Acta Informatica Pragensia
2026, Volume 15, Issue 3, 740–754
Special Issue: Selected Extended Papers from International Conferences CITA, ICRAMI, and ADACIS
https://doi.org/10.18267/j.aip.328

PRAGUE UNIVERSITY OF ECONOMICS AND BUSINESS

Miscellanea Open Access

Operationalising Large Language Models for Financial Sentiment Analysis: The PipeFinePT Modular Pipeline

Gonçalo Carnaz [1,2], **João M. Carvalho** [1,2], **Rui Pedro Marques** [1,3]

[1] Higher Institute of Accounting and Administration, University of Aveiro, Aveiro, Portugal
[2] Institute of Electronics and Telematics Engineering of Aveiro, University of Aveiro, Aveiro, Portugal
[3] Research Unit on Governance, Competitiveness and Public Policies, University of Aveiro, Aveiro, Portugal

Corresponding author: Rui Pedro Marques (ruimarques@ua.pt)

Editorial Record

First submission received:
March 9, 2026

Revisions received:
June 5, 2026
July 29, 2026

Accepted for publication:
August 10, 2026

Special Issue Editor:
Dalel Kanzari
University of Sousse, Tunisia

Academic Editor:
Zdenek Smutny
Prague University of Economics and Business, Czech Republic

This article was accepted for publication by the Academic Editor upon evaluation of the reviewers' comments.

How to cite this article:
Carnaz, G., Carvalho, J. M., & Marques, R. P. (2026). Operationalising Large Language Models for Financial Sentiment Analysis: The PipeFinePT Modular Pipeline. *Acta Informatica Pragensia*, 15(3), 740–754. https://doi.org/10.18267/j.aip.328

Abstract

Background: Financial decision-making increasingly relies on automated analysis of large volumes of textual information. However, financial sentiment analysis resources remain scarce for low-resource linguistic contexts such as European Portuguese, limiting reproducible evaluation and adoption of advanced AI techniques within financial information systems.

Objective: This study aims to design and evaluate a modular pipeline capable of operationalizing large language models (LLMs) for sentiment classification of Portuguese financial news while assessing performance, robustness, and deployment trade-offs across different model categories.

Methods: A domain-specific dataset of 200 European Portuguese financial news articles was constructed and manually annotated by three independent annotators using a positive, neutral, and negative sentiment scheme. The proposed PipeFinePT pipeline integrates data acquisition, preprocessing, prompt-based annotation, and standardized evaluation. Multiple proprietary and open-weights LLMs were evaluated using identical prompts and experimental conditions, with performance measured through accuracy, precision, recall, and F1-score metrics, including robustness testing via reversed input ordering.

Results: Results show that LLMs can perform financial sentiment classification in European Portuguese with moderate reliability, achieving up to 0.781 accuracy, although pairwise differences among the three models with valid outputs were not statistically significant. Higher-tier proprietary models such as GPT-4.1 demonstrated stronger semantic interpretation but at substantially higher inference cost, whereas lighter proprietary variants and the open-weight LLaMA-4-Maverick-17B offered improved efficiency with reduced contextual sensitivity. Robustness experiments revealed generally stable model rankings but performance variability under altered input ordering.

Conclusion: The PipeFinePT pipeline provides a reproducible framework for integrating LLM-based sentiment analysis into applied financial information systems. The findings highlight both the feasibility and limitations of deploying LLMs in low-resource financial language environments and support their use as operational analytical components in multilingual financial analytics workflows.

Index Terms

Sentiment analysis; Financial news; LLM; Portuguese; Prompt engineering; Natural language processing.

1 INTRODUCTION

The increasing digitalization of financial markets has transformed economic environments into highly information-intensive ecosystems, where decisions are continuously influenced by large volumes of textual data originating from news platforms, social media, analyst reports, and online publications.

In such environments, the ability to automatically process and interpret unstructured textual information has become a critical component of modern financial information systems. Investors, institutions, and analysts increasingly rely on computational tools capable of transforming qualitative information into structured signals that support decision-making processes.

Sentiment analysis has emerged as a key technique within applied informatics and natural language processing (NLP) for extracting opinions, attitudes, and expectations from textual sources. In financial contexts, sentiment indicators derived from news and public discourse have been shown to influence market expectations, risk perception, and investment behaviour. Consequently, automated sentiment analysis has evolved from a purely linguistic task into an essential analytical capability embedded within financial decision-support systems and data-driven economic platforms.

Recent advances in artificial intelligence, particularly the emergence of large language models (LLMs), have significantly expanded the possibilities of text analytics. Unlike traditional machine learning approaches that depend heavily on domain-specific training datasets, LLMs enable prompt-driven inference, contextual reasoning, and flexible adaptation across tasks without extensive retraining. These characteristics position LLMs as foundational components for next-generation information processing pipelines capable of supporting complex analytical workflows.

Despite these advances, important challenges remain. Most research and available resources in financial sentiment analysis are concentrated on English-language datasets and large financial markets. This creates a structural limitation for smaller linguistic ecosystems, where domain-specific annotated resources are scarce and existing models may struggle with linguistic nuances, idiomatic expressions, and context-dependent financial terminology. As a result, organizations operating in underrepresented language contexts face barriers in adopting advanced AI-driven analytical tools within their information systems.

The Portuguese language represents a particularly relevant case of a low-resource setting in financial NLP. Throughout this paper, the term "low-resource" is used in its NLP sense, denoting the scarcity of annotated datasets, pretrained models, and evaluation benchmarks rather than computational or infrastructure constraints. It encompasses two distinct dimensions: language-level scarcity, since European Portuguese is underrepresented in financial NLP benchmarks and pretraining corpora relative to English; and domain-level scarcity, since financial sentiment analysis resources for European Portuguese are limited even relative to general-purpose Portuguese language resources. While Portuguese is widely spoken globally, dedicated datasets and evaluation frameworks for financial sentiment analysis remain limited, especially for European Portuguese. This gap restricts reproducibility, benchmarking, and systematic evaluation of modern LLMs capabilities within localized financial information environments.

To address this limitation, this paper presents an extended and revised version of a study originally presented at the *IEEE International Conference on Advances in Data-Driven Analytics and Intelligent Systems*, introducing PipeFinePT (Carnaz et al., 2025). PipeFinePT is a modular sentiment annotation pipeline designed to evaluate and operationalize LLMs for financial news analysis in European Portuguese. The proposed approach integrates data extraction, preprocessing, prompt-based annotation, and standardized evaluation into a reproducible architecture aimed at supporting applied financial analytics workflows.

