BANCA D'ITALIA
EUROSISTEMA

Mercati, infrastrutture, sistemi di pagamento

(Markets, Infrastructures, Payment Systems)

Sentiment analysis with large language models for credit risk assessment

by Manuel Cugliari and Giulio Gariano

September 2026

Number 91

BANCA D'ITALIA
EUROSISTEMA

Mercati, infrastrutture, sistemi di pagamento

(Markets, Infrastructures, Payment Systems)

Sentiment analysis with large language models for credit risk assessment

by Manuel Cugliari and Giulio Gariano

Number 91 – September 2026

The papers published in the 'Markets, Infrastructures, Payment Systems' series provide information and analysis on aspects regarding the institutional duties of Banca d'Italia in relation to the monitoring of financial markets and payment systems and the development and management of the corresponding infrastructures in order to foster a better understanding of these issues and stimulate discussion among institutions, economic actors and citizens.

The views expressed in the papers are those of the authors and do not necessarily reflect those of Banca d'Italia.

The series is available online at www.bancaditalia.it.

Printed copies can be requested from the Paolo Baffi Library: richieste.pubblicazioni@bancaditalia.it.

ISSN 2724-6418 (online)
ISSN 2724-640X (print)

Banca d'Italia
Via Nazionale, 91 - 00184 Rome - Italy
+39 06 47921

Designed and printing by the Printing and Publishing Division of Banca d'Italia

SENTIMENT ANALYSIS WITH LARGE LANGUAGE MODELS FOR CREDIT RISK ASSESSMENT

by Manuel Cugliari* and Giulio Gariano*

Abstract

Recent advancements in natural language processing techniques have significantly expanded the possibilities for integrating unstructured textual data into quantitative analyses. In this paper, large language models (LLMs) are used to construct credit sentiment indicators based on financial news sourced from the Dow Jones Factiva database. These indicators are then integrated into Banca d'Italia's In-house Credit Assessment System (ICAS), which is part of the Eurosystem Credit Assessment Framework (ECAF). The results show that LLM-driven sentiment analysis enhances the ability to distinguish solvent firms from insolvent ones, offering a robust approach to incorporating unstructured textual data into credit risk evaluation.

JEL Classification: C14, C52, C55, G21, G24, G32.

Keywords: credit risk, machine learning, sentiment analysis, large language models.

Sintesi

I recenti progressi nelle tecniche di elaborazione del linguaggio naturale hanno ampliato significativamente le possibilità di integrare dati testuali non strutturati all'interno di analisi quantitative. In questo lavoro, modelli linguistici di grandi dimensioni (Large Language Models, LLM) vengono utilizzati per costruire indicatori di sentiment creditizio, basati su notizie finanziarie tratte dalla banca dati Dow Jones Factiva. Successivamente, tali indicatori vengono integrati nel sistema interno di valutazione del merito di credito (ICAS) della Banca d'Italia, che costituisce parte dello schema di riferimento dell'Eurosistema per la valutazione della qualità creditizia. I risultati mostrano che l'analisi del sentiment basata sugli LLM migliora la capacità di distinguere le imprese solvibili da quelle insolventi, offrendo un approccio robusto per l'integrazione di dati testuali non strutturati nella valutazione del rischio di credito.

* Banca d'Italia, Financial Risk Management.

CONTENTS

1. Introduction

The acceptance of bank loans to non-financial companies has long been a cornerstone of the collateral framework employed by the Eurosystem's monetary policy operations (Auria *et al.*, 2021). To accurately assess the collateral value of bank loans, central banks rely on three credit assessment sources: rating agencies, internal ratings-based (IRB) systems operated by banks and validated by the competent supervisory authority and In-house Credit Assessment System (ICAS) developed by the national central banks themselves[1].

Banca d'Italia's ICAS consists of two components: a statistical rating model, which evaluates information from financial statements and the National Credit Register to estimate the firm's one-year probability of default (PD)[2], and an expert judgment process conducted by financial analysts (Giovannelli *et al.*, 2020). This combination yields a comprehensive company rating. The expert judgment on each firm involves several data types, including qualitative information that is available online, such as press news, interviews, reviews, and corporate websites. However, a structured analysis of qualitative information is currently not available.

Natural language processing (NLP) techniques have greatly expanded the possibilities for integrating unstructured textual data into a quantitative framework. For instance, Lu *et al.* (2016) show that sentiment scores derived from financial news significantly enhance the predictive accuracy of CDS spreads, showing the value of qualitative data in complementing traditional financial indicators. More recently, large language models (LLMs) have emerged as powerful tools, offering unprecedented capabilities in processing and interpreting complex linguistic data. These tools are useful to discern nuanced sentiment within financial narratives (Zhang and Zhao, 2024).

This study leverages state-of-the-art LLMs, including Bidirectional Encoder Representations from Transformers (BERT), Large Language Model Meta AI 3 (LLaMA 3), Microsoft's Phi 3 and Google's Gemma 2 to extract sentiment indicators from a dataset of company-related news sourced from the Dow Jones Factiva database. These models bridge the gap between qualitative and quantitative analyses, enabling a comprehensive evaluation of corporate creditworthiness. While these models provide a flexible framework for processing textual data, their outputs should not be interpreted as fully objective measures of sentiment,

[1] Not all central banks have developed an ICAS. Moreover, when an ICAS is available, it can be used for purposes beyond collateral valuation. Banque de France, for example, is recognized as an external credit assessment institution (ECAI) for its company rating activity.

[2] The definition of default used by national ICASs is quite specific. Overlooking some details, a company is considered in 'ICAS default' when non-performing exposures (bad debts, unlikely to pay and past-due loans) exceed 5 percent of the total exposure for three consecutive months, when considering exposures to the entire banking system.

but rather as model-based approximations that may reflect both the structure of the input data and the characteristics of the underlying training process.

While advanced predictive models based on large language models (LLMs) offer great potential, one concern is the risk of manipulation if the qualitative data sources they rely on become widely known. In theory, this could lead to attempts to distort the information environment in order to influence the model's outputs. However, in the specific context of our study, this risk is minimal. This is due to the nature of the data sources, the statistical aggregation method we use, the capabilities of the NLP models involved, and the institutional framework of the rating system[3].

By applying a compression technique that reduce the computational load of these models, while preserving their performance to a considerable extent, this study achieves a balance between analytical precision and computational efficiency, while also ensuring compliance with Banca d'Italia's operational and confidentiality standards. Our findings reveal that sentiment indicators derived from LLMs enhance the discriminatory power of the Banca d'Italia's ICAS and establish a robust framework for incorporating unstructured data into predictive models.

The remainder of the paper is structured as follows: Section 2 reviews the relevant literature, tracing the evolution of textual analysis in credit risk assessment. Sections 3 through 5 describe the data, the LLMs and the methodology. Section 6 presents the results. Section 7 concludes with insights into the implications of our findings and suggestions for future research.

2. Literature review

Traditional credit risk models are based on quantitative indicators derived from company balance sheets and market data (Altman, 1983). These indicators can be easily included in multivariate regressions or, more recently, in machine learning models (Altman *et al.* 2017; Ciampi and Gordini, 2013). In contrast, a wide range of qualitative information contained in notes and comments on company financial statements, in company press releases and in news articles is not exploited by the above models.

[3] Several factors help reduce this risk in our case. First, the news sources used are reputable national outlets with strong editorial and fact-checking standards. Second, sentiment analysis is based on a large and time-distributed set of articles, which helps dilute the impact of any outliers. Third, modern NLP models can detect and downweight texts that are potentially biased or of low quality. Moreover, the ICAS ratings produced by the Bank of Italy are not publicly available, which limits the incentive for targeted manipulation. Finally, in the context of Italian SMEs, economic motivations for such manipulation are weak, due to the low share of listed companies and the importance of direct, long-term business relationships.

Recent literature shows that the use of textual information can increase the accuracy of credit risk models. Lu *et al.* (2016) construct a sentiment score based on quarterly reports of US companies with quoted CDS as well as related press reports and show that this score improves the forecast of CDS spreads. Similarly, Cathcart *et al.* (2020) show that a news sentiment score obtained via the Thomson Reuters News Analytics service is a significant variable in forecasting sovereign CDS spreads.

The use of alternative textual bases such as social networks and search engines is the object of a rich literature. González-Fernández and González-Velasco (2020) use the Google Trends service to build a sentiment index for bank credit risk, in alternative to classic CDS-based indices. Lau *et al.* (2018) extract an 'emotion indicator' from Twitter messages that can complement the usual financial indicators to forecast company ratings. Bozzon *et al.* (2020) extract information from the Facebook pages of Dutch SMEs to build indicators that complement financial data in a credit risk model.

Gariano and Viggiano (2022) show that credit sentiment indicators derived from press news and tweets using a dictionary approach improve the discriminating power of the Banca d'Italia's ICAS model.

The incorporation of financial news as a source of unstructured data has been shown to enhance the performance of credit risk models. Tran-The (2020) shows that incorporating news sentiment into predictive models significantly improves the ability to forecast credit downgrades, achieving significant gains along several accuracy metrics. Similarly, Lei *et al.* (2024) propose a framework combining linguistic features from financial news with quantitative indicators, which outperforms traditional models in predicting credit risk. These findings show the growing importance of textual data, particularly press news, in complementing conventional financial indicators.

In parallel, the application of NLP techniques has emerged as a pivotal advancement in harnessing unstructured textual data for credit risk assessment. Lis *et al.* (2023) employ NLP-based clustering methods to validate credit risk models, demonstrating their efficacy in uncovering latent patterns and addressing structural inconsistencies. These techniques have proven to be more effective than traditional approaches, especially in processing large-scale textual datasets. Dolphin *et al.* (2024) further show the potential of augmented NLP methods in extracting structured insights from financial news, offering innovative ways to build sentiment indicators that enrich the analytical framework of credit risk models.

More recently, LLMs have shown a better performance than classical NLP techniques in analysing textual data for credit risk applications. Sanz-Guerrero and Arroyo (2024) employ BERT to extract sentiment indicators from loan descriptions on peer-to-peer lending platforms, demonstrating superior predictive power compared to traditional NLP models. Mo *et al.* (2024) fine-tune the Gemma 2-7B model for analysing

financial headlines, achieving significant improvements in accuracy over existing NLP-based classifiers. These advancements underline the ability of LLMs to capture complex linguistic nuances, providing deeper insights into financial narratives and enhancing the predictive capabilities of credit risk models.

3. Data

We employ three data sources: ICAS, InfoCamere, and Factiva. ICAS provides credit risk quantitative indicators and default information on a monthly basis for a sample of Italian companies. InfoCamere offers detailed company registry information. Factiva supplies news, articles and data from journalistic sources, news agencies and business publications on a daily basis.

To enable a comparison with previous work (Gariano and Viggiano, 2022), the analysis spans the same five-year period, from January 2016 to December 2020.

3.1 ICAS

The sample companies are those that receive a comprehensive ICAS rating in the survey period: about 6,000 companies, of which 500 became insolvent, according to the ICAS definition, at least once in the sample period (Table 1)[4]. The one-year average default rate in the sample is around 1.6 percent (roughly 100 companies defaulting each year). Taking into account the monthly frequency of ICAS data (each company is considered for the 60 months in the sample period[5]), the number of records is above 350,000.

