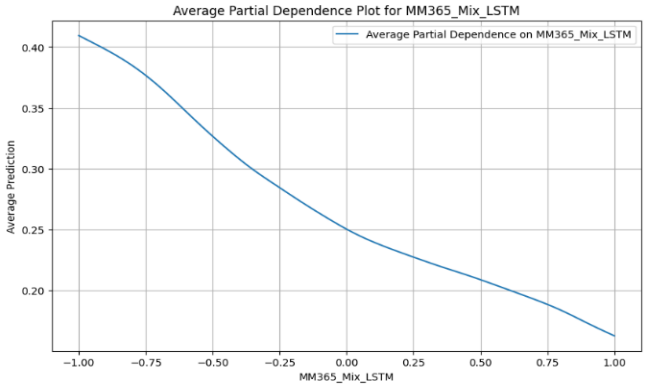


Figure 11: Partial Dependence Plot for MM365_Mix_LSTM in the DNN model

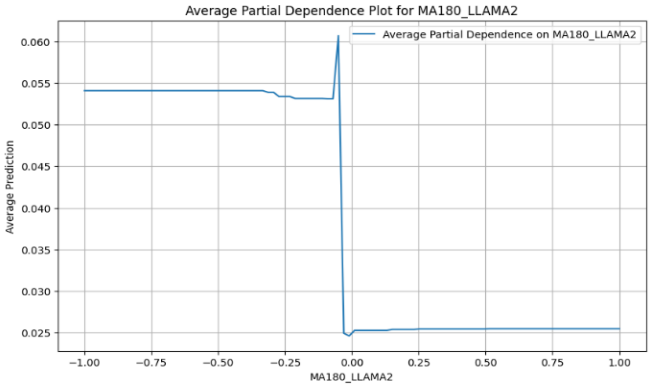


Figure 12: Partial Dependence Plot for MA180_LLAMA2 in the XGB model

From a comparative perspective, with reference to Figure 9 and Figure 10, the charts reveal that the two models share few important features, indicating a high degree of independence between them. Furthermore, the XGB model is predominantly impacted by features related to the pure count of consecutive negative days or discretised features derived from moving averages or medians, mainly related to pre-trained agents (primarily Mistral, but also Llama, Financial Roberta, AFINN, etc.). On the contrary, the DNN model is majorly affected by scaled counts or pure moving medians or averages referred to self-developed deep-learning agents (LSTM and CNN). Therefore, by means of the Shapley Value analysis, there is an evident complementarity between the XGB and DNN models, as they achieve similar performance while relying on distinctly different sets of features. In view of this, a natural unbiased choice for a final model selection consists in an ensembled model given by the uniform average of the XGB and the DNN models. This selection guarantees greater features coverage, allowing the final ensembled model to capture aspects that will be otherwise neglected by the single models, while the close similarity in terms of performances gives evidence for an equal weight when aggregating them.

f) Final model experimental results

To assess the accuracy and robustness of the final model, we conducted a back-testing analysis on key banking institutions. One notable example is Credit Suisse, one of Switzerland's largest banks, which faced a significant crisis in March 2023, culminating in its acquisition by UBS, another Swiss bank. As shown in Table 10, distress signals began to emerge in the news coverage as early as nine months before its eventual collapse. However, in today's environment of constant and rapid information flows, distinguishing actionable signals from background noise remains a significant challenge. The collapse of Credit Suisse highlights the critical role of early detection systems for bank distress, enabling regulators to identify and connect key risk signals amidst the noise.

Date	Headline	General Sentiment
4th August 2022	Credit Suisse mulls cutting thousands of jobs globally, report says	-0.56
27th July 2022	Credit Suisse posts bigger than expected Q2 loss	-0.98
2nd September 2022	Credit Suisse eyes massive job cuts	-1
3rd October 2022	Concern over Credit Suisse escalates further	-0.80
22nd September 2022	Credit Suisse could split into three, the stock falls	-0.65
2nd September 2022	Credit Suisse looking at cutting around 5,000 jobs	-0.93
3rd October 2022	Credit Suisse: credit default swap at its highest since 2009	-0.91
3rd October 2022	UBS credit default swaps hit decade high after Credit Suisse blow out	-0.84
4th October 2022	Credit Suisse - the collapse of a Swiss flagship	-0.54

Table 10: Selection of Credit Suisse news prior to its collapse.