The main contributions of this work are as follows:

- the design of a modular LLM-driven pipeline for financial sentiment annotation tailored to low-resource linguistic contexts;
- the construction of a manually annotated dataset of Portuguese financial news enabling controlled evaluation;
- a comparative assessment of open-weights and proprietary LLMs under identical experimental conditions;
- an empirical analysis of model robustness, including sensitivity to input ordering effects;
- a reproducible framework supporting future research and practical deployment of sentiment analysis within financial information systems.

To guide the empirical evaluation and position the proposed pipeline within applied financial informatics, this study addresses the following research questions:

- **RQ1:** To what extent can LLMs reliably perform sentiment classification on financial news written in European Portuguese, a low-resource linguistic context?
- **RQ2:** How do different categories of LLMs (proprietary versus open-weights models) compare in terms of classification performance, linguistic handling, and operational efficiency?
- **RQ3:** How robust are prompt-based LLMs sentiment classifications when experimental conditions vary, particularly regarding input ordering effects?
- **RQ4:** Can a modular, prompt-driven pipeline provide a reproducible framework for integrating LLM-based sentiment analysis into financial information systems?

These research questions frame the design and evaluation of the proposed PipeFinePT pipeline and guide the experimental analysis presented in the following sections. In this context, the PipeFinePT modular pipeline is positioned as an applied informatics artefact that enables LLMs to operate as analytical components within financial information systems, supporting reproducible experimentation and practical deployment in multilingual market environments.

The remainder of this paper is organized as follows. Section 2 reviews related work on financial sentiment analysis and LLMs. Section 3 presents the PipeFinePT architecture and experimental methodology. Section 4 discusses the evaluation results and analysis. Finally, Section 5 concludes the paper and outlines future research directions.

2 LITERATURE REVIEW

Early sentiment analysis approaches relied on lexicon-based and rule-driven techniques, later replaced by supervised machine learning and deep learning architectures capable of capturing contextual dependencies (Pang & Lee, 2008; Devlin et al., 2019). The emergence of transformer-based LLMs represents a further shift toward contextual and instruction-driven interpretation, enabling the analysis of complex financial discourse (Brown et al., 2020; Zhao et al., 2026).

The increasing availability of digital financial content has motivated extensive research on sentiment analysis as a mechanism for extracting actionable insights from textual data. Within financial information systems, sentiment analysis supports decision-making processes by transforming qualitative narratives into quantitative indicators that can complement traditional market signals. Early approaches relied primarily on classical machine learning techniques applied to structured textual representations, including Naïve Bayes, Support Vector Machines, and neural network architectures.

Carosia et al. (2019) demonstrated that sentiment extracted from Portuguese-language social media content can correlate with stock market movements, highlighting the relevance of textual sentiment as an economic signal. These works established the feasibility of integrating sentiment indicators into financial analytics workflows but remained dependent on manually engineered features and domain-specific datasets. While effective in controlled contexts, traditional approaches often struggle to capture implicit sentiment, and contextual nuances present in financial discourse.

As financial ecosystems became increasingly data-driven, limitations related to scalability, adaptability, and linguistic diversity motivated the exploration of more flexible natural NLP paradigms. To address the complexity of financial language, researchers introduced domain-adapted language models trained on specialized corpora, like FinBERT (Huang et al., 2022) that represents an evolution, incorporating financial-domain knowledge into transformer-based architectures to improve contextual understanding of analyst reports and financial statements. Similarly, large-scale initiatives such as BloombergGPT (Wu et al., 2023) demonstrated that domain-specialized training substantially improves performance across financial NLP benchmarks while preserving general linguistic capabilities.

More recent approaches focus on enhancing performance through targeted fine-tuning strategies. Models such as FinSoSent (Delgadillo et al., 2024) and FinSentGPT (Ardekani et al., 2024) leverage pretrained LLMs and adapt them to financial sentiment prediction tasks, achieving competitive performance across multilingual datasets. These studies highlight the growing importance of domain adaptation as a mechanism for improving semantic interpretation in finance-oriented applications.

Research focusing specifically on Portuguese-language financial sentiment analysis remains comparatively limited but has begun to emerge in recent years. Tavares et al. (2021) proposed a rule-based approach for automatic polarity detection in economic news texts, demonstrating that linguistic heuristics can effectively capture sentiment signals in Portuguese financial reporting. Building on transformer-based architectures, Santos et al. (2023) developed a sentiment analysis model for Portuguese financial news using the BERT neural network framework, highlighting the potential of pretrained language models for domain-specific financial text classification. Complementing these modelling efforts, Meier et al. (2026) introduced a dataset supporting the development of FinBERTimbau, a transformer model adapted to Brazilian Portuguese financial texts, providing an important resource for training and evaluating financial sentiment models in Portuguese-language settings.

However, most domain-specific financial language models continue to rely primarily on English-language corpora, which limits their applicability in multilingual or region-specific financial environments. This limitation also restricts the availability of standardized datasets and evaluation frameworks for systematically assessing sentiment analysis approaches in languages such as European Portuguese.

2.1 Large Language Models as Analytical Infrastructure

The emergence of LLMs has fundamentally altered the role of NLP within applied informatics. Rather than acting solely as predictive classifiers, LLMs increasingly function as general-purpose analytical components capable of performing reasoning, summarization, and structured information extraction through prompt-based interaction.

Recent research explores LLMs as generators of weak supervision signals and annotation assistants. Deng et al. (2023) proposed a semi-supervised pipeline where LLMs generate financial sentiment labels, later used to train smaller models, demonstrating how LLMs can reduce manual annotation costs. Retrieval-augmented architectures further enhance performance by combining external knowledge sources with instruction-tuned models, improving contextual reasoning and classification accuracy (Zhang et al., 2023b).

Instruction-tuning approaches, such as Instruct-FinGPT (Zhang et al., 2023a), show that converting labelled financial datasets into instruction-based training examples significantly enhances model performance in numerically and contextually complex financial tasks. These developments position LLMs not merely as models but as flexible infrastructures capable of supporting modular analytical pipelines within information systems.

2.2 Financial NLP in Low-Resource and Portuguese Contexts

Despite rapid progress in LLM-based financial analytics, research addressing low-resource languages remains limited. Portuguese-language financial NLP represents an underexplored domain, particularly for European Portuguese, where publicly available datasets and evaluation benchmarks are scarce.