Table 1. Companies dataset

(2016-2020)

Number of companies	6,006
Number of companies defaulting at least once in sample	500
Number of records	351,755
Number of records of defaulting companies	5,490

The dataset includes, for each company and reference date: (i) the statistical ICAS score and (ii) the default 'flag' of the company. The former is a monotone function of the probability of default, while the latter is a

[4] Unlike bankruptcy, insolvency is a reversible event.
[5] Some company are considered for less than 60 months, for example if they were rated only after January 2016 or if they defaulted before December 2020.

binary variable indicating whether the company defaults within the subsequent 12 months (1 for default, 0 otherwise).

The dataset predominantly features medium and large companies, which represent 80 percent of the sample. About 75 percent of these firms are located in Northern Italy, in line with the geographic concentration of economic activity within the region (Table 2).

Table 2. Company distribution by size and geographical area

(percentage values)

	North-West	**North-East**	**Centre**	**South**	Total
Micro	3	2	1	1	7
Small	6	4	2	1	13
Medium	19	16	7	5	48
Large	13	11	5	3	32
Total	41	34	15	10	

3.2 InfoCamere

Infocamere offers detailed company registry information, including the list of managers of the companies in our sample. The names of the managers are important since some of them might be involved in legal proceedings cited in news articles. As shown in Gariano and Viggiano (2022), news articles on legal proceedings are among the most relevant news for credit analysis.

On average, approximately 11 managers by company are identified in the sample period, for a total of approximately 66,000 managers (Table 3).

Table 3. Managers dataset

Number of managers	66,415
Minimum number of managers by company	1
Maximum number of managers by company	62
Average number of managers by company	11

3.3 Factiva

The Factiva database contains approximately 800,000 press articles per year, obtained from over 60 news sources. This dataset is extracted from the Dow Jones Factiva service, applying a preliminary filter to include

only articles related to economics and finance, thereby making its size manageable for our purposes (for a similar approach, see Aprigliano *et al*., 2021).

Each article provides access to the title, summary, main text, date, and source, along with some metadata designed to facilitate analysis, including a list of companies mentioned in the article.

To ensure that the analysis considers only relevant articles, some exclusions are applied, such as: (i) articles containing an extensive number of numerical values, and (ii) articles referencing a large number of companies. Such filters are implemented to exclude generic articles, such as those summarising daily stock market closures, which are frequent in the database but not informative for the purposes of this study.

Following this initial screening, all articles mentioning one of the companies in our sample are identified. As in Gariano and Viggiano (2022), this identification relies on metadata analysis when available. Where metadata are not provided, a raw search of the company name or company managers within the text is conducted.

Table 4 provides a breakdown for each year of the total number of articles available, the number of articles mentioning at least one company in our dataset, and the number of distinct company-date pairs. While the total number of articles exceeds 4 million, only 211,435 of them (about 5 percent) explicitly mention at least one company from our sample. Since companies, when mentioned, typically appear in multiple articles within the same month—on average, eight articles per month—the number of distinct company-month pairs is even lower (26,829). This measure is particularly relevant, as the default flag is defined on a monthly basis, meaning that all articles referencing the same company within the same month must be aggregated. Overall, we find news for 2,772 companies, out of 6,006 companies in our dataset.

Table 4. Factiva dataset

Year	Number of articles	Number of articles on sample companies	Number of company-date pairs
2016	590,117	23,134	3,587
2017	612,303	31,011	4,777
2018	659,798	43,981	5,994
2019	1,138,076	56,796	6,510
2020	1,255,528	56,513	5,961
Total	4,255,822	211,435	26,829

Before processing the articles from the Factiva database with LLM for inference, a comprehensive pre-processing pipeline is implemented to ensure data quality and consistency. Using the SpaCy[6] library, several steps are undertaken to clean and structure the text.

4. Large language models

LLMs are a cornerstone in the evolution of NLP and artificial intelligence. Their inception dates to the emergence of pre-trained transformer-based architectures[7], most notably BERT, introduced by Devlin *et al.* in 2019. BERT pioneered the use of bidirectional context during training, marking a paradigm shift in text representation and understanding. Since then, the field has witnessed the rapid proliferation of more advanced and larger models, including LLaMA 3, Phi 3, and Gemma 2, each making advancements to the state of the art. These innovations testify the relentless drive towards models capable of capturing increasingly complex linguistic structures, enabling them to understand and generate text with human-like proficiency.

The transformative capabilities of LLMs lie in their ability to leverage vast training datasets. By ingesting trillions of tokens from diverse sources, they develop a sophisticated understanding of syntax, semantics, and pragmatics. This pre-training phase equips LLMs with the ability to generalize across domains, making them adaptable to a wide range of downstream tasks. One area where LLMs have demonstrated exceptional utility is sentiment analysis. Financial news, in particular, present a unique challenge due to the specialized language, the nuanced sentiment conveyed, and the significant implications for decision-making. LLMs not only extract sentiment with remarkable precision, but also integrate these insights into predictive models for credit risk

[6] SpaCy is an advanced open-source library for NLP in Python, designed for efficient and scalable text processing. It provides a wide range of features, including tokenization, named entity recognition, part-of-speech tagging, dependency parsing, and text vectorization, making it a versatile tool for linguistic analysis and pre-processing tasks. Special characters, extraneous whitespace, and HTML tags are removed to eliminate noise that could affect model performance. URLs and embedded hyperlinks are stripped, while redundant or irrelevant metadata commonly found in Factiva articles, such as publication dates and author credits, are filtered out. Furthermore, text encoding inconsistencies are resolved to standardize the input format. This preparation ensures that the text fed into the LLM is optimized for accurate sentiment analysis, to extract meaningful insights from company-related news.

[7] Transformer-based architectures are a class of neural networks that rely on self-attention mechanisms to process and generate sequences of data. They have become the foundation for modern natural language processing models due to their ability to capture contextual relationships between words in a sequence efficiently, regardless of distance, making them highly effective for tasks such as machine translation, text summarization, and sentiment analysis. Attention mechanisms are a core element of transformer models. They enable the model to focus on the most relevant parts of a sentence when interpreting its meaning, by assigning higher importance to certain words depending on the context. This allows the model to capture relationships between words even when they are far apart in the text. See Vaswani *et al.* (2017) for further details.

assessment, enhancing the discriminatory power of traditional metrics (Chen *et al*., 2023; Zhang & Zhao, 2024).

Despite their strong performance, LLMs also present a number of limitations that are relevant in this context. First, their outputs may be sensitive to the phrasing of the input and to the specific configuration parameters, even when these are carefully standardised. Second, although recent models exhibit improved robustness, they may still produce inconsistent classifications in the presence of ambiguous or context-dependent language, which is common in financial news. Third, the training data of LLMs, often dominated by general-purpose corpora, may not fully capture the specificities of financial reporting, potentially affecting the interpretation of domain-specific expressions. These aspects suggest that LLM-based sentiment indicators should be interpreted with caution and, where possible, combined with other sources of information.

Sentiment analysis in finance requires a granular approach. The language employed in financial news is replete with subtleties and implicit cues that traditional models often fail to capture. LLMs excel by leveraging their contextual awareness, trained on extensive *corpora* that include financial texts. This contextual depth allows them to discern the latent sentiment embedded in narratives.

The architecture of an LLM is governed by several settings that define its functionality and adaptability. Context length is a pivotal factor, determining the volume of preceding text the model can consider when generating responses. A longer context window enables the model to maintain coherence over extended narratives. This is particularly valuable in financial discourse where connections between events and implications often span multiple sentences or paragraphs. Tokenization is another foundational aspect, converting text into discrete units, or tokens[8], that the model processes. This step ensures that the model can handle complex linguistic constructs while maintaining computational efficiency. Temperature, that controls the randomness of output generation, allows fine-tuning of the model responses. Adjusting the temperature can yield output that range from deterministic to creative, a flexibility that is especially useful in exploratory sentiment analysis.

LLMs are computationally demanding. Their inference requires substantial processing power, posing challenges for real-time applications or deployments in resource-constrained environments. The sheer number

[8] In NLP, a token is a discrete unit of text, such as a word, sub word, or character, that serves as the fundamental element for analysis and processing in language models. Tokenization is the process of splitting text into these units to enable computational manipulation and interpretation.

of parameters in these models, often running into billions, necessitates high memory bandwidth and parallel processing capabilities. A pragmatic solution to this requirement is quantisation, a compression technique that maps high precision parameters to lower precision ones, reducing the computational load while preserving the model performance to a considerable extent. This study employs quantisation techniques to enable the deployment of advanced models like Gemma 2 and Phi 3 within operational constraints, achieving a balance between accuracy and efficiency. This approach, however, involves a trade-off. Lower precision representations may lead to a partial loss of information, potentially affecting the stability and accuracy of the model outputs, especially in borderline classification cases. In this study, quantization is adopted as a practical compromise between computational feasibility and analytical performance, in line with the operational constraints of the application. The evolution of LLMs is intrinsically tied to advancements in hardware and software. The advent of graphics processing units (GPUs)[9] and tensor processing units (TPUs)[10] has been instrumental in scaling these models, enabling training and inference at an unprecedented scale. Concurrently, software frameworks like PyTorch[11] and TensorFlow[12] have streamlined model development, fostering innovation and accessibility.

One notable challenge in deploying LLMs for non-English contexts lies in their training data composition. Most LLMs are predominantly trained on English-language *corpora*, leading to potential biases and limitations when applied to other languages. This study addresses this gap by employing multilingual models which offer robust support for Italian language alongside other languages, ensuring relevance and accuracy in the analysis of Italian company-related news.

This study leverages four state-of-the-art LLMs: BERT, LLaMA 3, Phi 3 and Gemma 2. These models are selected for their advanced capabilities, robust architecture, and proven performance in sentiment analysis. Their training on extensive and diverse datasets ensures adaptability to the intricacies of financial language,

[9] Graphics Processing Units (GPUs) are specialized hardware initially designed for rendering graphics but now widely adopted in parallel computing tasks, including training and inference of deep learning models due to their ability to handle large-scale matrix operations efficiently.
[10] Tensor Processing Units (TPUs) are application-specific integrated circuits (ASICs) developed by Google, designed to accelerate machine learning workloads, particularly for tensor-based operations in neural networks, offering both efficiency and scalability for large models.
[11] PyTorch is an open-source machine learning framework developed by Facebook's AI Research lab. It provides flexibility for building and training deep learning models, offering dynamic computation graphs and strong support for research and production environments.
[12] TensorFlow is an open-source machine learning framework developed by Google Brain. It is widely used for developing, training, and deploying machine learning models, offering tools for scalable computation on a range of platforms from mobile devices to high-performance clusters.

while their tuning for the Italian context addresses the linguistic challenges inherent in this domain. By combining these models, this study harnesses their collective strengths.

To ensure data confidentiality and compliance with the Banca d'Italia's operational standards, all computations involving the LLMs used in this study are conducted locally. This approach mitigates the risk of sensitive information being transmitted to external servers, thereby complying with stringent data protection protocols. Inference for the LLMs is conducted with LM Studio[13] 0.3.5, a robust platform for deploying LLMs in localised environments. The inference process complies with OpenAI standards[14], ensuring consistency and precision in the model's output. The computational experiments involving all LLMs employed in this study are conducted on an NVIDIA® GeForce RTX™ 4080[15], a high-performance graphics processing unit (GPU) with advanced computational capabilities.

Overall, the models employed in this study should be viewed as practical tools for extracting structured information from unstructured text, rather than as fully reliable representations of underlying sentiment.

The following sub-sections provide essential information on the four LLMs used in the study.

