From Human to Scalable Annotation: Teaching LLMs to mimic experts labeling on Economic News Sentiment

Michele Petrocelli[1,2], Andrea Rollin[1], Matteo Berta[3,*], Francesca Zafonte[3], Simone Monaco[3], Salvatore Lo Sardo[3], Daniele Apiletti[3], Daria Scacciatelli[4] and Tania Cerquitelli[3]

[1]*Ministry of the Economy and Finance, Roma*
[2]*Guglielmo Marconi University, Roma*
[3]*Politecnico di Torino, Department of Control and Computer Engineering (DAUIN), Corso Castelfidardo, 34/d, 10138 Torino TO*
[4]*Sogei, Roma*

Abstract

Economic news sentiment offers a timely window into how the public perceives current economic conditions. These perceptions shape expectations, economic behavior, financial markets, and policy decisions, making news-based sentiment a valuable proxy for tracking and anticipating broader economic trends. Most existing approaches to economic news sentiment analysis rely on automated or weakly supervised labeling strategies to ensure scalability. However, the reliance on such training data, together with a strong dependence on English-centric sentiment lexicons, limits their ability to capture the heterogeneous expression of sentiment across economic subdomains in the Italian language. To address these limitations, the proposed work introduces a human-annotated corpus of Italian economic news, provides a systematic comparison between long-context encoder–decoder models and instruction-tuned language models, and develops an instruction-tuned model adapted via LoRA that achieves strong agreement with expert annotations.

Keywords

Economic Sentiment, Text Representation, News Annotation

1. Introduction

Sentiment analysis of economic texts has become an increasingly important tool for understanding macroeconomic dynamics and market behavior. Media coverage of economic events captures real-time changes in narratives, expectations, and perceptions, offering valuable qualitative information that complements traditional quantitative indicators [1]. By analyzing how economic conditions are discussed and framed in news articles, text-based approaches can provide timely insights into evolving economic trends [2].

Traditional economic indicators remain fundamental, but they are typically released with substantial delays [1, 3]. As a result, they may fail to capture rapid changes and are often slow to reflect the impact of exogenous shocks and unforeseen events. In this context, economic news sentiment provides a timely and complementary source of information, enabling the detection of emerging trends as they unfold. In particular, sentiment analysis can capture immediate shifts in expectations and reactions to unexpected developments, offering early signals of turning points that are not yet observable in conventional macroeconomic indicators.

Despite its relevance, measuring economic sentiment from text remains challenging. Conventional approaches to sentiment analysis in economics often rely on lexicon-based methods or shallow models [4], which struggle to capture the complexity, ambiguity, and context-dependence of economic language.

Published in the Proceedings of the Workshops of the EDBT/ICDT 2026 Joint Conference (March 24-27, 2026), Tampere, Finland
*Corresponding author.
michele.petrocelli@mef.gov.it (M. Petrocelli); andrea.rollin@mef.gov.it (A. Rollin); matteo.berta@polito.it (M. Berta); francesca.zafonte@polito.it (F. Zafonte); simone.monaco@polito.it (S. Monaco); salvatore.losardo@polito.it (S. L. Sardo); daniele.apiletti@polito.it (D. Apiletti); dscacciatelli@sogei.it (D. Scacciatelli); tania.cerquitelli@polito.it (T. Cerquitelli)
https://matteoberta.github.io/ (M. Berta); https://simonemonaco.github.io/ (S. Monaco)
0009-0009-3046-0386 (M. Berta); 0000-0003-4948-6120 (S. Monaco); 0000-0003-0538-9775 (D. Apiletti); 0000-0002-9039-6226 (T. Cerquitelli)