Recent efforts such as DeB3RTa (Pires et al., 2025) demonstrate the benefits of training transformer-based architectures on Portuguese financial corpora, achieving improved performance compared to multilingual baseline models. These results confirm that linguistic specialization plays a crucial role in accurately capturing financial semantics. Nevertheless, the development and evaluation of LLM-driven pipelines tailored to Portuguese financial content remain largely unexplored. Despite these advances, sentiment analysis remains particularly challenging in financial contexts due to domain-specific terminology, implicit sentiment expression, multilingual variability, and sensitivity to linguistic constructs such as negation and sarcasm (Loughran & McDonald, 2011; Malo et al., 2014). These challenges are amplified in low-resource languages, where limited annotated datasets hinder systematic evaluation and reproducibility (Araci, 2019).

Moreover, existing research frequently emphasizes model development rather than standardized evaluation pipelines, limiting reproducible comparison across models and deployment contexts (Rogers et al., 2020; Bommasani et al., 2021). Taken together, these studies demonstrate rapid progress in financial NLP but reveal limited attention to reproducible evaluation pipelines for low-resource financial languages.

2.3 Research Gap and Positioning of This Study

The reviewed literature reveals three main trends: (i) the integration of sentiment analysis into financial decision-support systems, (ii) the emergence of domain-specialized language models, and (iii) the increasing use of LLMs as flexible analytical components. However, a clear gap persists at the intersection of these streams.

Existing studies predominantly focus on English-language datasets or highly specialized proprietary models, while reproducible pipelines designed for low-resource financial languages remain scarce. Furthermore, limited attention has been given to evaluating LLM robustness and operational behaviour under controlled experimental conditions in multilingual financial contexts.

To address this gap, the present study introduces PipeFinePT, a modular and reproducible pipeline designed to operationalize LLM-based sentiment analysis for European Portuguese financial news. By combining dataset construction, prompt-driven annotation, and systematic benchmarking across multiple models, this work contributes an applied informatics perspective that bridges advances in LLMs with practical financial information system applications.

3 METHODOLOGY AND SOLUTION PROPOSAL

3.1 Research Design

This study adopts an applied informatics research design aimed at evaluating the feasibility and robustness of LLMs for financial sentiment analysis in a low-resource linguistic context. The research is structured around the development and empirical assessment of a modular analytical pipeline, enabling controlled experimentation and reproducible evaluation across multiple models.

The methodological approach follows four main stages: (i) dataset construction and annotation, (ii) pipeline design and implementation, (iii) prompt-driven model evaluation, and (iv) comparative performance analysis. These stages were designed to systematically address the research questions defined in Section 1 by assessing classification performance, operational characteristics, and robustness under controlled experimental conditions.

The overall workflow is operationalized through the PipeFinePT pipeline, described in the following subsection.

3.2 PipeFinePT Architecture

To operationalize sentiment analysis within a financial information-processing workflow, this study proposes PipeFinePT, a modular pipeline designed to transform raw financial news into structured analytical outputs suitable for downstream financial applications.

The architecture follows a modular design and consists of four main components:

1. **Data Acquisition Module** – responsible for extracting financial news articles from publicly available online sources using a web crawler.
2. **Preprocessing Module** – performs cleaning, normalization, and sentence extraction to ensure consistency and quality of textual inputs.
3. **LLM Annotation Module** – interacts with pre-trained language models through structured prompts, generating entity identification and sentiment classification outputs.
4. **Evaluation Module** – compares model predictions with human-annotated ground truth using standardized performance metrics.

The functional architecture of PipeFinePT is illustrated in Figure 1. As illustrated in Figure 1, extracted news content is first standardized and stored in a financial sentence repository. These inputs are continuously supplied to the annotation module, where LLM process each instance according to predefined prompting instructions. The resulting structured outputs are subsequently evaluated using standardized performance metrics.

This modular structure enables extensibility and reproducibility while allowing independent adjustment of pipeline components. The design reflects an applied informatics perspective in which LLMs function as analytical services embedded within an information system rather than standalone predictive models.

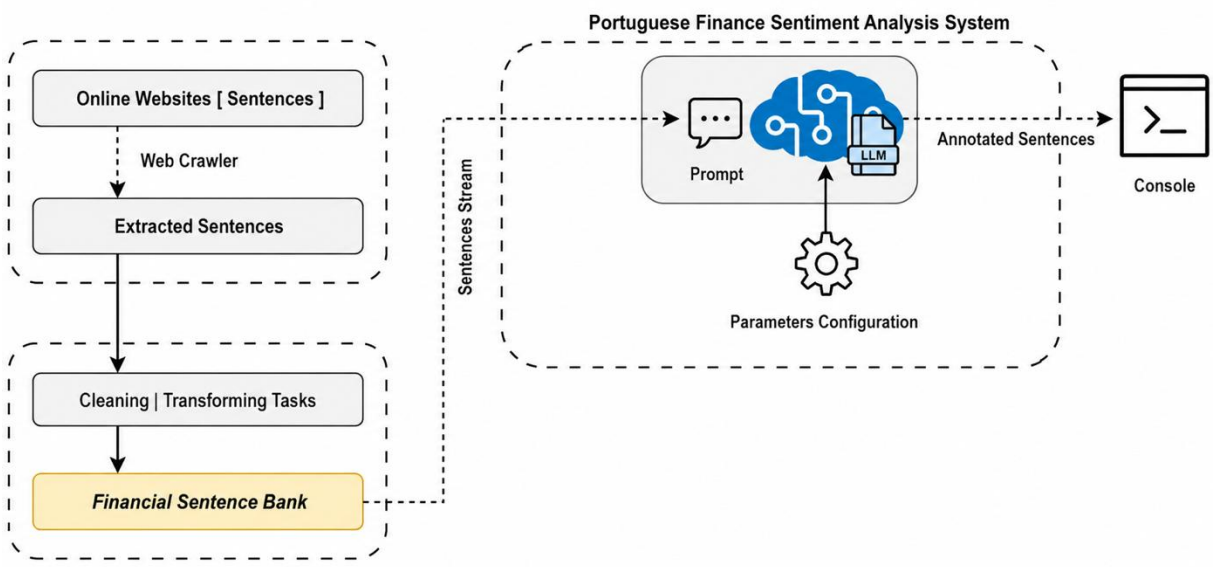

Figure 1. *PipeFinePT Functional Architecture.*

Unlike general-purpose LLM benchmarking pipelines, whose contribution lies mainly in the breadth of models or tasks covered, the novelty of PipeFinePT is not architectural but domain- and language-specific: it operationalises the full evaluation loop (crawling, Portuguese-specific cleaning, strictly formatted prompt-based annotation, and standardized scoring) for European-Portuguese financial news, a low-resource setting for which no reproducible evaluation pipeline previously existed. Its modular design additionally allows each component (data source, preprocessing rules, model, and metric) to be swapped independently, so the pipeline doubles as a reusable benchmarking harness rather than a one-off experiment.