4.1 BERT

BERT is a major milestone in the development of natural language processing and large language models. Introduced by Devlin *et al.* in 2019 as part of Google's research efforts, BERT changed the way text is represented by applying a bidirectional training method, which allows the model to learn context from both earlier and later words in a sentence. This advancement overcomes the limits of traditional unidirectional models and enables a better understanding of linguistic structure, idioms, and contextual meaning.

---

[13] LM Studio is an open-source platform designed for deploying and interacting with localized LLMs. It provides a user-friendly interface and robust backend capabilities for running models on constrained hardware, supporting advanced configurations such as quantization and parameter optimization. LM Studio facilitates efficient inference while maintaining data privacy by ensuring all computations occur locally, eliminating the need for cloud-based processing.

[14] OpenAI standards refer to a set of best practices and protocols for implementing LLMs, ensuring consistency, reliability, and reproducibility in their output. These standards encompass guidelines for model configuration, including temperature settings, token limits, and inference batch sizes, enabling developers to achieve optimal performance tailored to specific tasks.

[15] The NVIDIA GeForce RTX 4080 is a high-performance graphics processing unit (GPU) designed for demanding computational tasks, including machine learning and deep learning applications. Equipped with 16GB of GDDR6X memory and advanced tensor cores, it delivers exceptional performance in handling complex models and large datasets. The RTX 4080 supports efficient parallel processing and is optimized for AI frameworks, making it a suitable choice for tasks such as LLM inference and training.

The main idea behind BERT is based on its two training tasks: masked language modelling (MLM) and next sentence prediction (NSP). In MLM, some words in a sentence are masked, and the model learns to predict them from surrounding context, improving its knowledge of grammar and meaning. NSP helps the model understand if two sentences are connected, reinforcing its ability to model relationships in text. Combined with the attention mechanisms of the transformer, these tasks help BERT capture detailed connections between words and phrases.

Several studies highlight BERT's strong performance in extracting sentiment from a variety of textual sources, including financial news and social media. For instance, Chen *et al.* (2023) show that BERT-based models are effective in identifying subtle sentiment changes in financial texts, helping to anticipate market movements. Likewise, Yang and Yu (2020) show that it can detect fine sentiment shifts, which is especially important in credit risk analysis where small changes in tone may carry significant meaning.

In this study, a multilingual version of BERT, fine-tuned for sentiment analysis, is employed. This model, available through the Hugging Face[16] platform, supports six languages: English, Dutch, German, French, Spanish, and Italian. Its adaptability ensures the seamless integration of non-English textual data, a crucial feature given the multilingual nature of financial reporting and news coverage.

4.2 LLaMA 3

LLaMA 3 represents a recent advancement in Meta's family of large language models, designed to deliver state-of-the-art performance while maintaining computational efficiency. As shown by Touvron *et al.* (2024), LLaMA 3 introduces architectural innovations that optimise the balance between model size and usability, addressing the increasing demand for powerful yet accessible LLMs. It is designed with a decoder-only architecture[17], that simplifies the inference process by focusing on generating output conditioned solely on the input sequence, reducing the computational overhead that is typically associated with bidirectional architectures.

---

[16] Hugging Face is an open-source platform and library for natural language processing, offering pre-trained language models, tools, and datasets. It facilitates the development, fine-tuning, and deployment of LLMs for a wide range of applications, enabling researchers and practitioners to leverage state-of-the-art NLP technologies efficiently.

[17] Decoder-only architecture refers to a transformer design where only the decoder component is used, focusing on generating output conditioned on the input sequence. Unlike the encoder-decoder architecture, which processes input through an encoder before the decoder generates the output, decoder-only models are streamlined for tasks like text generation, reducing computational overhead while maintaining contextual understanding.

The 8-billion parameter[18] variant of LLaMA 3, employed in this study, is a multilingual fine-tuned version of the model, as outlined in Devine (2024). This specific configuration offers a balanced solution, combining computational feasibility with the advanced capabilities required for sophisticated text analysis.

To adapt LLaMA 3 for local deployment, a quantised version with 4-bit precision is employed, significantly reducing its computational demands. This precision enables the model to perform inference efficiently without compromising on accuracy. The model configuration (see Appendix 1 for more details) is carefully calibrated to optimise its output for the specific use case. The temperature is set at 0.30, prioritising deterministic responses while allowing limited variability in the output generation process. It should be noted that the notion of temperature is relevant for decoder-only models, where it directly controls token sampling during text generation. In contrast, encoder-only models such as BERT do not generate text and therefore do not rely on temperature in the same way; their outputs are deterministic conditional on the input and model parameters. In our framework, BERT produces probabilistic scores over sentiment categories, which are not affected by sampling variability.

4.3 Phi 3

Phi 3 is a key innovation within Microsoft's Phi family of models, that makes a significant leap in balancing computational efficiency with linguistic precision (Microsoft, 2024). Phi 3 offers multiple configurations ('variants') tailored to various computational and analytical needs, ranging from smaller models optimized for efficiency to larger models designed for high precision. The medium variant used in this study, featuring 14 billion parameters, strikes a balance between these extremes, providing advanced capabilities for resource-constrained environments.

The architecture of Phi 3 is rooted in a decoder-only structure, optimised for generating coherent and contextually accurate text output. Unlike bidirectional models, the decoder-only approach prioritises streamlined processing, focusing on efficient generation without compromising on contextual understanding.

[18] In LLMs, parameters are the learned weights that define how the model processes and generates text. These values are adjusted during training to optimise the model's predictions, enabling it to capture linguistic nuances and complex patterns. The number of parameters directly influences the model's capacity to handle intricate tasks; larger models with billions of parameters generally show greater precision and versatility. However, this also increases computational requirements for training and inference. The 8-billion parameter variant of LLaMA 3 is a carefully balanced configuration, offering significant analytical power while maintaining efficiency, making it suitable for applications that demand both accuracy and resource-conscious deployment.

The open design of Phi 3 emphasizes its versatility, offering pre-trained and instruction-tuned variants to suit a wide range of applications.

The Phi 3 model employed here features an 8-bit quantisation setup, a strategic choice to minimise computational overhead. Specific settings (see Appendix 1 for more details) are optimally configured to enhance the model performance for the objectives of this study. The temperature is set at 0.20, ensuring deterministic and precise output with minimal randomness.

4.4 Gemma 2

The last of the four LLMs employed in this study is developed by Google. This family of lightweight, state-of-the-art LLMs stems from the same research and technology behind the Gemini models[19], though tailored for broader accessibility. As a text-to-text, decoder-only LLM, Gemma 2 is specifically designed to excel in diverse tasks such as question answering, summarisation, and logical reasoning. Its design prioritises versatility and scalability, allowing for the deployment on limited-resource environments. This approach fosters innovation while ensuring broader access to cutting-edge AI technologies (Gemma Team, Google DeepMind, 2024).

The Gemma 2 model used in this study, with 27 billion parameters, strikes a balance between scale and practicality, offering high performance without the computational overhead of larger models. Its open-source nature enhances its adaptability, providing both pre-trained and instruction-tuned variants to suit a variety of use cases. To optimise the local deployment of Gemma 2, a quantised version with 4-bit precision is employed. Key configurations are fine-tuned to optimise performance for this study (further details on the iterative optimisation process and sensitivity to parameter choices are provided in Appendix 1). The temperature is set at 0.25, minimising randomness in the model responses and prioritising determinism.

5. Methodology

This section describes how we construct credit sentiment indicators derived from articles with the LLMs described in sections 4.1-4.4: BERT, LLaMa, Phi and Gemma[20].

---

[19] Gemini is Google's advanced generative AI systems designed for a wide range of applications, including conversational AI, content creation and complex reasoning.

[20] For brevity, the specific versions of the LLMs are omitted in the text from this point onward. However, all references to these models pertain exclusively to the configurations described in Sections 4.2–4.4.

We ask LLMs to classify each article as: ‘very good’, ‘good’, ‘neutral’, ‘bad’ or ‘very bad’. For completeness, and to assess the distribution of sentiment classifications across categories, Appendix 3 reports the empirical frequency distributions for each model.

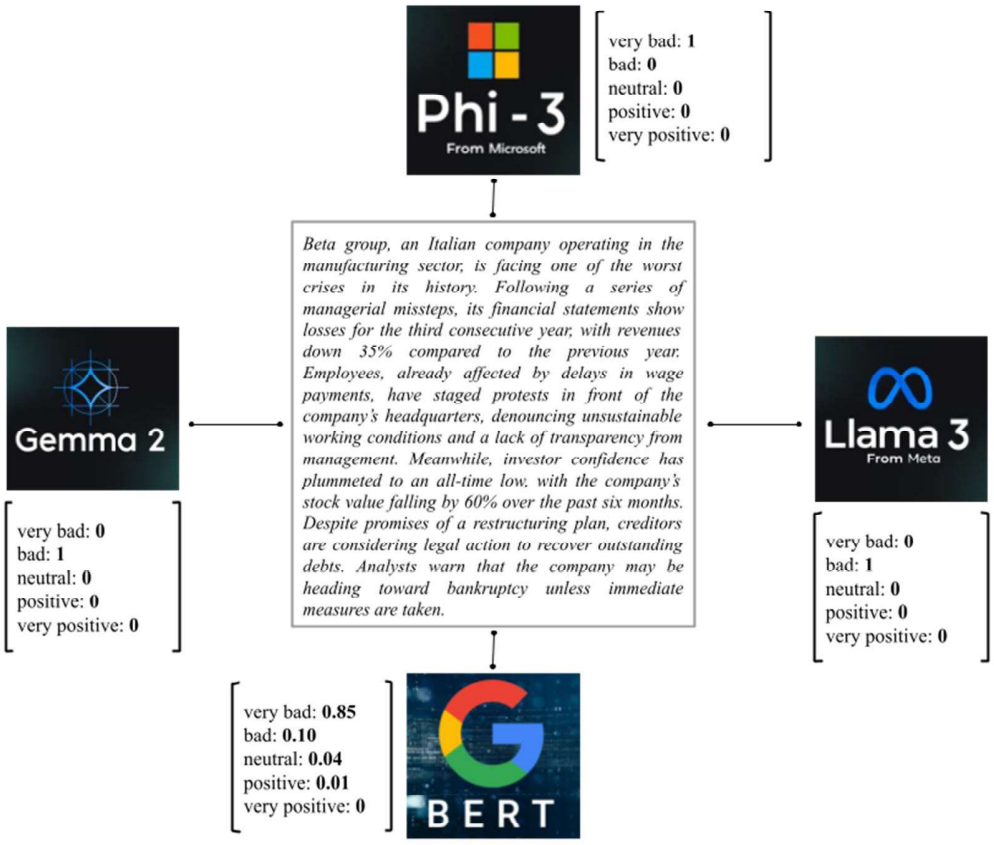
When the length of the article exceeds the LLM's maximum length, the text is divided into shorter segments to ensure compatibility with the model input constraints. Each segment is processed independently by the LLM, generating separate sentiment predictions. To obtain a final sentiment prediction for the entire *corpus*, an ensemble[21] method is employed, aggregating the predictions from all segments. This approach allows the model to handle extensive textual data while maintaining the integrity of the analysis.

While the output of LLaMa, Phi and Gemma is the actual classification, the output of BERT is a vector of five probabilities: the probability that the article is ‘very good’, the probability that it is ‘good’, etc. In Figure 1[22] we present a comparative example illustrating the sentiment classification outcomes generated by the LLMs. Gemma and LLaMA classify the example article as ‘bad’, while Phi categorizes it as ‘very bad’, showing a slight divergence in sentiment interpretation, possibly due to variations in their fine-tuning strategies or training *corpora*. BERT, unlike the other models, provides probabilistic output for each sentiment category. In this instance, the highest probability, 85 per cent, is assigned to ‘very bad’, showing a strong agreement with Phi while also offering a more nuanced breakdown of sentiment probabilities across all categories.