While isolated news items may trigger alarms, our model emphasizes that the real insights lie in analysing general trends. In this case, Cassandra's probability of distress estimate for Credit Suisse began to climb rapidly in early 2021 and maintained this upward trajectory throughout 2022, as illustrated in Figure 13. The imminent collapse, happening in Q1 2023, was correctly predicted by Cassandra.

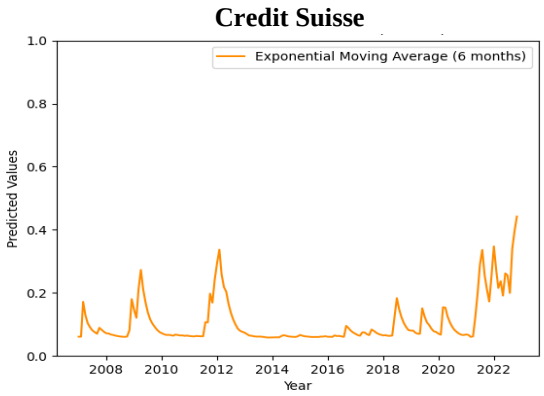


Figure 13: Credit Suisse predicted probability of distress. Y-axis represents 6-month exponential moving average of the predicted probability

Another notable example is the series of defaults by National Bank of Greece (NBG). The first occurred in 2012 following a government bail-out and subsequent restructuring, during which private creditors were required to accept haircuts on Greek bonds. The second happened in 2015 when capital controls were imposed in Greece after a default on an IMF payment, leading to widespread bank closures and another bail-out. It can clearly be seen in Figure 14 that Cassandra's predicted probability of distress spiked well in advance of these events. Additionally, it is also worth noting that after each of those events, the observations were removed, as outlined in section 5)b) (sample selection), to avoid biasing the model towards negative sentiments. Furthermore, the model successfully captured the improvement in NBG's financial health after significant deleveraging and the sell-off of lower-quality loans in the period 2019-2022, reflected in a marked decrease in its predicted probability of distress.

The model's sensitivity to events beyond strictly financial matters is evident in Figure 15, where the estimated probability of distress for Deutsche Bank is represented. Cassandra captured not only broad periods of financial turbulence, such as the 2008 Great Financial Crisis and the 2012 EU sovereign debt crisis, but also Deutsche Bank's legal challenges between 2016 and 2019. These challenges included litigation charges, police raids related to the Panama Papers investigations, and allegations of money laundering.

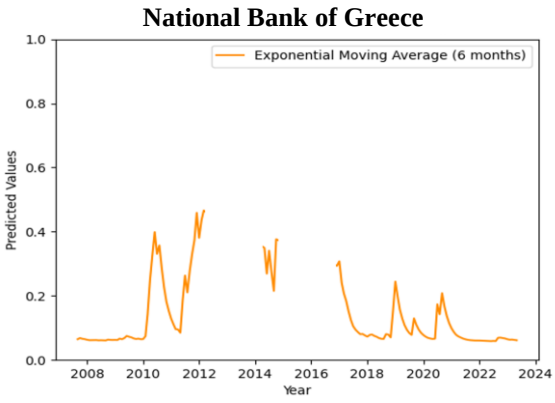


Figure 14: National Bank of Greece predicted probability of distress. Y-axis represents 6-month exponential moving average of the predicted probability