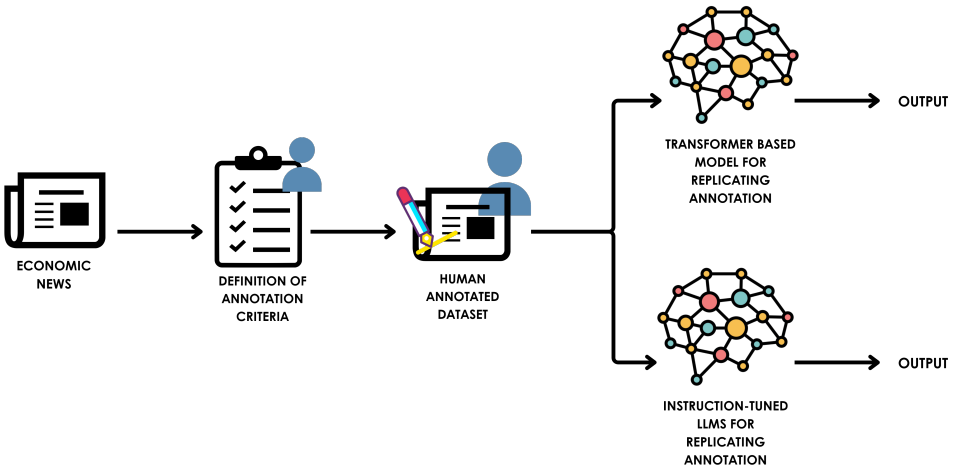

Figure 1: Overview of the annotation and modeling pipeline. Economic news articles are first collected and used to define annotation criteria, which guide the creation of a human-annotated dataset. This dataset is then used to train different transformer-based models aimed at replicating the human annotation process.

These limitations are particularly pronounced in specialized domains and for languages other than English, where domain-specific lexical resources and high-quality annotated datasets remain scarce [5].

To scale sentiment analysis to large corpora, many existing studies adopt automated or weakly supervised labeling strategies. While efficient, these approaches typically rely on generic or noisy labels and impose strong assumptions on sentiment expression, limiting their ability to reflect nuanced sentiment variations across economic subtopics. As a consequence, the quality and reliability of sentiment annotations remain a key bottleneck for downstream economic analysis.

Recent advances in transformer-based architectures have substantially improved the modeling of long and complex texts, while also increasing computational and memory requirements [6, 7]. Large language models benefit from extended contextual representations and strong performance across many language understanding tasks, but their practical adoption is often constrained in domain-specific and low-resource settings.

An alternative strategy is to adapt smaller transformer models to specific tasks using parameter-efficient fine-tuning methods such as Low-Rank Adaptation (LoRA) [8]. By updating only a limited number of parameters while keeping the backbone model fixed, these techniques reduce training costs and memory usage, enabling the specialization of open-source models for economic text analysis without full retraining.

Beyond scalability, an important challenge is whether language models can reproduce expert human annotation of economic news.

For tasks involving long input sequences, architectures explicitly designed for long documents provide an additional advantage. Longformer-based encoder–decoder models [9] enable the processing of entire news articles without truncation, preserving access to long-range contextual information. When combined with parameter-efficient fine-tuning, such models remain usable on standard hardware.

In this work, we study whether annotations in Italian economic news can be replicated using transformer-based models. We consider two complementary modeling strategies: (i) long-context encoder–decoder architectures and (ii) instruction-based language models adapted through parameter-efficient fine-tuning. The evaluation focuses on agreement with human annotators, the impact of different label aggregation strategies, and generalization across heterogeneous types of economic news (Figure 1).

Our analysis is conducted on a human-annotated corpus of approximately 800 Italian news articles published between 2007 and 2024. Articles are annotated for sentiment on a five-point ordinal scale and assigned to one of six economic topics. The results show that instruction-tuned language models achieve strong agreement with both individual and aggregated human annotations. In contrast, the Longformer Encoder–Decoder (LED) model [9] exhibits a systematic tendency toward regression to the

mean, limiting its ability to capture variations. These findings indicate that, in low-data settings, while long-context models are effective at processing extended documents, smaller instruction-based models adapted through parameter-efficient techniques offer more robust performance for replicating expert economic annotation.

2. Related Works

This section outlines the role of textual sentiment in economic analysis (2.1), discusses the main methodological approaches to sentiment extraction (2.2), and motivates the use of large language models as tools for replicating expert sentiment and topic annotations in economic news (2.3).

2.1. Economic Sentiment Analysis

Traditional indicators provide objective measures of economic activity, but they present two significant limitations: publication lags and limited forward-looking information. This temporal gap has motivated researchers to explore the use of textual data, which offers real-time availability and forward-looking information. Previous studies have shown that economic sentiment extracted from textual sources can serve as an effective proxy for anticipating key macroeconomic indicators, including GDP (Gross Domestic Product) growth, unemployment, and inflation, thereby providing valuable information for economic forecasting and policy analysis [5, 10, 11]. Text-based measures of sentiment capture expectations, uncertainty, and perceived economic conditions that are often not immediately reflected in traditional quantitative indicators [12, 13]. Recent econometric studies further support the predictive value of textual information: text-enhanced factor models incorporating news content improve GDP forecasting accuracy [14], while text-augmented VAR (Vector AutoRegression) and dynamic factor models highlight the role of central bank communication in predicting macroeconomic outcomes [15].