The defined pipeline follows the state of the art regarding LLMs evaluation with the certain adaptations (such as, in the cleaning task, certain punctuations must be removed or adapted) linked to the Portuguese language.

3.3 Dataset Construction and Annotation

To support the proposed framework and address the scarcity of Portuguese-language financial datasets, a domain-specific sentiment analysis dataset was constructed. While English-language resources for financial sentiment analysis exist, datasets targeting European Portuguese remain limited, creating challenges for systematic evaluation.

The dataset was created by crawling a publicly accessible online financial platform and extracting economic news items written in European Portuguese. The collected texts were curated and processed to ensure linguistic consistency and suitability for sentiment classification tasks. Depending on the model used, the dataset consists of approximately 120k to 130k tokens.

A total of 200 financial news items reflecting the perceived impact of each news item on the referenced company's stock performance were manually annotated using a three-class sentiment scheme: 112 positive reviews (56%); 48 negative reviews (24%); 40 neutral reviews (20%). Each item was independently labelled by three annotators, and final ground-truth labels were obtained through majority voting to reduce individual bias. The average pairwise Cohen's Kappa of inter-annotator agreement was of 0.89 and the Fleiss' Kappa was 0.81. Although relatively small, the dataset size was considered adequate for controlled benchmarking experiments focused on comparative model behaviour rather than large-scale training. Although the dataset is not publicly available, the authors are ready to share it via email, for research purposes, as stated in the Data Availability statement.

Table 1 summarizes the main characteristics of the dataset used throughout all experiments.

Table 1. *Dataset summary.*

Item	Description
Total number of queries	200

Item	Description
Source	Public online platform (open source)
Language	Portuguese
Annotation scheme	Positive / Negative / Neutral
Annotators per query	3
Annotation type	Manual
Ground truth creation	Majority voting from annotations
Class distribution	Positive: 112 (56%); Negative: 48 (24%); Neutral: 40 (20%)
Inter-annotator agreement	Cohen's Kappa = 0.89; Fleiss' Kappa = 0.81

The resulting dataset serves both as input to the PipeFinePT pipeline and as a reproducible benchmarking resource for evaluating LLM performance in a low-resource financial language setting. Instead of fine-tuning models, this study employs a prompt-driven inference strategy reflecting realistic deployment scenarios where retraining large models may not be feasible. The annotation workflow evaluates each LLM's ability to interpret financial news and produce structured outputs under identical prompting conditions.

Each model received financial news instances containing an identifier, title, and full article content written in European Portuguese. The models were instructed to perform two tasks simultaneously:

- identify the company referenced in the news article;
- classify sentiment as positive, neutral, or negative according to its expected impact on stock market performance.

To ensure methodological transparency and reproducibility, all models were required to return results exclusively in CSV format containing three fields: id, entity, and classification. This strict formatting enabled automated comparison with the human-annotated ground truth while minimizing variability caused by response formatting differences.

The structured prompt used consistently across all experiments is presented below, in Prompt 1.

Prompt 1. *LLM Input Instructions.*

```
Given the following economic news, from a portuguese website (written in Portuguese from Portugal), identify which is the company that the article is mentioning and classify the news as "positive", "neutral" or "negative", given how what is being mentioned might affect the shares of that company on the stock market.

For each news you will receive an id, a title and the contents of the news. In the end, provide a CSV with three columns:

- id: with the id of the original news

- entity: the company being discussed in the article

- classification: "positive", "neutral" or "negative"

Do not include anything else in your reply besides the CSV itself!
```

An example of the standardized output generated by the models is shown in Listing 1.

Listing 1. *Example of model output in CSV format.*

```
0, Apple, neutral
1, Tesla, positive
2, Banco Santander, negative
3, EDP, neutral
4, Galp Energia, positive
( . . . )
```

Each output row corresponds to a single news item and includes the original identifier, detected entity, and predicted sentiment classification. This standardized representation simplifies automated evaluation and cross-model comparison.

Using prompt-based interaction allows evaluation of models as deployable analytical components without additional training, directly addressing practical deployment scenarios in financial information systems (RQ4).

3.4 Model Selection and Experimental Setup

Multiple LLMs were evaluated to analyse differences in performance and operational characteristics (RQ2). The selected models represent diverse deployment paradigms, including proprietary models and open-weight architectures. The evaluation deliberately focused on general-purpose LLMs rather than models fine-tuned for financial sentiment analysis, because general-purpose assistants are the tools most readily adopted by non-specialist end users and by organizations that lack the resources to train or fine-tune dedicated models. Assessing such off-the-shelf models under a uniform, prompt-based protocol therefore reflects a realistic deployment scenario and isolates the models' native linguistic and reasoning capabilities in European Portuguese, without the confounding effect of task-specific adaptation.

All models were executed using identical prompts, datasets, and evaluation procedures to ensure fairness and comparability. For simplicity and to guarantee a uniform execution environment, all models — both proprietary and open-weight — were accessed through the OpenRouter API. All experimental runs were executed on 20 May 2026, with each model receiving the complete batch of news items in a single request per run (Section 3.5). This does not restrict the open-weight models to cloud use: because their weights are publicly available, models such as LLaMA-4-Maverick-17B and Mistral-Medium-3 can equally be deployed on local, GPU-accelerated infrastructure given sufficient computational resources, allowing full control over inference parameters and reproducibility while keeping all data within the organization. Local deployment is nevertheless subject to each model's licence terms. In particular, the Llama 4 Community License Agreement does not grant its rights to individuals domiciled, or companies with a principal place of business, in the European Union (Meta, 2025), which currently precludes local deployment of LLaMA-4 models by EU-based organizations; Mistral models carry no such geographic restriction. The implications of this constraint are discussed in Sections 4.2 and 5.

Table 2 provides an overview of the evaluated models, including deployment type, operational characteristics, and observed linguistic handling capabilities.

Table 2. *Comparison of LLMs used for financial sentiment analysis.*

Model	Company	Inference	Parameters	PT Language Handling	Remarks
Gemini-2.5-Flash	Google	Proprietary	Not publicly disclosed	Moderate	Optimized for speed; some loss in nuance
GPT-4.1-mini	OpenAI	Proprietary	Not publicly disclosed	Good	Lightweight version; stable outputs
GPT-4.1-nano	OpenAI	Proprietary	Not publicly disclosed	Fair	Fastest GPT variant, but shallow reasoning
GPT-4.1	OpenAI	Proprietary	Not publicly disclosed	Excellent	Strong semantic understanding
LLaMA-4-Maverick-17B	Meta	Open-Weights	17B	Good	Balanced performance
Mistral-Medium 3	Mistral	Open-Weights	141B	Very good	Competitive performance on instruction tasks
O3	OpenAI	Proprietary	Not publicly disclosed	Good	Flexible runtime, adjusting reasoning effort

Notes: Parameter counts are reported as "Not publicly disclosed" because the providers (OpenAI, Google, and Mistral) do not officially release the parameter sizes of these models; the figures circulating online are unverified and are therefore not used as evidence. Indicative per-token API costs are reported separately in Table 3.