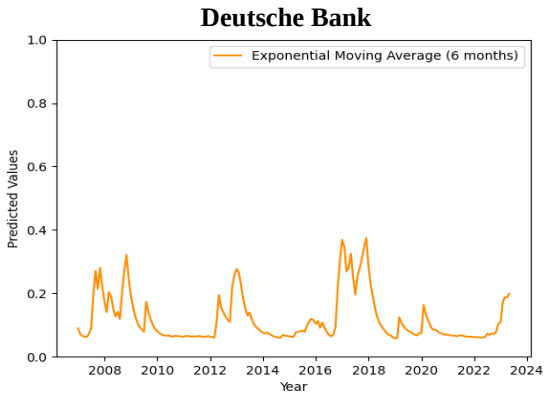


Figure 15: Deutsche Bank predicted probability of distress. Y-axis represents 6-month exponential moving average of the predicted probability

Cassandra also demonstrates its reliability by avoiding false positives for banks in stronger financial positions. For example, Figure 16, shows Cassandra's low predicted probability of distress for Banco Santander. While there were notable spikes in 2012 due to the EU sovereign debt crisis, Cassandra's estimates reflect the bank's underlying financial stability, reflecting its resilience during and after such periods of turmoil.

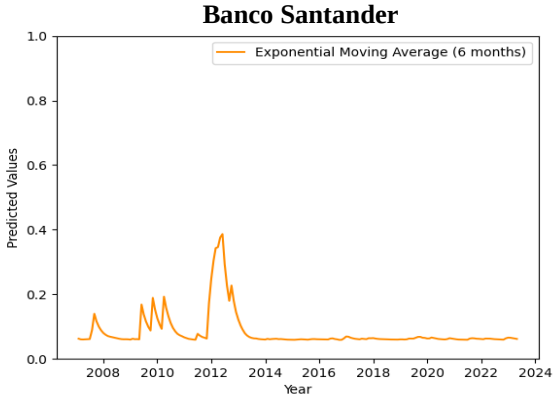


Figure 16: Banco Santander predicted probability of distress. Y-axis represents 6-month exponential moving average of the predicted probability

The proposed framework provided by Cassandra offers a sophisticated approach to financial stability by aggregating individual bank distress probabilities into a singular, overarching probability of systemic distress. This capability allows supervisors to detect events with widespread implications, enabling them to raise signals that may indicate broader systemic risks.

The primary benefit of this tool lies in its capacity to transform supervisory strategies from reactive to proactive. With early detection of systemic threats, supervisors can implement timely interventions, adjust regulatory measures, and tailor oversight efforts to mitigate risks before they materialize into crises. This proactive stance enhances the resilience of the financial system, reducing the likelihood of widespread disruptions and safeguarding economic stability. Furthermore, by providing a quantitative basis for decision-making, the model supports data-driven policies and fosters a more transparent and accountable supervisory process.

In Figure 17, we present the aggregated estimated probability of distress for the largest banks, identified by their size (measured in Risk-Weighted Assets). The included banks are Deutsche Bank, Banco Santander, BNP Paribas, BPC, Crédit

Agricole, Société Générale, UniCredit, and ING. To derive a single, cohesive signal reflecting system-wide trends, the probabilities are weighted by the banks' Risk-Weighted Assets.

Notably, the periods of heightened systemic risk—such as the Great Financial Crisis in 2008, the European sovereign debt crisis in 2012, and the COVID-19 pandemic in 2020—are reflected as clear increases in the aggregated estimated probability. he relatively lower and stable values between these peaks suggest periods of reduced systemic distress or more stable financial conditions. The gradual decline in prediction values post-2020 could indicate recovery efforts and improved economic conditions in subsequent years.