[21] Ensemble refers to a technique in machine learning where multiple models or predictions are combined to improve overall accuracy and robustness. In the context of sentiment analysis with LLMs, ensemble methods aggregate predictions from text segments, ensuring that the final output reflects the broader context while minimizing biases or inconsistencies from individual predictions. This approach is widely used to enhance performance in complex tasks. In this study, greater weight is assigned to segments classified with negative sentiment, as adverse signals in financial news are often more informative for credit risk assessment. This weighting approach ensures that potential distress signals receive appropriate emphasis in the final sentiment indicator.

[22] The original prompt submitted to the LLMs is in Italian. It has been translated into English to ensure linguistic uniformity throughout the text.

Figure 1 – Illustration of sentiment classification outcomes produced by LLMs



For convenience, we treat the output of LLaMa, Phi and Gemma as a vector of five binary variables, one for each state, all equal to zero except one. Therefore, for each article, we have 20 variables: 5 probabilities from BERT, 5 binary variables from LLaMA, 5 binary variables from Phi, 5 binary variables from Gemma. In the following, we refer to these variables as $X^{i,j,a}$, where $i \in$ {bert, llama, phi, gemma}, $j \in$ {very bad, bad, neutral, good, very good} and a indexes the article.

Before moving on, we need to address the potential issue of hallucination. Hallucination occurs when an LLM generates output that appears plausible but is factually incorrect or inconsistent with the input data (Huang *et al.*, 2023). In the context of sentiment analysis for credit risk assessment, this issue could compromise the reliability of the predictions, particularly when dealing with unstructured textual data. To mitigate this concern, a robustness check is implemented by comparing sentiment predictions derived from two levels of textual representation: the full text (*corpus*) of articles and their corresponding snippets (the concise summaries) provided by the Dow Jones Factiva database. This approach allows for an assessment of

consistency in sentiment classification across different textual granularities, thereby reducing the risk of spurious output. The results of this analysis are presented in Appendix 2.

In the next step, we aggregate the information from different articles. Since the default flag is defined on a monthly basis, we need to aggregate all articles mentioning the same company in the same month. Additionally, one could assume that older articles (e.g. those of the previous month, or those of two months before, etc.) are predictive of the default as well. In line with Gariano and Viggiano (2022), we aggregate the articles of the last six months. This means that each article is used for six consecutive months to predict the defaults of the company. The length of the time window (six months) is chosen taking into account the discriminatory power of the resulting model.

The aggregation is based on the average and maximum functions. More specifically, for each model *i*, for each state *j*, for each company *c*, and for each month *m* we compute:

$$A^{i,j,c,m} = \text{Average } (X^{i,j,a})$$

$$M^{i,j,c,m} = \text{Max } (X^{i,j,a})$$

where the average and maximum functions are computed on the articles of the last six months (from month *m* until month *m-5*, included) mentioning company *c*.

The purpose of the 'average' variables is to capture the average sentiment of the available articles, while the purpose of the 'maximum' variables is to detect the presence of extreme articles, which might otherwise get diluted among several neutral articles.

For Gemma, LLaMA and Phi, the interpretation of these variables is straightforward: the variables $A^{i,j,c,m}$ show the fraction of articles mentioning company *c* in the last six months that model *i* classifies in state *j*, while the variables $M^{i,j,c,m}$ are binary variables which are equal to 1 only if in the last six months there is at least one article (mentioning company *c*) classified by model *i* in state *j*.

For BERT, whose output is a set of probabilities, the interpretation is less straightforward[23]: the variables $A^{bert,j,c,m}$ are the average probabilities that the articles mentioning company *c* in the last six months are of type

[23] Alternatively, we could have converted the five probabilities of BERT into a single categorical variable looking at the highest probability (e.g. 'very bad' in the example of Figure 1). This approach, however, would be recommended only if BERT tends to assign high probability to one state and negligible probabilities to the other four (like in the example of Figure 1). Unfortunately, this is not the case: in the large majority of cases BERT tends to produce similar probabilities for different states: the highest probability of the five is lower than 30 per cent for more than half of the sample, and it is lower than 50 per cent for 95 per cent of the sample.

j, while the variables $M^{bert,j,c,m}$ are the maximum probabilities that the articles mentioning company c in the last six months are of type j.

Other aggregation approaches are also considered, such the use of the mode[24] or the conversion of the categorical variable into a scalar[25]. These alternatives have been discarded due to their lower discriminatory power compared to the proposed approach.

6. Results

In section 6.1-6.4 we analyse the models on a standalone basis, while in section 6.5 we analyse the models jointly. We focus on discriminatory power since the predictive power, the other relevant feature of a credit assessment system, is not significantly affected by LLMs[26].

We find that, when considered individually, the BERT model does not add information to the statistical ICAS, while the other three models (marginally) do. However, the best model, capable of materially improving the ICAS, is the LLMs ensemble obtained by combining all the LLMs together.

The dataset used for the estimation consists of 60,571 records, of which 967 are defaults[27]. These records are a relatively small fraction (about 17 per cent) of all records in the original ICAS dataset (which are 351.755, see Table 1). This is mainly due to two reasons:

(i) 54 per cent of companies are never mentioned in the news articles available in the database;

(ii) the remaining 46 per cent of companies are mentioned, but not for all dates in our sample (on average, taking into account the six months' window of the articles, each company is mentioned in 22 out of 60 months).

---

[24] Under this alternative approach, for each company and for each month, we take the classification ('very bad', 'bad', etc..) that appears more often. For BERT, we take the classification with the highest aggregate probability.

[25] Under this alternative approach, for each article, the classification is converted into a scalar (eg. 'very bad' = 2, 'bad' = 1, 'good' = -1, 'very good' = -2, or other values). For BERT, the scalar is obtained as weighted average. Subsequently, the average/maximum functions are computed on these scalars.

[26] Broadly speaking, discriminatory power is the ability of a system to sort the companies by riskiness, while predictive power is the ability of a system to assign probabilities of default aligned with the observed default rates. The former is generally measured by the area under the ROC curve (AUROC), while the latter is usually verified with binomial tests.

[27] Consistently with the use of qualitative information in expert judgment, the absence of news is treated as absence of textual information rather than as neutral sentiment. Therefore, records without news were removed from the sample.

Since sometimes the LLM fails to classify an article, the dataset shrinks a little further in the next sections, depending on the model used.

In the following sections we use a stepwise regression to select the significant variables (we only keep variables with a *p-value* < 5 per cent). Furthermore, variables with the wrong sign[28] are removed from the estimation as well. Variables selected by the stepwise procedure but displaying a sign inconsistent with their ex ante economic interpretation are excluded from the final specification. This restriction is not intended to maximise model performance, but to preserve the interpretability of the sentiment indicators in a credit assessment framework. In particular, negative sentiment variables are expected to be positively associated with default probability, while positive sentiment variables are expected to display the opposite sign. Variables violating this condition are therefore considered difficult to interpret from an economic standpoint and are not retained in the baseline specification[29].

We use pre-trained versions of LLMs, without a fine tuning on our default data. Therefore, our dataset can be considered completely out-of-sample.

6.1 BERT

In this section we show the results for the BERT model described in section 4.1.

We consider the two following logistic models:

$$g(P[def_c = 1]) = \alpha_0 + \sum_j \left(\beta_j \cdot A^{bert,j,c} + \gamma_j \cdot M^{bert,j,c}\right) \quad (1)$$

$$g(P[def_c = 1]) = \alpha_0 + \alpha_1 \cdot S_c + \sum_j \left(\beta_j \cdot A^{bert,j,c} + \gamma_j \cdot M^{bert,j,c}\right) \quad (2)$$

where c is the company, *def* is the default binary variable (equal to 1 if and only if an ICAS default has occurred), $g(\pi) = log\left(\pi/(1-\pi)\right)$ is the logistic function, S is the statistical ICAS score, $A^{bert,j,c}$ and $M^{bert,j,c}$

---

[28] Sign must be positive for $j \in$ {very bad, bad} and negative for $j \in$ {good, very good}. Any sign for j = neutral is considered acceptable.

[29] Among the four LLMs, only BERT produced variables with incorrect signs, which were therefore removed. The other three LLMs (LLaMA, Phi, and Gemma) did not exhibit variables with incorrect signs. Consequently, their corresponding sections would remain unchanged even under a pure stepwise selection procedure.

are the two credit sentiment indicators described in section 5, with j ∈ {very bad, bad, neutral, good, very good}.

The first model uses the credit sentiment indicators to predict the defaults, while the second model uses both the statistical ICAS and the credit sentiment indicators. The purpose of the first model is to test whether the credit sentiment indicators built with BERT have good predictive power, while the purpose of the second model is to test whether the credit sentiment indicators, in addition to having good predictive power, are also able to add discriminatory power to the statistical ICAS.

BERT fails to classify 14 per cent of the articles. As a result, the dataset used for the regressions shrinks by 3 percent to 58,953 records[30], of which 962 are defaults.

Table 5 summarizes the results, reporting estimates, standard errors and p-values for the selected variables and the area under the ROC curve (AUROC)[31].

Table 5. BERT models

	def	
	(1) (News)	(2) (News + ICAS)
α_0	-4.0805*** (0.1391)	1.1653*** (0.1805)
α_1		1.2145*** (0.0279)
β_{bad}	3.4645*** (0.4098)	2.2640*** (0.4466)
$\beta_{neutral}$	-3.6995*** (0.7017)	-4.6090*** (0.7597)
AUROC	0.580	0.864
Note: * p-value < 0.01; ** p-value < 0.001; *** p-value < 0.0001		

[30] An article in six months is enough to keep a record, therefore a 14 per cent reduction in the number of articles results in a much lower (3 per cent) reduction in the number of records.

[31] The ROC curve (receiver operating characteristic curve) is a graphical representation of model performance. It plots the true positive rate versus the false positive rate at various thresholds. The area under the ROC curve (AUROC) summarizes the performance with a number between 0 and 1, where a model whose predictions are 100 percent wrong achieves 0, while a model whose predictions are 100 percent correct achieves 1. A random classifier achieves an AUROC of 0.5, which is considered a lower bound for any classification model. Credit scoring models based on financial statement data typically reach AUROC values around 0.7; the ICAS statistical model, using both financial statement and Credit Register data, reaches a value around 0.86.

The first model shows weak performance, with an AUROC of only 58 per cent. An analysis of the variables confirms this outcome. Maximum variables ($M^{bert,j,c}$) are not selected. Average variables ($A^{bert,j,c}$) are selected only for 'bad' and 'neutral' articles. The fact that we have a 'bad' variable without a 'very bad' variable is suspicious (in principle, 'very bad' articles should be more informative in predicting default than simply 'bad' articles). The fact that we have a 'neutral' variable with negative sign (it reduces the probability of default) without 'good' and 'very good' variables is suspicious as well (in principle, 'good' and 'very good' articles should be more informative in predicting non-default than simply 'neutral' articles).

The second model considers the BERT variables in addition to the statistical ICAS. The selected variables are the same of the previous model: only the average variables for 'bad' and 'neutral' articles add information to the statistical ICAS: the overall AUROC increases by 0.2 per cent (from 86.2 to 86.4 per cent) with respect to the statistical ICAS alone.