Outputs were generated in standardized CSV format containing three attributes: news identifier, detected entity, and sentiment classification. This diversity of models enables systematic comparison between efficiency, semantic understanding, and practical deployment considerations within applied financial informatics contexts.

Table 3. *Model token pricing (in USD) of input and output of the different LLMs, when used via API via OpenRouter, as recorded on 20 May 2026.*

Model	Price/1M input tokens	Price/1M output tokens
gemini-2.5-flash	$0.30	$2.50
gpt-4.1	$2	$8
gpt-4.1-mini	$0.40	$1.60
gpt-4.1-nano	$0.10	$0.40
llama-4-maverick-17b	$0.15	$0.60
mistral-medium-3	$0.30	$2.50
o3	$2	$8

Notes: Prices correspond to the rates displayed by OpenRouter on 20 May 2026 for the endpoints used in our experiments. OpenRouter is an aggregator that routes requests to upstream providers; the specific upstream provider serving each request was not logged during the experiments, and the same model may be offered at different prices by other providers (e.g., OpenAI direct, Microsoft Azure, AWS Bedrock) or by OpenRouter itself at different times. API pricing is revised frequently, so these figures are time-sensitive and are reported solely to contextualize the relative cost tiers of the evaluated models; readers should consult current provider documentation for up-to-date rates.

3.5 Evaluation Metrics and Robustness Analysis

Model performance was assessed using widely adopted classification metrics: accuracy rate (percentage of predictions that are correct), precision (of the items predicted positive, the proportion that is indeed positive), recall (also known as *sensitivity* - proportion of actual positives that were correctly detected), and F1-score (the harmonic mean of precision and recall, balancing both). These metrics collectively evaluate predictive correctness, class discrimination capability, and balance between false positives and false negatives.

To investigate robustness (RQ3), experiments were repeated using reversed input ordering. This additional evaluation tested whether sequence effects or hidden contextual dependencies influenced model behaviour, an aspect particularly relevant for prompt-based LLMs evaluation. In both runs the complete set of 200 articles was submitted within a single batch prompt, and robustness was assessed by reversing the order of the items in that batch (from news id 0, 1, …, 198, 199 to news id 199, 198, …, 1, 0) while keeping their content unchanged. Because all items share a single context window, the relative position of an article can influence the model's output, which is precisely the effect this experiment isolates. As the test relies on a single paired comparison per model rather than repeated sampling, its results should be interpreted as indicative of ordering sensitivity rather than as precise effect-size estimates.

By maintaining identical experimental conditions across runs, differences in performance could be attributed to model characteristics rather than dataset variation. The methodological design directly operationalizes the research questions:

- RQ1 is addressed through evaluation on a manually annotated European Portuguese dataset;
- RQ2 through comparative benchmarking of heterogeneous LLMs architectures;
- RQ3 through robustness testing using reversed query ordering;
- RQ4 through the design and validation of the PipeFinePT modular pipeline.

This structured alignment ensures methodological coherence between research objectives, experimental procedures, and analytical outcomes. Having established the dataset, architectural design, and evaluation framework, the PipeFinePT modular pipeline provides the foundation for analysing how LLMs behave as operational analytical components within applied financial information systems.

4 RESULTS ANALYSIS AND DISCUSSION

From an applied informatics perspective, the results illustrate how the PipeFinePT modular pipeline enables LLMs to function as operational analytical components within financial information systems. This section presents the experimental results obtained through the PipeFinePT pipeline and discusses their implications in relation to the research questions defined in Section 1. The analysis focuses on three complementary dimensions: (i) sentiment classification performance in a low-resource linguistic context, (ii) comparative behaviour across different categories of LLMs, and (iii) robustness of prompt-based inference under varying experimental conditions.

All evaluated models processed identical inputs using the same dataset and structured prompt configuration described in Section 3. Model outputs were automatically compared against the human-annotated ground truth, enabling consistent and reproducible evaluation across heterogeneous LLMs architectures.

The analysis proceeds in three stages. First, overall classification performance is examined to evaluate whether LLMs can reliably perform financial sentiment analysis in European Portuguese (RQ1). Second, performance differences across deployment paradigms are analysed (RQ2). Finally, robustness is assessed through reversed input ordering experiments (RQ3), followed by a synthesis of implications for applied financial information systems (RQ4).

4.1 Sentiment Classification Performance

Table 4 summarizes the classification performance of the evaluated models using accuracy, precision, recall, and F1-score metrics. These measures collectively assess predictive correctness, discrimination capability between sentiment classes, and balance between false positives and false negatives.

Table 4. *Classification metrics comparison by model.*

Model	Accuracy	Precision	Recall	F1
Gemini-2.5-Flash	(1)	(1)	(1)	(1)
GPT-4.1	0.781	0.735	0.776	0.747
GPT-4.1-mini	0.756	0.703	0.727	0.710
GPT-4.1-nano	(1)	(1)	(1)	(1)
LLaMA-4-Maverick-17B	0.697	0.677	0.704	0.671
Mistral-Medium 3	(1)	(1)	(1)	(1)
O3	(1)	(1)	(1)	(1)

To compute the evaluation metrics, we used the results extracted from each LLM. However, the files identified with "1" in Table 4 contained noisy data as output (LLM doesn't followed the user prompt), so we could not be certain of their validity; therefore, they were discarded. Specifically, Gemini-2.5-Flash, GPT-4.1-Nano, Mistral-Medium-3, and O3 repeatedly returned outputs that did not conform to the requested CSV-only format, which prevented a reliable automated comparison; these four models were therefore excluded from the quantitative analysis. Accordingly, the discussion below focuses on the three models that produced valid, conforming outputs across both runs (GPT-4.1, GPT-4.1-Mini, and LLaMA-4-Maverick-17B), and the inability of the remaining models to follow a strict output format is itself reported as an operational finding. No model-specific prompt engineering (e.g., few-shot output examples or reformulated instructions) was attempted to recover conforming outputs from these models, since the evaluation protocol deliberately fixed a single, identical zero-shot prompt for all models; any per-model prompt adaptation would have compromised the strict comparability that the protocol was designed to guarantee. Investigating whether adjusted prompts could recover valid outputs from the excluded models is identified as future work (Section 5).