A parallel research effort has been devoted to define *sentiment*. Prior work has typically approached sentiment from two complementary perspectives: (i) the evaluation of sentiment expressed in the semantic content of the text, and (ii) the analysis of tone and emotional nuances conveyed through linguistic framing and stylistic choices. While both perspectives have proven informative in different contexts, this study adopts the second approach and focuses on sentiment as tone and emotional nuance in economic news articles. This choice reflects the central role of media framing in shaping economic narratives and public perceptions, independently of the underlying economic facts reported.

2.2. Lexicon-Based Approaches versus Deep Learning Methods

Lexicon-based approaches represent one of the most widely used methods for measuring economic sentiment. These approaches rely on the construction of domain-specific dictionaries designed to capture recurrent linguistic patterns in economic and financial texts. Prominent examples include the news-based sentiment measures proposed by Shapiro et al. [4], the financial sentiment dictionary developed by Loughran and McDonald [16], and related dictionary-based methods applied to news and market analysis [12, 17].

Lexicon-based methods are attractive due to their simplicity, transparency, long-standing use in the literature, and low computational cost [18]. However, these approaches present well-documented limitations, including strong dependence on domain and language specific vocabularies, limited ability to account for context, negation, and semantic composition, and reduced effectiveness when applied to complex or ambiguous texts [19, 20, 21]. These limitations are particularly pronounced for specialized domains and for languages different from English, where domain-specific lexical resources and annotated datasets remain scarce [5].

In response to these limitations, deep learning methods based on transformer architectures have been increasingly adopted. The literature continues to debate whether traditional lexicon-based approaches or more recent neural models provide superior performance [22, 23]. While lexicon-based models

remain competitive in some settings, transformer-based architectures are generally better suited to capture complex linguistic patterns and contextual dependencies when trained on high-quality data.

2.3. Large Language Models as Annotators

Recent advances in Large Language Models (LLMs) have demonstrated strong capabilities in sentiment analysis, achieving robust zero-shot and few-shot performance across diverse domains. Their ability to follow natural language instructions and leverage broad linguistic and contextual knowledge has made them promising candidates for tasks traditionally requiring expert human judgment, such as text annotation. However, these strengths do not extend uniformly to more complex sentiment analysis settings: as shown in [24] and [25], LLMs underperform smaller, task-specific supervised models on tasks involving ambiguity or fine-grained sentiment distinctions, underscoring the need for careful evaluation of LLM-generated annotations.

In parallel, a growing literature has explored the use of LLMs as annotators or evaluators for tasks traditionally performed by humans. Several studies show that, with clear guidelines and appropriate prompting, LLMs can approximate or match crowdworker-level performance in classification and span annotation tasks, offering substantial gains in scalability and cost efficiency [26], [27]. Nonetheless, alignment with human judgments remains imperfect, with LLMs typically achieving only moderate agreement and exhibiting sensitivity to prompt design, task structure, and annotation granularity.

Moreover, recent work highlights that LLMs inherit and may amplify systematic biases, including verbosity, authority, and aesthetic biases, raising concerns about their reliability as judges [28], [29]. While LLMs can correlate well with average human judgments, they struggle to capture annotator disagreement and demographic heterogeneity in subjective tasks where disagreement is informative rather than noise [30]. Accordingly, several studies caution against treating LLMs as drop-in replacements for expert annotators and advocate agreement-based, statistically grounded evaluation frameworks rather than accuracy against majority labels [31], [25].

3. Methodology

This section describes the construction of the annotated dataset, the human annotation protocol and agreement assessment (3.1), and the training and evaluation procedures adopted to compare long-document encoder–decoder models and instruction-tuned large language models in replicating humans sentiment annotations (3.2).

3.1. Human Annotation of Economic News

The study is based on a corpus of Italian economic news articles collected from major national newspapers through the Factiva database. A total of 800 articles were retrieved using a predefined set of economy-related keywords that has been defined with a group of economic experts and are related to several thematic areas (e.g, inflation, economic forecasting, social impact, politics and government measures, central banks and geopolitics) by avoiding narrow or sentiment-driven selection.