6.2 LLaMA

In this section we show the results of the LLaMA model described in section 4.2. We consider the same logistic models of previous section, except that LLaMA variables are used instead of BERT ones:

$$g(P[def_c = 1]) = \alpha_0 + \sum_j \left(\beta_j \cdot A^{llama,j,c} + \gamma_j \cdot M^{llama,j,c}\right) \quad (3)$$

$$g(P[def_c = 1]) = \alpha_0 + \alpha_1 \cdot S_c + \sum_j \left(\beta_j \cdot A^{llama,j,c} + \gamma_j \cdot M^{llama,j,c}\right) \quad (4)$$

LLaMA fails to classify 17 per cent of the articles[32]. As a result, the dataset used for the regressions shrinks by 4 per cent to 58,166 records (a few hundred records less than the BERT one), of which 944 are defaults.

Table 6 summarizes the results, reporting estimates, standard errors and p-values for the selected variables and the AUROC.

[32] Failure to classify an observation occurs when the model does not return a valid label among the predefined sentiment categories. In practice, this may happen for different reasons depending on the model architecture. For decoder-only models (LLaMA, Phi, Gemma), this typically corresponds to outputs that do not match any of the expected categories (e.g. longer textual responses, ambiguous statements, or empty outputs), despite the prompt requesting a single label. For BERT, which produces probabilistic scores over the five categories, failure may occur when the output cannot be properly mapped into a valid classification due to technical issues in the inference process or missing predictions. In all such cases, the observation is excluded from the analysis.

Table 6. LLaMA models

	def	
	(3) (News)	(4) (News + ICAS)
α_0	-3.8643*** (0.0749)	0.6683*** (0.1063)
α_1		1.2129*** (0.0280)
$\beta_{neutral}$	-0.3024 (0.1310)	
$\beta_{very\ good}$	-0.8601*** (0.1172)	-0.5150*** (0.1215)
$\gamma_{very\ bad}$	0.3267*** (0.0704)	0.3736*** (0.0733)
γ_{good}	-0.2593*** (0.0665)	
AUROC	0.604	0.868
Note: * p-value < 0.01; ** p-value < 0.001; *** p-value < 0.0001		

Four variables are selected for the first model: one for the 'very good' articles, one for the 'good' articles, one for the 'neutral' articles and one for the 'very bad' articles. The variable for the 'neutral' articles has negative sign (it reduces the probability of default), and – unlike the previous BERT models - its presence is justified by the fact that 'good' and 'very good' articles are also considered. The fact that the coefficient of the 'very good' articles is bigger (in absolute value) than that of the 'neutral' articles is expected as well (in principle, 'very good' articles should be more informative in predicting non-default than 'neutral' articles). The AUROC is 60.4 per cent, better than BERT but still not high.

When considering LLaMA variables together with the statistical ICAS, only two variables remain: one for the 'very good' articles and one for the 'very bad'. The less extreme articles ('neutral', 'good') add no information to the statistical ICAS. The AUROC increases by 0.4 per cent (from 86.4 to 86.8 per cent[33]) with respect to the statistical ICAS alone.

[33] The AUROC of the statistical ICAS was 86.2 per cent in previous section and 86.4 per cent in this one. This difference is due to the fact that the datasets slightly differ in the two sections (the latter is a little bit smaller than the former).

6.3 Phi

In this section we show the results of the Phi model described in section 4.3. We consider the same logistic models of previous sections, except that Phi variables are used here:

$$g(P[def_c = 1]) = \alpha_0 + \sum_j \left(\beta_j \cdot A^{phi,j,c} + \gamma_j \cdot M^{phi,j,c}\right) \quad (5)$$

$$g(P[def_c = 1]) = \alpha_0 + \alpha_1 \cdot S_c + \sum_j \left(\beta_j \cdot A^{phi,j,c} + \gamma_j \cdot M^{phi,j,c}\right) \quad (6)$$

As LLaMA, Phi fails to classify 17 per cent of the articles. As a result, the dataset used for the regressions shrinks by 4 per cent to 58,096 records, of which 941 are defaults.

Table 7 summarizes the results, reporting estimates, standard errors and p-values for the selected variables and the AUROC.

Table 7. Phi models

	Def	
	(5) (News)	(6) (News + ICAS)
α_0	-4.4334*** (0.0569)	0.4031** (0.1085)
α_1		1.2058*** (0.0280)
$\beta_{\text{very bad}}$	0.6200*** (0.1501)	0.4665* (0.1641)
β_{bad}	0.4798** (0.1343)	
$\beta_{\text{very good}}$	-1.0706*** (0.2753)	-0.8660** (0.2712)
γ_{bad}	0.4082*** (0.0971)	0.5565*** (0.0723)
AUROC	0.616	0.868
Note: * p-value < 0.01; ** p-value < 0.001; *** p-value < 0.0001		

Four variables are selected for the first model: one for the 'very good' articles, one for the 'very bad' articles, two for the 'bad' articles. The fact that the coefficient of the 'very bad' articles is bigger (in absolute value) than that of the 'bad' articles is as expected. The AUROC is 61.6 per cent, marginally better than LLaMA.

When considering Phi variables together with the statistical ICAS, three variables remain: one for the 'very good' articles, one for the 'very bad' articles and one for the 'bad' articles. The AUROC increases by 0.4 per cent (from 86.4 to 86.8 per cent) with respect to the statistical ICAS.

6.4 Gemma

For Gemma (see section 4.4) we consider the same logistic models of the previous sections, except that Gemma variables are used here:

$$g(P[def_c = 1]) = \alpha_0 + \sum_j \left(\beta_j \cdot A^{gemma,j,c} + \gamma_j \cdot M^{gemma,j,c}\right) \quad (7)$$

$$g(P[def_c = 1]) = \alpha_0 + \alpha_1 \cdot S_c + \sum_j \left(\beta_j \cdot A^{gemma,j,c} + \gamma_j \cdot M^{gemma,j,c}\right) \quad (8)$$

Gemma fails to classify 19 per cent of the articles. As a result, the dataset used for the regressions shrinks by 5 per cent to 57,622 records (a few hundred records less than LLaMA and Phi), of which 922 are defaults.

Table 8 summarizes the results, reporting estimates, standard errors and p-values for the selected variables and the AUROC.

Table 8. Gemma models

	def	
	(7) (News)	(8) (News + ICAS)
α_0	-3.9169*** (0.0596)	0.6043*** (0.1020)
α_1		1.2081*** (0.0282)
β_{good}	-0.3663*** (0.0922)	
$\beta_{very\ good}$	-3.0401*** (0.3727)	-2.2111*** (0.3736)
$\gamma_{very\ bad}$		0.2013* (0.0800)
γ_{bad}	0.2750*** (0.0722)	0.3191*** (0.0793)
AUROC	0.613	0.871
Note: * p-value < 0.01; ** p-value < 0.001; *** p-value < 0.0001		

Three variables are selected for the first model: one for the 'very good' articles, one for the 'good' articles and one for the 'bad' articles. The fact that the coefficient of the 'very good' articles is bigger (in absolute

value) than that of the 'good' articles is as expected. The fact that we have a 'bad' variable without a 'very bad' variable is suspicious (in principle, 'very bad' articles should be more informative in predicting default than simply 'bad' articles). The AUROC is 61.3 per cent, slightly lower than Phi's.

When considering Gemma variables together with the statistical ICAS, three variables are selected: one for the 'very good' articles, one for the 'bad' articles and one for the 'very bad' articles. The AUROC increases by 0.5 per cent (from 86.6 to 87.1 per cent[34]) with respect to the statistical ICAS.

6.5 LLMs ensemble

In this section we use all LLMs described in section 4.

We consider the following logistic models:

$$g(P[def_c = 1]) = \alpha_0 + \sum_{i,j}\left(\beta_{i,j} \cdot A^{i,j,c} + \gamma_{i,j} \cdot M^{i,j,c}\right) \quad (9)$$

$$g(P[def_c = 1]) = \alpha_0 + \alpha_1 \cdot S_c + \sum_{i,j}\left(\beta_{i,j} \cdot A^{i,j,c} + \gamma_{i,j} \cdot M^{i,j,c}\right) \quad (10)$$

Table 9 summarizes the results, reporting estimates, standard errors and p-values for the selected variables and the AUROC.

Table 9. LLMs ensemble

	def	
	(9) (News)	(10) (News + ICAS)
α_0	-3.6050*** (0.1028)	1.2718*** (0.1511)
α_1		1.1940*** (0.0279)
$\gamma_{bert,neutral}$	-3.4365*** (0.4475)	-3.8454*** (0.5267)
$\beta_{phi,very\ good}$	-0.5778 (0.2578)	-0.5168 (0.2611)
$\gamma_{phi,very\ bad}$	0.3282*** (0.0790)	0.3232*** (0.0851)
$\gamma_{phi,bad}$	0.6938*** (0.0722)	0.5939*** (0.0769)

[34] The AUROC of the statistical ICAS was 86.2 per cent in section 6.1, 86.4 per cent in sections 6.2-6.3 and 86.6 per cent in this section. These differences are due to the fact that the datasets slightly differ in the four sections.

$\gamma_{phi,neutral}$		0.1934 (0.0865)
$\beta_{gemma,neutral}$	0.2336* (0.0929)	
$\beta_{gemma,very\ good}$	-2.5259*** (0.3737)	-1.9767*** (0.3622)
AUROC	0.651	0.871
Note: * p-value < 0.01; ** p-value < 0.001; *** p-value < 0.0001		

Six variables are selected for the first model: two for the 'very good' articles, two for the 'neutral' articles, one for the 'bad' articles and one for the 'very bad' articles. Based on their significance, we have one BERT variable, three Phi variables and two Gemma variables. We note that no LLaMA variables are selected, while one BERT variable is, despite the fact that LLaMA is stronger in terms of discriminatory power than BERT on a standalone basis. This means that, when combining all LLMs together, BERT still has something to say (i.e., it is less correlated to the other models), while LLaMA seems to be redundant. The AUROC is 65.1 per cent, a fairly decent value, and better than those of the previous sections.

When considering all variables together with the statistical ICAS, six variables are selected again, mostly those of the previous model, except that a 'neutral' variable from the Gemma model is replaced by a 'neutral' variable from the Phi model. Overall, we have four Phi variables out of six. The AUROC increases by 0.9 per cent (from 86.2 to 87.1 per cent) with respect to the statistical ICAS.

Each model analysed in previous sections demonstrates unique strengths in extracting and interpreting sentiment, but their combined use yields the most robust results, reflecting the synergetic potential of the LLM approach.

We recall that our LLMs are not trained on default data. This means that, for example, LLMs may classify as 'bad' or 'very bad' some articles whose content is overall negative, but without relevance to the possible default of the company. Despite this limitation, LLMs are able to increase the AUROC of the ICAS by almost one percentage point, as shown in this section. We expect that LLMs trained on ICAS default data can perform even better; this is left for future research.

The results should be interpreted in light of the objective of the exercise. The purpose of the analysis is not to build a stand-alone classifier based on news, but to assess whether LLM-based sentiment indicators contain additional information that can complement the statistical ICAS. For this reason, the main performance metric is the AUROC, which evaluates discriminatory power of the model.

7. Conclusions

This study shows the ability of large language models (LLMs) to enhance the accuracy of credit risk assessment systems through the integration of sentiment indicators derived from company-related news. By systematically processing vast amounts of unstructured textual data, the models employed are able to capture nuanced sentiment cues that traditional methods often overlook. The integration of these indicators into the Banca d'Italia's ICAS framework complements quantitative metrics and addresses the complexities inherent in qualitative assessments.