The results indicate that LLMs are capable of performing sentiment classification in European Portuguese with moderate reliability, despite the limited availability of domain-specific linguistic resources. GPT-4.1 achieved the highest overall performance, achieving an accuracy of 0.781 and a balanced F1-score of 0.747, suggesting strong consistency between predicted and ground-truth labels, although, as reported below, the pairwise differences among the three valid models are not statistically significant.

GPT-4.1-Mini followed closely, with competitive accuracy (0.756) but slightly lower precision and recall, indicating a somewhat higher incidence of misclassifications. The open-weight LLaMA-4-Maverick-17B achieved the lowest accuracy among the three valid models (0.697), suggesting a more limited capability in handling contextual and domain-specific sentiment interpretation.

LLaMA-4-Maverick-17B, the only open-weight model that produced valid outputs, showed stable intermediate performance across metrics; the remaining open-weight model, Mistral-Medium-3, could not be evaluated quantitatively because its outputs did not conform to the required CSV format. Observations concerning open-weight models in this study therefore refer to LLaMA-4-Maverick-17B specifically and cannot be generalized to open-weight models as a category.

To assess whether the observed differences between the three valid models were statistically significant, pairwise McNemar tests were conducted on the instance-level predictions of each model, with p-values adjusted for multiple comparisons using the Holm-Bonferroni correction. No statistically significant differences were found for any pair: GPT-4.1 versus GPT-4.1-Mini, GPT-4.1 versus LLaMA-4-Maverick-17B, and GPT-4.1-Mini versus LLaMA-4-Maverick-17B), with all adjusted p-values above 0.05. Although GPT-4.1 produced a higher number of exclusively correct predictions in some comparisons, the performance differences reported in Table 4 should therefore be interpreted as indicative rather than conclusive, particularly given the limited statistical power afforded by the 200-item dataset.

These findings are consistent with prior studies demonstrating that LLMs can achieve competitive performance in financial sentiment analysis without task-specific fine-tuning (Deng et al., 2023; Zhang et al., 2023a). However, the moderate performance levels observed in this study reinforce previous observations that domain adaptation and linguistic specialization remain critical factors for achieving higher accuracy in financial NLP tasks (Huang et al., 2022; Wu et al., 2023).

Overall, the findings confirm that LLM-based sentiment classification remains feasible in low-resource financial language settings, directly addressing RQ1, while also revealing substantial performance variability across the evaluated model architectures.

4.2 Comparative Analysis Across Model Categories

Beyond quantitative metrics, qualitative differences emerged across model families in terms of semantic interpretation, operational efficiency, and linguistic handling. It should be noted at the outset that, because only three models produced valid outputs — two proprietary OpenAI models and a single open-weight model — the proprietary-versus-open-weight contrast discussed in this subsection effectively rests on one representative per category among valid outputs. The observations below should therefore be read as model-level findings rather than categorical conclusions, and the underlying pairwise performance differences are not statistically significant (Section 4.1).

Higher-tier proprietary models such as GPT-4.1 — also among the most expensive to operate in our evaluation, at USD 2.00 and 8.00 per million input and output tokens, respectively (Table 3) — demonstrated stronger capabilities in interpreting implicit sentiment and complex financial phrasing during qualitative inspection of model outputs. These models were particularly effective when sentiment was expressed indirectly or embedded within nuanced economic narratives. However, their substantially higher inference cost may limit their applicability in cost-sensitive analytical environments. Conversely, lightweight API-based models such as GPT-4.1-Mini operate at a substantially lower cost tier (Table 3) but occasionally struggled to recognise sentiment nuances or correctly identify entities in ambiguous textual contexts.

LLaMA-4-Maverick-17B, the only open-weight model with valid outputs, provided a balanced trade-off between computational efficiency and semantic performance. Although all models were accessed in this study through the OpenRouter API, open-weight models of this kind can also be deployed on local, GPU-accelerated infrastructure given sufficient resources; this makes them particularly attractive for organizations that require data privacy and full control over their inference infrastructure, since no data needs to leave their environment. This potential advantage is, however, conditional on the applicable licence terms: the Llama 4 Community License Agreement does not grant its rights to individuals domiciled, or companies with a principal place of business, in the European Union (Meta, 2025). EU-based financial organizations therefore cannot currently rely on local deployment of LLaMA-4

models and would need to consider open-weight alternatives without geographic restrictions, such as Mistral models — although, in our evaluation, Mistral-Medium-3 did not produce format-conforming outputs.

Differences were also observed in handling European Portuguese linguistic characteristics. GPT-4.1 exhibited the most stable interpretation of idiomatic expressions and finance-specific terminology, whereas smaller models occasionally introduced translation bias or semantic inconsistencies. These observations highlight the importance of linguistic adaptation when deploying LLM-based analytics in multilingual financial ecosystems.

This trade-off between semantic depth and computational efficiency aligns with observations reported in recent financial LLMs research, where larger models tend to exhibit stronger contextual reasoning capabilities at the cost of increased computational demands (Bommasani et al., 2021; Zhao et al., 2026). The results therefore support the emerging view of LLMs as configurable analytical infrastructures rather than universally optimal solutions.

Together, these findings provide a partial answer to RQ2, suggesting that deployment paradigm and model tier can influence operational suitability within applied financial informatics contexts. However, given that the valid results include only one open-weight representative and that the pairwise performance differences are not statistically significant, these comparative observations should be regarded as indicative rather than conclusive.

4.3 Robustness Analysis: Input Ordering Effects

To evaluate robustness, experiments were repeated using reversed input ordering while maintaining identical experimental conditions. This test examined whether prompt-based inference remained stable under variations in data presentation sequence. The resulting performance metrics are presented in Table 5.

Table 5. *Classification metrics with reversed query order.*

Model	Accuracy	Precision	Recall	F1
Gemini-2.5-Flash	(1)	(1)	(1)	(1)
GPT-4.1	0.700	0.696	0.716	**0.682**
GPT-4.1-mini	0.715	0.665	0.697	0.673
GPT-4.1-nano	(1)	(1)	(1)	(1)
LLaMA-4-Maverick-17B	0.690	0.654	0.669	0.652
Mistral-Medium 3	(1)	(1)	(1)	(1)
O3	(1)	(1)	(1)	(1)

Overall, model rankings remained relatively consistent, indicating that input order had limited influence on global performance trends. However, several models exhibited measurable performance variation.