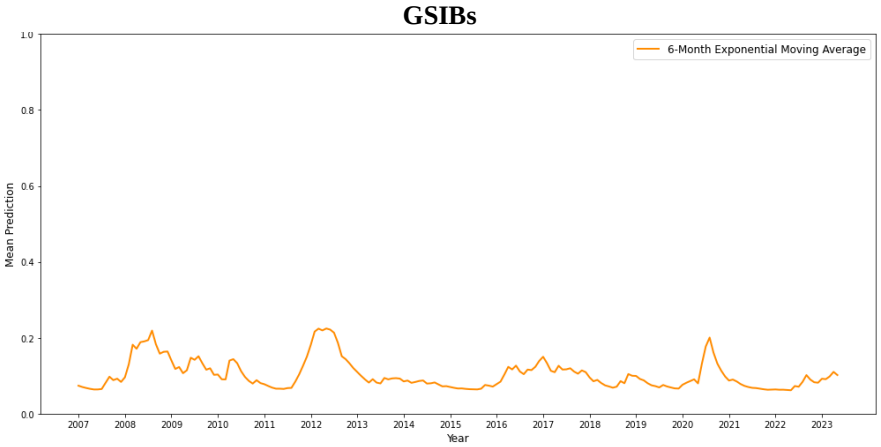


Figure 17: G-SIB predicted probability of distress. Y-axis represents 6-month exponential moving average of the predicted probability

6) Conclusions and future work

This article offers the proof-of-concept of an automated early warning system capable of identifying and flagging banks at risk of distress by utilizing exclusively qualitative sentiment information of publicly available financial news obtained by using industry-leading AI tools. Importantly, the automated system proposed in this study provides regulators and supervisors with a comprehensive and scalable tool for ongoing surveillance, enabling real-time and data-driven decision-making in the banking ecosystem.

In this analysis, we have demonstrated the application of Named Entity Recognition (NER) to a large corpus of financial news, enabling the identification of relevant entities. Following this, we explored various methods to generate sentiment scores for these news articles, ranging from basic approaches to more advanced techniques, culminating in the development of custom agents tailored to our specific objectives. From these sentiment scores, we derived a diverse set of features, which were subsequently used to implement machine learning models to create a unique, bank-level signal of probability of distress in the next two months.

The main contributions of this empirical study and its clear differences from other studies in the related literature of bank failures can be summarized in three layers. First, the sole use of qualitative text-based data, such as financial news, for conducting analysis of bank distress events marks a significant departure from the traditional literature, which predominantly relies on quantitative indicators. This approach not only sheds light on the expansion of the scope of the modelling dataset but also introduces a new, forward-looking source of information into the regulatory and supervisory resources. Second, it is notable the use of an innovative sentiment analysis framework. Indeed, we implemented industry-leading Large Language Models (LLMs) and artificial intelligence (AI) technologies to process a vast dataset of news articles, assigning sentiments in a novel automatic way in the regulatory banking sector. In particular, we introduced three key innovations in this regard: self-developed agents (custom-built, leveraging AI models to refine sentiment classification tailored to financial contexts), regulatory agents (grounded in regulatory documentation, enabling sentiment analysis specific to compliance and supervisory concerns), and risk-weighted agents (through a zero-shot classification approach, these agents categorized news sentiment based on various business thematic sectors, incorporating domain-specific risk profiles). Finally, another core aspect of this project is the predictive modelling of bank distress. This involves the application of sophisticated statistical methods to develop models for forecasting bank distress. We employed deep learning techniques, comparing their performance against traditional machine learning algorithms and standard statistical models. This approach enabled us to capture nuanced relationships between sentiment patterns and distress events and to select the models that better capture this relationship, allowing us to develop a whole framework optimised for accuracy.

One of the primary challenges we encountered was the inherent imbalance in the dataset, characterized by infrequent distress events compared to the non-distress *status quo*. To address this issue, we applied the Borderline SMOTE oversampling technique, which proved effective in mitigating class imbalance and improving model performance. Another limitation of the framework is the reliance on the amount of bank-specific news. Cassandra is less adaptive in situations where the news coverage of a specific bank is low and shows strong performance when the availability of news is abundant.

We think there are three promising directions for future research emerging from this work. First, the inclusion of non-English news sources could significantly expand the dataset, allowing for a more diversified and representative sample of global banking news. While this would introduce translation and cross-linguistic sentiment challenges, it could also provide a richer, country-specific context that could enhance predictive accuracy. Second, a more general area of improvement consists of enriching the dataset with a larger set of jurisdictions. This development would contribute to a more robust infrastructure and model, while also helping to address the data imbalance previously discussed. Future studies could explore the integration of more granular financial metrics and regulatory data to complement sentiment analysis, further refining the forecasting models. Third, the increase of sources and news coverage will lead Cassandra to be more adaptive to sudden and recent economic events, allowing for more tailored estimation capabilities.