The results indicate that LLM-driven sentiment analysis improves the discriminatory power of the Banca d'Italia's statistical ICAS, by one percentage point in terms of area under the ROC curve. This result is significant, considering that improvements in AuROC become more and more difficult as the starting level increases, and that the current model, with an AuROC of around 85 percent, is already above the levels generally reported in the literature or by other credit assessment systems.

The application of quantised versions ensures computational efficiency while maintaining high levels of accuracy, complying with operational constraints and data protection standards. Each model shows individual strengths in extracting and interpreting sentiment. Their combined use yields the most significant results, reflecting the synergetic potential of this approach.

Future developments could involve the integration of even larger LLMs and the application of fine-tuning techniques, such as training on default data. Further developments could also involve sectorial analysis, rather than company-based analysis. In particular, the extraction of information on economic sectors from the news could overcome the main limitation of the current approach, namely the fact that we did not find articles for a large share of companies.

More generally, as technology advances, the role of LLMs in credit risk assessment is likely to become even more transformative, paving the way for finer and innovative financial decision-making.

References

Altman, E.I. (1983), A complete guide to predicting, avoiding, and dealing with bankruptcy. *Corporate Financial Distress, New York*.

Altman, E., Barboza, F. and Kimura, H. (2017), Machine learning models and bankruptcy prediction, *Expert Systems with Applications*, *83*, 405-417.

Aprigliano, V., Emiliozzi, S., Guaitoli, G., Luciani, A., Marcucci, J., and Monteforte, L. (2021), The power of text-based indicators in forecasting the Italian economic activity, *Banca d'Italia Working Papers,* No. 1321.

Auria, L., Bingmer, M., Charavel, C., Gavilá, S., Graciano, C., Iannamorelli, A., Levy, A., Maldonado, A., Mateo, C., Resch, F., Rossi, A.M. and Sauer, S. (2021), Overview of Central Banks' In-House Credit Assessment Systems in the Euro Area, *European Central Bank Occasional Paper Series*, No. 284.

Bozzon, A., Putra, S.G.P., Joshi, B. and Redi, J. (2020), A Credit Scoring Model for SMEs Based on Social Media Data, in *International Conference on Web Engineering* (pp. 113-129), Springer, Cham.

Cathcart, L., Gotthelf, N.M., Uhl, M., and Shi, Y. (2020), News sentiment and sovereign credit risk, *European Financial Management*, *26*(2), 261-287.

Chen, Y., Liu, Z., & Wang, H. (2023). Financial Sentiment Analysis Using BERT-Based Models. *Journal of Financial Data Science*, 5(2), 45–58.

Ciampi, F. and Gordini, N. (2013), Small Enterprise Default Prediction Modeling through Artificial Neural Networks: An Empirical Analysis of Italian Small Enterprises, *Journal of Small Business Management*, *51*(1), 23-45.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1, 4171–4186.

Devine, P. (2024). Tagengo: A Multilingual Chat Dataset. arXiv preprint arXiv:2405.12612.

Dolphin, R., Dursun, J., Chow, J., Blankenship, J., & Adams, K. (2024). Extracting Structured Insights from Financial News: An Augmented LLM Driven Approach. *arXiv preprint arXiv:2407.15788*.

Gariano, G. and Viggiano, G. (2022), Press news and social media in credit risk assessment: the experience of Banca d'Italia's In-house Credit Assessment System, *Markets, Infrastructures, Payment Systems' series*.

Gemma Team, Google DeepMind (2024). Gemma 2: Improving Open Language Models at a Practical Size. *arXiv preprint arXiv:2408.00118*.

Giovannelli, F., Iannamorelli, A. Levy, A. and Orlandi, M. (2020), The In-house credit assessment system of Banca d'Italia, *Banca d'Italia Occasional Paper*, No. 586.

González-Fernández, M. and González-Velasco, C. (2020), An alternative approach to predicting bank credit risk in Europe with Google data, *Finance Research Letters*, *35*, 101281.

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., Liu, T. (2023). A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *arXiv preprint arXiv:2311.05232.*

Lau, R.Y., Li, C., Yuan, H. and Wong, M.C. (2018), Mining Emotions of the Public from Social Media for Enhancing Corporate Credit Rating, in *2018 IEEE 15th International Conference on e-Business Engineering (ICEBE)* (pp. 25-30), IEEE.

Lei, Y., Wang, Z., Liu, C., Wang, T., & Lee, D. (2024). FinLangNet: A Novel Deep Learning Framework for Credit Risk Prediction Using Linguistic Analogy in Financial Data. *Journal of Computational Finance*, 18(3), 123–140.

Lis, S., Kubkowski, M., Borkowska, O., Serwa, D., & Kurpanik, J. (2023). Analyzing Credit Risk Model Problems through NLP-Based Clustering and Machine Learning: Insights from Validation Reports. *Proceedings of the International Conference on Financial Engineering and Risk Management*, 112–120.

Lu, H.M., Hung, M.W. and Tsai, F.T. (2016), The impact of news articles and corporate disclosure on credit risk valuation, *Journal of Banking and Finance*, *68*, 100-116.

Microsoft (2024). Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone, *arXiv preprint arXiv*: 2404.14219.

Mo, K., Liu, W., Xu, X., Yu, C., Zou, Y., & Xia, F. (2024). Fine-Tuning Gemma 2-7B for Enhanced Sentiment Analysis of Financial News Headlines. *arXiv preprint arXiv:2406.13626*.

Sanz-Guerrero, M., & Arroyo, J. (2024). Credit Risk Meets Large Language Models: Building a Risk Indicator from Loan Descriptions in P2P Lending. *arXiv preprint arXiv:2401.16458*.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., Lample, G. (2024). LLaMA 3: Open and Efficient Foundation Language Models. *arXiv preprint arXiv:2405.00360*.

Tran-The, T. (2020). Modeling Institutional Credit Risk with Financial News. *arXiv preprint arXiv*:2004.08204.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is All You Need. arXiv preprint arXiv:1706.03762.

Yang, X., & Yu, H. (2020). BERT-based Sentiment Analysis for Financial News. *Procedia Computer Science*, 174, 122–127.

Zhang, L., & Zhao, X. (2024). Leveraging LLaMA 3 for Sentiment Analysis in Financial News. International *Journal of Computational Finance*, 12(1), 78–92.

Appendix 1 – LLMs prompt and model settings

In this appendix, we describe the model settings employed for the large language models (LLMs) used in this study: LLaMA 3, Phi 3 and Gemma 2. These configurations are carefully selected and fine-tuned to balance computational efficiency and predictive accuracy, ensuring reliable sentiment analysis. Among the key elements, the temperature controls the randomness of the model output, with lower values such as 0.2 producing more deterministic predictions suitable for this context (see Table 13 below).

To ensure consistent and reliable sentiment analysis, the prompt construction is carefully standardized across all LLMs. It is explicitly designed to instruct the models to perform a sentiment classification task, minimizing ambiguity and encouraging deterministic output. The instruction requires the model to classify the sentiment of a given text into one of five categories: 'molto negativo,' 'negativo,' 'neutrale,' 'positivo,' or 'molto positivo.' An example of the prompt is as follows:

You are a sentiment analysis model in the financial domain. Analyze the following TEXT and respond with one of the following categories: 'very negative', 'negative', 'neutral', 'positive', or 'very positive'. If possible, reply with the category only.
TEXT: *Beta group, an Italian company operating in the manufacturing sector, is facing one of the worst crises in its history. Following a series of managerial missteps, its financial statements show losses for the third consecutive year, with revenues down 35% compared to the previous year. Employees, already affected by delays in wage payments, have staged protests in front of the company's headquarters, denouncing unsustainable working conditions and a lack of transparency from management. Meanwhile, investor confidence has plummeted to an all-time low, with the company's stock value falling by 60% over the past six months. Despite promises of a restructuring plan, creditors are considering legal action to recover outstanding debts. Analysts warn that the company may be heading toward bankruptcy unless immediate measures are taken.*

The carefully structured nature of this prompt ensures that the models delivers concise and relevant output, in line with the study's objective of sentiment classification. The temperature plays a crucial role in regulating the randomness and variability of the output generated by LLMs. When the temperature is set at a low value (e.g., 0.2), the models prioritize the most likely and deterministic response and gives a direct sentiment classification. Conversely, higher temperature values (e.g., 0.8 or 1.0) introduce variability in token sampling, often leading to longer output or less constrained responses.

Table A1 illustrates simulated output from the three models when tasked with analysing the example text at different temperature settings:

Table A1. Impact of temperature on LLMs output for sentiment analysis

Temperature	LLaMA 3	Phi 3	Gemma 2
0.2	'negativo'	'molto negativo'	'negativo'
0.8	The company's outlook seems decidedly negative given its ongoing crisis	The financial situation indicates significant distress and a negative outlook	Conditions suggest the company is experiencing severe financial challenges

This example shows how the temperature setting influences the models' response, with lower values ensuring concise and reliable sentiment labels, and higher values allowing for more descriptive output. For the purposes of this study, a low temperature was adopted to prioritize deterministic and consistent predictions, in line with the requirement for precise sentiment analysis in credit risk assessment.

The maximum number of tokens defines the length of the output, tailored to capture sentiment concisely while accommodating necessary detail. The batch size for prediction, which determines the number of input samples processed simultaneously during a single inference pass, is uniformly set at 1024 across all models. This choice strikes a balance between computational efficiency and throughput, allowing multiple inputs to be processed in parallel. A minimum token limit of 5 was specified to ensure the brevity of output, while a maximum token threshold of 2048 allows the models to handle complex textual input. The top-p value, used for nucleus sampling, determines the diversity of the output by considering the cumulative probability of top token options. Both the presence penalty and frequency penalty were set at 0.0 to avoid penalizing token repetition excessively, ensuring the generation of coherent and contextually relevant text. Additional elements, such as the stopping criteria, are kept at their default values to maintain neutrality and to comply with the OpenAI standard for configuring large language models.

The optimal configurations for LLaMA, Phi, and Gemma are determined through an iterative process of experimentation and fine-tuning. Starting from default settings, adjustments are made based on the models' performance on a validation subset of the dataset. Temperature, maximum token limit, and batch size were systematically optimized to balance computational efficiency, inference speed, and the accuracy of sentiment predictions. This approach ensures that each model's configuration is in line with the study objectives and the characteristics of the input data. Table A2 shows the results.

Table A2. LLMs settings

	LLaMA 3	Phi 3	Gemma 2
Temperature	0.30	0.20	0.25
Max tokens	14	7	7
Predicted batch size	1024	1024	1024
N minimal tokens	5	5	5
N maximal tokens	2048	2048	2048
Top-p	0.80	0.85	0.90
Presence penalty	0	0	0
Frequency penalty	0	0	0
Stop	None	None	None
Logit bias	0	0	0
N	1	1	1
Best of	1	1	1

The optimisation process follows a pragmatic and iterative approach, based on repeated evaluations over a validation subset of the dataset. Starting from default configurations, key parameters such as temperature, maximum number of tokens and top-p are progressively adjusted within a restricted range of plausible values. The objective is not to maximise performance on a specific metric, but to ensure stable and consistent sentiment classification across different texts and models.

In particular, the temperature parameter is calibrated to balance determinism and variability in the output. Lower values (between 0.2 and 0.3) are preferred, as they reduce randomness and favour consistent classification, which is crucial in a credit risk context. Similarly, the maximum token setting is chosen to ensure concise outputs aligned with the classification task, while still allowing sufficient flexibility to handle longer or more complex inputs.