Among the three models that produced valid outputs, LLaMA-4-Maverick-17B was the most stable, its accuracy decreasing only marginally from 0.697 to 0.690 under reversed ordering. GPT-4.1-Mini showed a moderate decline (from 0.756 to 0.715), while GPT-4.1 exhibited the largest change (from 0.781 to 0.700), indicating that even higher-tier models are not immune to input-ordering effects when all items are processed within a shared batch context.

Overall, these results suggest that robustness to input ordering is not strictly tied to model tier: the open-weight LLaMA-4-Maverick-17B proved the most stable configuration, whereas the proprietary GPT-4.1 was the most affected by the reversal.

The observed sensitivity to input ordering also resonates with recent discussions regarding variability and reproducibility challenges in prompt-based LLM evaluation (Rogers et al., 2020). These findings highlight the importance of controlled experimental pipelines when assessing LLMs behaviour in applied analytical environments. These findings address RQ3, emphasizing the importance of stability testing when evaluating prompt-based LLM systems and highlighting potential sensitivity to input sequencing in practical deployments.

4.4 Discussion and Synthesis

Taken together, the results demonstrate that LLMs can effectively support financial sentiment analysis in low-resource linguistic environments, though performance depends strongly on model architecture, deployment strategy, and linguistic capabilities. From an applied informatics perspective, the objective of the experimental evaluation is not solely to identify the best-performing model but to understand how large language models behave as operational analytical components within a controlled financial information system pipeline. Consequently, the results are interpreted in terms of reproducibility, robustness, and deployment trade-offs rather than isolated performance metrics.

The PipeFinePT pipeline enables systematic benchmarking by standardizing dataset preparation, prompting conditions, and evaluation procedures. Its modular design allows LLMs to function as interchangeable analytical components within financial information systems, facilitating reproducible experimentation and practical adoption.

The comparative analysis reveals a trade-off between semantic depth and operational efficiency: higher-tier models provide stronger contextual reasoning, whereas smaller or optimized models offer advantages in deployment cost. This trade-off reflects practical constraints faced by financial organizations when balancing analytical accuracy with infrastructure cost and response-time requirements. These comparative observations should nevertheless be interpreted with caution, as the pairwise differences among the three valid models are not statistically significant (Section 4.1) and the open-weight category is represented by a single valid model. Robustness testing further shows that prompt-based inference may exhibit sensitivity to experimental configuration, reinforcing the need for controlled evaluation frameworks.

Overall, the findings validate the PipeFinePT architecture as a reproducible approach for integrating LLM-based sentiment analysis into financial analytics workflows, thereby addressing RQ4 and bridging advances in large language models with real-world applied informatics applications.

5 CONCLUSIONS

This study introduces PipeFinePT**,** a modular pipeline designed to operationalize LLMs within applied financial information systems for sentiment analysis in European Portuguese, addressing challenges associated with low-resource linguistic environments. By combining dataset construction, prompt-driven annotation, and controlled benchmarking across multiple models, the proposed approach enables systematic evaluation of LLMs as operational analytical components while contributing an applied informatics perspective to financial text analytics.

The results provide evidence that LLMs can perform sentiment classification in European Portuguese with moderate reliability, thereby addressing RQ1. Despite the limited availability of domain-specific resources, several models demonstrated the ability to capture contextual sentiment signals in financial news, confirming the feasibility of applying LLM-based analysis beyond dominant English-language settings.

Regarding RQ2, the comparative evaluation could only be partially addressed: of the seven models evaluated, only three produced valid outputs, and the open-weight category was ultimately represented by a single model. Within this restricted set, the proprietary GPT-4.1 achieved the strongest semantic understanding, while the open-weight LLaMA-4-Maverick-17B offered a favourable balance between performance and operational efficiency; however, the pairwise performance differences were not statistically significant, so these observations are model-specific and cannot be generalized to proprietary and open-weight models as categories. These findings nevertheless highlight the importance of considering deployment context, licensing conditions, and computational constraints when integrating LLMs into real-world analytical workflows.

The robustness analysis addressed RQ3 by examining the impact of input ordering on classification outcomes. While overall model rankings remained relatively stable, performance fluctuations observed in certain models indicate sensitivity to experimental configuration. This result emphasizes the necessity of robustness validation when adopting prompt-based inference in applied informatics scenarios.

Finally, the development and evaluation of the PipeFinePT architecture respond directly to RQ4, demonstrating that a modular, prompt-driven pipeline can provide a reproducible framework for incorporating LLM-based sentiment analysis into financial information systems. The proposed approach supports systematic experimentation while enabling practical deployment without model fine-tuning, bridging research advances and operational application.

The proposed PipeFinePT pipeline demonstrates how LLMs can be integrated into financial information systems to support automated monitoring of market-relevant news and sentiment signals. Financial institutions and fintech organizations may leverage similar modular architectures to enhance decision-support tools while maintaining flexibility regarding model deployment and infrastructure constraints, subject to the licensing conditions applicable to each model (Section 4.2). Furthermore, the reproducible evaluation framework enables organizations to systematically compare LLMs solutions before operational adoption, supporting more reliable integration of AI-driven analytics into multilingual financial monitoring environments.

Despite these contributions, several limitations should be acknowledged. The dataset size remains relatively small and focuses exclusively on financial news articles, which may limit generalization across broader financial communication channels. Additionally, prompt-based evaluation introduces inherent variability associated with probabilistic model behaviour. Moreover, only three of the seven evaluated models produced format-conforming outputs, and the open-weight category was ultimately represented by a single valid model (LLaMA-4-Maverick-17B); the proprietary-versus-open-weight comparison addressed in RQ2 is therefore severely constrained, and its conclusions should be interpreted with corresponding caution. The small dataset also limits the statistical power of the model comparisons, and the observed pairwise differences were not statistically significant (Section 4.1). Furthermore, the model selection did not include models specifically recognized for strong multilingual capabilities or trained predominantly on Romance-language corpora, which may have influenced the observed performance in European Portuguese. Future work should therefore expand dataset coverage, incorporate additional financial text sources, and explore domain-adapted fine-tuning strategies for Portuguese-language models. A further direction is the systematic comparison of prompting strategies, such as zero-shot versus few-shot formulations: the present study deliberately fixed a single zero-shot prompt so that all models were evaluated under identical conditions, and a controlled study of prompt sensitivity is left to future work, including an assessment of whether adjusted prompts (e.g., few-shot output examples) could recover valid outputs from the models excluded for format non-compliance.