In conclusion, the framework we have developed contributes to the growing intersection of AI, sentiment analysis, and banking supervision. The methods and insights presented here offer a foundation for continued innovation in regulatory monitoring, risk assessment, and financial stability, providing a valuable tool for both academic research and practical regulatory application.

References

Alhoshan, W. F. (2023). Zero-shot learning for requirements classification: An exploratory study. *Information and Software Technology, 159*.

Bekkar, M., Djemaa, H., & Alitouche, T. (2013). Evaluation Measures for Models Assessment over Imbalanced Data Sets. *Journal of Information Engineering and Applications, Vol. 3, No.10*.

Betz, F., Oprică, S., Peltonen, T., & Sarlin, P. (2014). Predicting distress in European banks. *Journal of Banking and Finance, 45*, 225-241.

Blümke, O. (2017). Out-of-Time Validation of Default Probabilities within the Basel Accord: A Comparative Study. *SSRN*.

Climent, F., Momparler, A., & Carmona, P. (2019). Anticipating bank distress in the Eurozone: An Extreme Gradient Boosting approach. *Journal of Business Research, Volume 101*, 885-896.

Cole, R., & White, L. (2012). Déjà vu all over again: the causes of US commenrical bank failures this time around. *Journal of Financial Services Research, 42*, 5-29.

De Jesus, D., & Cássio da Nóbrega, B. (2023). Machine learning and sentiment analysis: Projecting bank insolvency risk. *Research in Economics, Volume 77, Issue 2,*, 226-238.

Demyanyk, Y., & Iftekhar, H. (2010). Financial crises and bank failures: A review of prediction methods. *Omega 38(5)*, 315-324.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR*.

Fryer, D., Strümke, I., & Nguyen, H. (2021). Shapley Values for Feature Selection: The Good, the Bad, and the Axioms. *IEEE Access (9)*, 144352-144360.

Goldstein, A., Kapelner, A., Bleich, J., & Pitkin, E. (2015). Peeking Inside the Black Box: Visualizing Statistical Learning With Plots of Individual Conditional Expectation. *Journal of Computational and Graphical Statistics, 24(1)*, 44-65.

Hægland, H. M. (2023). *Sentiment-XDistil-Uncased*. Retrieved from https://huggingface.co/hakonmh/sentiment-xdistil-uncased

Han, H., Wen-Yuan, W., & Bing-Huan, M. (2005). Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. *Advances in intelligent computing*, 878-887.

Hutto, C., & Gilbert, E. (2014). VADER: A Parsimonious Rule-based Model for. *Proceedings of the International AAAI Conference on Web and Social Media, 8(1)*, 216-225.

Jiang, T., & Zeng, A. (2023). Financial sentiment analysis using FinBERT with application in predicting stock movement. *arXiv preprint arXiv:2306.02136*.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., . . . Zettlemoyer, L. (2019). BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. *arXiv: 1910.13461*.

Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of finance, 66(1)*, 35-65.

Mayes, D., & Stemmel, H. (2014). The effectiveness of capital adequacy measures in predicting bank distress. *SUERF studies, No. 2014/1*.

Meta. (2023). *LLaMA-2-7b*. Retrieved from https://huggingface.co/meta-llama/Llama-2-7b

Mistralai. (2023). *Mistral-7B-Instruct-v0.2*. Retrieved from https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2

Nassirtoussi, A., Aghabozorgi, S., Wah, T., & Ngo, D. (2014). Text mining for market prediction: A systematic review. *Expert Systems with Applications, 41(16)*, 7653-7670.

Nielsen, F. Å. (2011). A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. *arXiv*.