Sensitivity checks indicate that moderate variations in these parameters do not materially affect the aggregate distribution of sentiment classifications, nor the main results presented in Section 6. Changes in temperature within a reasonable range may lead to marginal differences in individual classifications, particularly for borderline cases, but these effects tend to average out when aggregating information across articles and over time. Overall, the results appear robust to small perturbations of the configuration parameters, suggesting that the findings are not driven by a specific tuning choice.

Appendix 2 – Chi-squared test for sentiment consistency

A challenge for LLMs is hallucination, where output appears plausible but is factually incorrect or inconsistent with the input data. In credit risk analysis, this can undermine the reliability of sentiment indicators extracted from textual data. To address this issue, we compare the sentiment predictions generated from two representations of the articles: the full text (*corpus*) and the corresponding snippets (concise summaries) provided by the Dow Jones Factiva database. This comparison ensures consistency in sentiment classification across different text granularities.

The Chi-squared test used for this analysis is particularly suitable for comparing distributions of discrete variables, such as the sentiment levels considered in this study: very bad, bad, neutral, good, and very good. A p-value greater than 0.05 indicates no statistically significant difference between the two distributions, thereby supporting the consistency of predictions.

Table A3 presents the results for LLaMA. The Chi-squared statistics and corresponding p-values for all sentiment categories confirm that there are no statistically significant differences between the predictions based on the *corpus* and snippet analyses.

Table A3. Chi-squared test results for LLaMA sentiment distribution

	Chi-squared statistic	p-value
Very bad	1.87e-01	0.510
Bad	1.65e-01	0.883
Neutral	6.21e-01	0.445
Positive	3.58e-01	0.478
Very positive	1.63e-01	0.863

The results for Phi are shown in Table A4. Across all sentiment categories, the p-values are well above 0.05, indicating high consistency in predictions between the full texts and snippets.

Table A4. Chi-squared test results for Phi sentiment distribution

	Chi-squared statistic	p-value
Very bad	3.86e-01	0.290
Bad	9.63e-01	0.626
Neutral	2.73e-01	0.188
Positive	6.67e-01	0.104
Very positive	1.05e-01	0.478

For Gemma, as shown in Table A5, the p-values across all sentiment levels also exceeded the 0.05 threshold, further confirming that there are no statistically significant differences between the sentiment distributions of the *corpus* and snippets.

Table A5. Chi-squared test results for Gemma sentiment distribution

	Chi-squared statistic	p-value
Very bad	2.33e-01	0.788
Bad	1.99e-01	0.795
Neutral	3.58e-01	0.202
Positive	3.35e-01	0.292
Very positive	1.00e-01	0.936

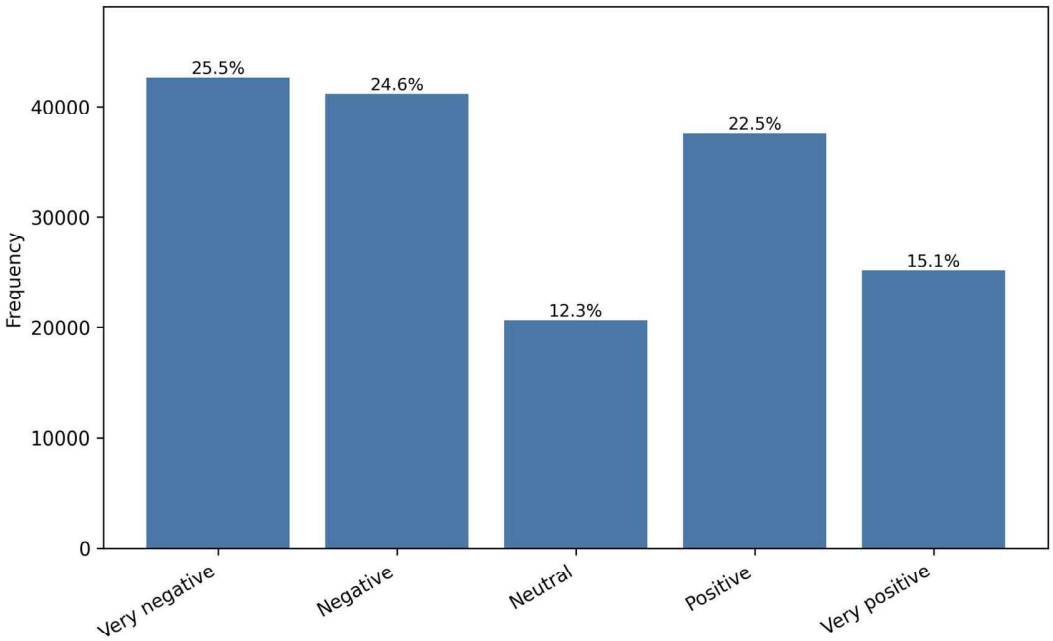
The results of the Chi-squared tests show that the sentiment predictions generated by the three LLMs - LLaMA, Phi, and Gemma - remain consistent across full texts and snippets. This analysis suggests that the output of the models is robust and reliable, validating their suitability for large-scale sentiment analysis tasks.

Appendix 3 – Distribution of sentiment classifications

In this appendix, we provide a detailed analysis of the frequency distributions of sentiment classifications produced by the four large language models (LLMs) employed in this study: BERT, LLaMA 3, Phi 3 and Gemma 2. The purpose of this analysis is to offer a descriptive assessment of the sentiment composition derived from financial news articles and to verify whether the classification exercise generates any imbalances that could affect the interpretation of the results presented in Sections 5 and 6.

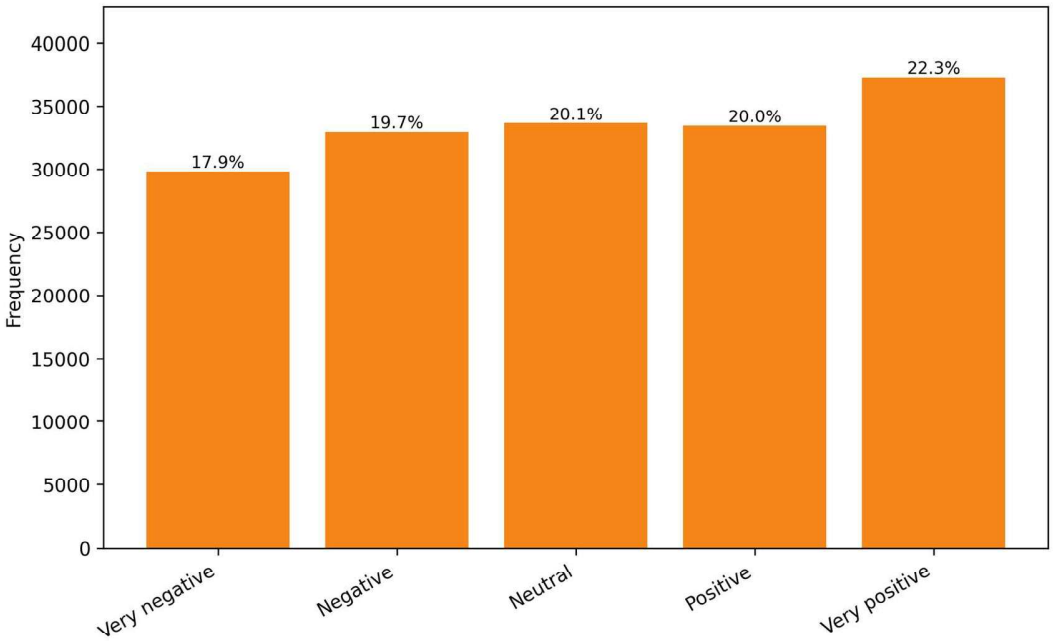
Figures 2 through 5 report the empirical distributions across the five sentiment categories for each model. A first relevant observation is that all categories are consistently populated in every case, with no evidence of empty or near-empty classes. This feature is important in light of the aggregation procedure described in Section 5, where sentiment indicators are constructed using averages and maxima across articles. The presence of observations in all categories ensures that both intermediate and extreme sentiment signals are retained in the dataset, thereby supporting the robustness of the subsequent statistical analysis.

Figure 2 – BERT frequency distributions of sentiment classification categories



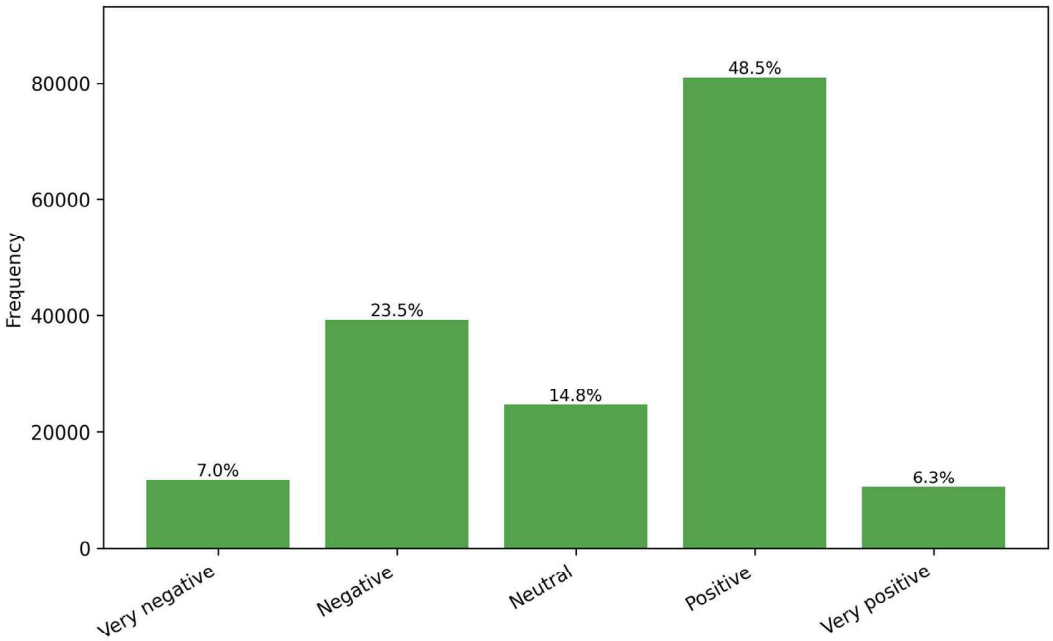
The distribution obtained with BERT, shown in Figure 2, displays a moderate prevalence of negative sentiment. The categories very negative and negative jointly account for approximately half of the observations, while positive and very positive represent a slightly smaller share. The neutral category remains clearly present, although less frequent. This configuration is broadly consistent with the nature of financial news, where adverse events tend to receive substantial coverage. At the same time, the presence of a significant proportion of positive observations indicates that the dataset is not dominated by negative reporting. It is also worth recalling that BERT produces probabilistic outputs, which are aggregated across categories. This characteristic contributes to a relatively smooth allocation of observations and reduces the likelihood of excessive concentration in specific classes.

Figure 3 – LLaMA frequency distributions of sentiment classification categories



The distribution for LLaMA, presented in Figure 3, is notably balanced across all five categories. Each class accounts for a similar proportion of the total, with shares ranging within a relatively narrow interval. This pattern suggests a classification behaviour that distributes observations more evenly across sentiment levels. From an analytical perspective, such a configuration reduces the risk of systematic bias toward a specific direction of sentiment and ensures that all categories contribute comparably to the construction of sentiment indicators.