Further research may also investigate the integration of named entity recognition modules and automated evaluation dashboards to enhance analytical precision and scalability. Extending the pipeline toward real-time monitoring environments represents another promising direction, enabling continuous sentiment tracking within digital financial ecosystems.

Overall, this study demonstrates how LLMs can be systematically evaluated and operationalized within applied financial informatics, contributing a reproducible pathway for adopting advanced AI techniques in multilingual and underrepresented market contexts.

ADDITIONAL INFORMATION AND DECLARATIONS

Conflict of Interests: The authors declare no conflict of interest.

Author Contributions: G.C.: Software, Data curation, Conceptualization, Investigation, Methodology, Writing – Original draft preparation, Writing – Reviewing and Editing. J.C.: Software, Data curation, Conceptualization, Investigation, Methodology, Writing – Original draft preparation, Writing – Reviewing and Editing. R.P.M.: Conceptualization, Investigation, Methodology, Writing – Original draft preparation, Writing – Reviewing and Editing.

Statement on the Use of Artificial Intelligence Tools: The authors declare that they didn't use artificial intelligence tools for text or other media generation in this article.

Data Availability: Additional data that support the findings of this study are available from the corresponding author.

REFERENCES

Araci, D. (2019). FinBERT: Financial sentiment analysis with pre-trained language models. arXiv preprint arXiv:1908.10063. https://arxiv.org/abs/1908.10063

Ardekani, A. M., Bertz, J., Bryce, C., Dowling, M., & Long, S. (2024). FinSentGPT: A universal financial sentiment engine? *International Review of Financial Analysis*, 94, 103291. https://doi.org/10.1016/j.irfa.2024.103291

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., ... Liang, P. (2021). On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258. https://arxiv.org/abs/2108.07258

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodei, D. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems 33 (NeurIPS 2020), (pp. 1877–1901). NIPS.

Carnaz, G., Carvalho, J. M., & Marques, R. P. (2025). Finance Sentiment Analysis using Large Language Models in the Portuguese Context. In *2025 IEEE International Conference on Advances in Data-Driven Analytics and Intelligent Systems (ADACIS).* IEEE. https://doi.org/10.1109/ADACIS65663.2025.11436272

Carosia, A. E. O., Coelho, G. P., & Silva, A. E. A. da. (2019). Analyzing the Brazilian Financial Market through Portuguese Sentiment Analysis in Social Media. *Applied Artificial Intelligence*, 34(1), 1–19. https://doi.org/10.1080/08839514.2019.1673037

Delgadillo, J., Kinyua, J., & Mutigwe, C. (2024). FinSoSent: Advancing Financial Market Sentiment Analysis through Pretrained Large Language Models. *Big Data and Cognitive Computing*, 8(8), 87. https://doi.org/10.3390/bdcc8080087

Deng, X., Bashlovkina, V., Han, F., Baumgartner, S., & Bendersky, M. (2023). What do LLMs Know about Financial Markets? A Case Study on Reddit Market Sentiment Analysis. In *WWW '23 Companion: Companion Proceedings of the ACM Web Conference 2023*, (pp. 107–110). ACM. https://doi.org/10.1145/3543873.3587324

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT),* (pp. 4171–4186). ACL. https://doi.org/10.18653/v1/N19-1423

Huang, A., Wang, H., & Yang, Y. (2022). FinBERT: A Large Language Model for Extracting Information from Financial Text. *Contemporary Accounting Research*, 40(2), 806. https://doi.org/10.1111/1911-3846.12832

Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of Finance*, 66(1), 35–65. https://doi.org/10.1111/j.1540-6261.2010.01625.x

Malo, P., Sinha, A., Takala, P., Korhonen, P., & Wallenius, J. (2014). Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology*, 65(4), 782–796. https://doi.org/10.1002/asi.23062

Meier, L. F., & Althaus Junior, A. A. (2026). FinBERTimbau: Curated Brazilian Portuguese financial text corpus and sentiment annotation data (Version 1.0) [Data set]. *Zenodo*. https://doi.org/10.5281/zenodo.18463339

Meta. (2025). Llama 4 Community License Agreement. *Meta Platforms*. https://www.llama.com/llama4/license/

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1–2), 1–135. https://doi.org/10.1561/1500000011

Pires, H., Paucar, V. L., & Carvalho, J. P. (2025). DeB3RTa: A Transformer-Based Model for the Portuguese Financial Domain. *Big Data and Cognitive Computing*, 9(3), 51. https://doi.org/10.3390/bdcc9030051

Rogers, A., Kovaleva, O., & Rumshisky, A. (2020). A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8, 842–866. https://doi.org/10.1162/tacl_a_00349

Santos, L. L., Bianchi, R. A. da C., & Reali Costa, A. H. (2023). FinBERT-PT-BR: Análise de sentimentos de textos em português do mercado financeiro. In *Proceedings of the Brazilian Workshop on Artificial Intelligence in Finance (BWAIF),* (pp. 144–155). Sociedade Brasileira de Computação. https://doi.org/10.5753/bwaif.2023.231151

Tavares, C., Ribeiro, R., & Batista, F. (2021). Sentiment analysis of Portuguese economic news. In *Proceedings of the 10th Symposium on Languages, Applications and Technologies (SLATE 2021)*. Schloss Dagstuhl–Leibniz-Zentrum für Informatik. https://doi.org/10.4230/OASIcs.SLATE.2021.17

Wu, S., İrsoy, O., Lu, S., Dabravolski, V., Dredze, M., Gehrmann, S., Kambadur, P., Rosenberg, D., & Mann, G. (2023). BloombergGPT: A Large Language Model for Finance. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2303.17564

Zhang, B., Yang, H., & Liu, X.-Y. (2023a). Instruct-FinGPT: Financial Sentiment Analysis by Instruction Tuning of General-Purpose Large Language Models. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4489831

Zhang, B., Yang, H., Zhou, T., Babar, M. A., & Liu, X.-Y. (2023b). Enhancing Financial Sentiment Analysis via Retrieval Augmented Large Language Models. In *ICAIF '23: Proceedings of the Fourth ACM International Conference on AI in Finance*, (pp. 349–356). ACM. https://doi.org/10.1145/3604237.3626866

Zhao, W. X., Zhou, K., Li, J., Tang, T., Dong, Z., Hou, Y., Zhang, B., Min, Y., Zhang, J., Liu, P., Wang, X., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., . . . Wen, J. (2026). A survey of large language models. *Frontiers of Computer Science*, 20(12), 2012627. https://doi.org/10.1007/s11704-026-60308-3

Acta Informatica Pragensia is published by the Prague University of Economics and Business, Czech Republic | eISSN: 1805-4951