Pakhale, K. (2023). Comprehensive overview of named entity recognition: Models, domain-specific applications and challenges. *arXiv:2309*.

Petropoulos, A., Siakoulis, V., Stavroulakis, E., & Vlachogiannakis, N. (2020). Predicting bank insolvencies using machine learning techniques. *International Journal of Forecasting 36(3)*, 1092-113.

Poghosyan, M., & Cihak, M. (2009). Distress in european banks: an analysis basedon a new dataset. *International Monetary Fund, working papers*, 1-37.

ProsusAI. (2021). *FinBERT*. Retrieved from https://huggingface.co/ProsusAI/finbert

Rachid, A. (2023). *FinancialBERT-Sentiment-Analysis*. Retrieved from https://huggingface.co/ahmedrachid/FinancialBERT-Sentiment-Analysis

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., . . . Liu, P. J. (2023). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *arXiv:1910.10683*.

Romero, M. (2020). *DistilRoBERTa-Finetuned-Financial-News-Sentiment-Analysis*. Retrieved from https://huggingface.co/mrm8488/distilroberta-finetuned-financial-news-sentiment-analysis

Rönnqvist, S., & Sarlin, P. (2017). Bank distress in the news: Describing events through deep learning. *Neurocomputing, volume 264*, 57-70.

Shah, A., Vithani, R., Gullapalli, A., & Chava, S. (2023). Finer: Financial named entity recognition dataset and weak-supervision model. *arXiv preprint arXiv:2302.11157*.

Sigma. (2023). *Financial-Sentiment-Analysis*. Retrieved from https://huggingface.co/Sigma/financial-sentiment-analysis

SirElkhatim, M., & Salim, N. (2015). Prediction of Banks Financial Distress. *SUST Journal of Engineering and Computer Sciences (JECS), Vol. 16, No. 1*.

Soleimanian. (2023). *Financial-RoBERTa-Large-Sentiment*. Retrieved from https://huggingface.co/soleimanian/financial-roberta-large-sentiment

Wang, S., Xu, R., Liu, B., Gui, L., & Zhou, L. (2014). Financial named entity recognition based on conditional random fields and information entropy. *International Conference on Machine Learning and Cybernetics*, (pp. 838-843).

Wheelock, D., & Wilson, P. (2000). Why do banks disappear? The determinants of US bank failures and acquisitions. *Review of Economics and Statistics, 82(1)*, 127-138.

Xia, Y. H. (2020). A DYNAMIC CREDIT SCORING MODEL BASED ON SURVIVAL GRADIENT BOOSTING DECISION TREE APPROACH. *Technological and Economic Development of Economy.*

Zhang, W., Deng, Y., Liu, B., Pan, S., & Bing, L. (2023). Sentiment analysis in the era of large language models: A reality check. *arXiv preprint arXiv:2305.15005*.

7) Annex

Accuracy: ratio of correctly predicted observations to all actual observations. In the case of Positive Accuracy, it measures the correctly predicted positive sentiment over all actual positives. The same logic follows for Neutral and Negative Accuracy.

$$Accuracy_i = \frac{Predicted_i}{Actual_i}; where\ i\ can\ be\ Positive, Neutral, Negative\ or\ Overall$$

Hit Rate: measures the proportion of actual positive instances that are correctly identified as positive by the model.

$$Hit\ Rate = \frac{True\ Positives\ (TP)}{True\ Positives\ (TP) + False\ Negatives\ (FN)}$$

False Alarm Rate: measures the proportion of actual negative instances that are incorrectly classified as positive by the model.

$$False\ Alarm\ Rate = \frac{False\ Positives\ (FP)}{False\ Positives\ (FP) + True\ Negatives\ (TN)}$$

Specificity: measures the proportion of actual negative instances that are correctly identified as negative by the model. It evaluates the model's ability to avoid false alarms by accurately classifying negatives.

$$Specificity = \frac{True\ Negatives\ (TN)}{True\ Negatives\ (TN) + False\ Positives\ (FP)} = 1 - False\ Alarm\ Rate$$