Figure 4 – Phi frequency distributions of sentiment classification categories



The distribution produced by Phi, shown in Figure 4, exhibits a higher concentration in the positive category, which represents the largest share of observations. The remaining categories are nevertheless well represented, including both negative and very negative classes. While this pattern differs from those observed for BERT and LLaMA, it does not imply a structural distortion. The model retains the ability to identify adverse sentiment, as evidenced by the non-negligible share of negative classifications, and the overall distribution remains sufficiently diversified to support the empirical analysis.

Figure 5 – Gemma frequency distributions of sentiment classification categories

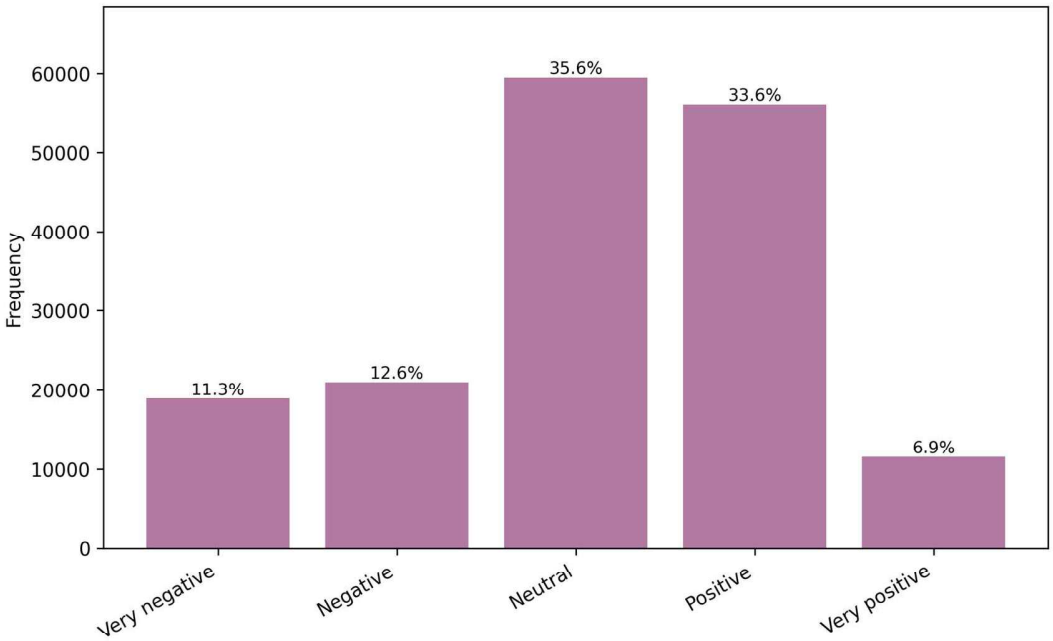


Figure 5 reports the distribution obtained with Gemma. In this case, the largest shares are associated with the neutral and positive categories, while the remaining classes are distributed across the spectrum with lower but still meaningful frequencies. This pattern suggests a relatively conservative classification approach, with a tendency to assign observations to intermediate sentiment levels. However, the presence of both negative and very positive observations confirms that the model continues to capture signals across the entire sentiment range.

Taken together, the four distributions do not indicate a systematic tendency toward negative sentiment in the dataset. While individual models exhibit different allocation patterns, these differences appear to be model-specific rather than driven by the underlying data. In particular, the presence of models with more balanced or even slightly positive-oriented distributions offsets the moderate prevalence of negative categories observed in others. This heterogeneity is consistent with the use of multiple LLMs with different architectures and training characteristics, as described in Section 4.

From a statistical perspective, the distributions do not raise concerns regarding the adequacy of sample sizes within individual categories. All sentiment classes are sufficiently populated, including the extreme ones, which are particularly relevant for credit risk analysis. This supports the reliability of the regression results reported in Section 6, where sentiment indicators derived from these classifications are used as explanatory variables.

It is important to note that the main contribution of the paper does not rely on differences in the marginal distributions across models. The empirical analysis focuses on the consistency of sentiment classification across different textual representations, as discussed in Appendix 2, and on the contribution of sentiment indicators to the discriminatory power of the ICAS. The distributional evidence presented here should therefore be interpreted as a complementary descriptive assessment, confirming that the classification exercise produces a sufficiently rich and balanced set of observations for the purposes of the study.

Recently published papers in the 'Markets, Infrastructures, Payment Systems' series

n. 51 Environmental data and scores: lost in translation, *by Enrico Bernardini, Marco Fanari, Enrico Foscolo and Francesco Ruggiero*

n. 52 How important are ESG factors for banks' cost of debt? An empirical investigation, *by Stefano Nobili, Mattia Persico and Rosario Romeo*

n. 53 The Bank of Italy's statistical model for the credit assessment of non-financial firms, *by Simone Narizzano, Marco Orlandi and Antonio Scalia*

n. 54 The revision of PSD2 and the interplay with MiCAR in the rules governing payment services: evolution or revolution?, *by Mattia Suardi*

n. 55 Rating the Raters. A Central Bank Perspective, *by Francesco Columba, Federica Orsini and Stefano Tranquillo*

n. 56 A general framework to assess the smooth implementation of monetary policy: an application to the introduction of the digital euro, *by Annalisa De Nicola and Michelina Lo Russo*

n. 57 The German and Italian Government Bond Markets: The Role of Banks versus Non-Banks. A joint study by Banca d'Italia and Bundesbank, *by Puriya Abbassi, Michele Leonardo Bianchi, Daniela Della Gatta, Raffaele Gallo, Hanna Gohlke, Daniel Krause, Arianna Miglietta, Luca Moller, Jens Orben, Onofrio Panzarino, Dario Ruzzi, Willy Scherrieble and Michael Schmidt*

n. 58 Chat Bankman-Fried? An Exploration of LLM Alignment in Finance, *by Claudia Biancotti, Carolina Camassa, Andrea Coletta, Oliver Giudice and Aldo Glielmo*

n. 59 Modelling transition risk-adjusted probability of default, *by Manuel Cugliari, Alessandra Iannamorelli and Federica Vassalli*

n. 60 The use of Banca d'Italia's credit assessment system for Italian non-financial firms within the Eurosystem's collateral framework, *by Stefano Di Virgilio, Alessandra Iannamorelli, Francesco Monterisi and Simone Narizzano*

n. 61 Fintech Classification Methodology, *by Alessandro Lentini, Daniela Elena Munteanu and Fabrizio Zennaro*

n. 62 The Rise of Climate Risks: Evidence from Expected Default Frequencies for Firms, *by Matilde Faralli and Francesco Ruggiero*

n. 63 Exploratory survey of the Italian market for cybersecurity testing services, *by Anna Barcheri, Luca Bastianelli, Tommaso Curcio, Luca De Angelis, Paolo De Joannon, Gianluca Ralli and Diego Ruggeri*

n. 64 A practical implementation of a quantum-safe PKI in a payment systems environment, *by Luca Buccella and Stefano Massi*

n. 65 Stewardship Policies. A Survey of the Main Issues, *by Marco Fanari, Enrico Bernardini, Elisabetta Cecchet, Francesco Columba, Johnny Di Giampaolo, Gabriele Fraboni, Donatella La Licata, Simone Letta, Gianluca Mango and Roberta Occhilupo*

n. 66 Is there an equity greenium in the euro area?, *by Marco Fanari, Marianna Caccavaio, Davide Di Zio, Simone Letta and Ciriaco Milano*

n. 67 Open Banking in Italy: A Comprehensive Report, *by Carlo Cafarotti and Ravenio Parrini*

n. 68 Report on the payment attitudes of consumers in Italy: results from ECB SPACE 2024 survey, *by Gabriele Coletti, Marialucia Longo, Laura Painelli, Emanuele Pimpini and Giorgia Rocco*

n. 69 A solution for cross-border and cross-currency interoperability of instant payment systems, *by Domenico Di Giulio, Vitangelo Lasorella, Pietro Tiberi*

n. 70 Do firms care about climate change risks? Survey evidence from Italy, *by Francesca Colletti, Francesco Columba, Manuel Cugliari, Alessandra Iannamorelli, Paolo Parlamento and Laura Tozzi*

n. 71 Demand and supply of Italian government bonds during the exit from expansionary monetary policy, *by Fabio Capasso, Francesco Musto, Michele Pagano, Onofrio Panzarino, Alfonso Puorro and Vittorio Siracusa*

n. 72 Statistics on tokenized financial instruments: A challenge for central banks, *by Riccardo Colantonio, Massimo Coletta, Riccardo Renzi*

n. 73 Credit Risk Assessment with Stacked Machine Learning, *by Francesco Columba, Manuel Cugliari, Stefano Di Virgilio*

n. 74 What if Ether Goes to Zero? How Market Risk Becomes Infrastructure Risk in Crypto, *by Claudia Biancotti*

n. 75 The Cyber Risk of Non-Financial Firms, *by Francesco Columba, Manuel Cugliari, Marco Orlandi, Federica Vassalli*

n. 76 Sustainability and financial innovation: The emerging role of Fintech for Good (F4G), *by Alessandro Lentini and Daniela Elena Munteanu*

n. 77 Hydrogeological and credit risk: the Italian firms' physical risk-adjusted probability of default, *by Manuel Cugliari, Simone Narizzano and Federica Vassalli*

n. 78 Liquidity Optimization in Gross Settlement Systems with Quantum Reordering: Application to TARGET2, *by Valerio Astuti, Adriano Baldeschi, Luca Bastianelli, Giuseppe Bruno, Ajit Desai, Danica Marsden and Riccardo Russo*

n. 79 The expert assessment within Banca d'Italia's in-house credit assessment system, *by Lorenzo Esposito, Massimo Guglielmi, Francesco Monterisi, Simone Narizzano and Marco Orlandi*

n. 80 Digital payments and economic performance: evidence from Italy, *by Guerino Ardizzi, Niccolò Lippi Boncambi, Cristina Demma and Alberto Leorati*

n. 81 The euro-area repo market: structure, participants and interlinkages, *by Lorenzo Caverni, Annalisa De Nicola and Mattia Persico*

n. 82 Siamese Networks for AI-powered Automated Banknote Quality Control, *by Salvatore Gentile, Andrea Luciani, Sabina Marchetti, Domenico Pansini and Marco Viticoli*

n. 83 Private Equity and Innovation in the Euro Area, *by Emanuele Degani, Simone Letta and Tommaso Perez*

n. 84 Social and Governance Factors in Default Risk Assessment, *by Manuel Cugliari, Stefano Di Virgilio, Enrico Foscolo and Alessandra Iannamorelli*

n. 85 Household Allocation of the Next Generation of Digital Money, *by Marcello Pericoli*

n. 86 Are Stablecoins Efficient for Remittances? Evidence from a Mystery Shopping Exercise by Banca d'Italia, *by Alberto Di Iorio, Enrica Di Stefano, Michele Mascioli and Giorgio Trebeschi*

n. 87 Who uses 'Buy Now, Pay Later' in Italy? Evidence from Household Survey Data, *by Daniele Figoli and Serena Palazzo*

n. 88 Fee-cap Models for Payment Instruments: a Policy Perspective, *by Nicola Branzoli, Stefano Pietrosanti and Alessia Vita*

n. 89 Send me money ASAP: Fast Payment Systems and the Cost of Remittances, *by Marco Brandi, Alberto Di Iorio and Andrea Nobili*

n. 90 Hedging at speed in sovereign bond markets, *by Onofrio Panzarino and Antonio Perrella*