The corpus spans two distinct periods, 2007-2014 and 2017-2024, allowing the analysis to capture different economic cycles and institutional contexts. To reduce potential temporal biases, articles were balanced across publication years.

Each article is treated as an independent observation and annotated at the document level, without sentence-level or paragraph-level labeling. This choice reflects the objective of capturing the overall tone and the dominant economic theme conveyed by each article.

Sentiment is annotated on a five-point ordinal scale, where 1 indicates very negative sentiment and 5 indicates very positive sentiment. Annotators are instructed to focus on the tone and emotional nuances of each article rather than its factual economic content.

To promote consistency across annotators, detailed written guidelines are provided prior to the annotation task. These guidelines define sentiment levels and include examples illustrating different

cases. In instances of ambiguity, annotators are instructed to select the closest sentiment based on the overall emphasis and framing of the article or to assign a neutral label when sentiment is mixed.

Inter-annotator agreement is computed on the subset of articles annotated by all seven raters using Fleiss' κ [32],

$$\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e},$$

.

The statistic κ measures the degree of agreement among annotators beyond what would be expected by chance. Here, $\bar{P}$ denotes the average observed agreement across items, while $\bar{P}_e$ represents the expected agreement under random assignment of categories. The ratio normalizes observed agreement by its maximum possible value above chance, yielding $\kappa = 1$ for perfect agreement, $\kappa = 0$ for chance-level agreement and $\kappa < 0$ for disagreement. For this reason it estimates provide an indicator of the consistency and reliability of the annotation process and serve as a reference point for interpreting subsequent model performance.

3.2. Model Training and Evaluation

We selected three transformer-based architectures for sentiment annotation of Italian economic news: one long-context encoder–decoder model and two large language models. Their differing architectures, context windows, and training strategies enable a comparative assessment of suitability.

The first architecture is a *Longformer Encoder–Decoder (LED)* model [9], included as a representative long-context encoder–decoder baseline to assess whether access to full document context alone is sufficient to replicate expert annotation in a limited-data setting. LED model follows a supervised fine-tuning paradigm and is trained directly on the annotated dataset. Its architecture enables the processing of entire news articles without truncation, which is particularly relevant for economic texts where contextual information may be distributed across long documents.

The final two models considered are *Gemma-2-9B-IT*[1] [33] and *Phi-3.5 Mini Instruct*[2] [34], both instruction-tuned large language models. While both models adopt a decoder-only Transformer architecture with grouped-query attention, they differ in parameter scale and context capacity, with *Gemma-2-9B-IT* featuring a larger parameter count and *Phi-3.5 Mini Instruct* designed for higher efficiency and extended context handling. These models were selected based on their strong performance in Italian language understanding tasks, as documented by the Evalita-LLM benchmark [35][3]. Evalita-LLM provides a systematic evaluation framework for Italian NLP, comparing a wide range of open-source LLMs across ten tasks spanning both textual and multimodal settings. The benchmark results support the selection of these models as representative state-of-the-art systems for Italian text processing.

3.2.1. Longform Encoder-Decoder Training

The chosen backbone is `allenai/led-base-16384` because it supports sparse attention and is widely used for long-document modeling [9]. The maximum input length is set to 4096 tokens to process entire article. Global attention is restricted to the first token, which provides a document-level aggregation mechanism in Longformer-based models and preserves computational efficiency for long sequences.

We evaluate multiple loss functions because the sentiment labels are ordinal and different objectives impose different inductive biases. In particular, we consider: *MSE* as a regression baseline, modeling sentiment as a continuous variable bounded to the annotation scale; *Cross-Entropy*, which treats sentiment prediction as a multi-class classification task; *Soft Cross-Entropy* which extends cross-entropy to soft targets to capture annotation uncertainty; *Coral*, which formulates sentiment prediction as an ordinal regression problem, explicitly modeling the ordered structure of labels and *Distribution*, which

[1] https://huggingface.co/google/Gemma-2-9B-IT
[2] https://huggingface.co/microsoft/Phi-3.5-mini-instruct
[3] https://huggingface.co/spaces/evalitahf/evalita_llm_leaderboard

models sentiment as a discrete probability distribution and penalizes prediction errors according to their ordinal distance.

3.2.2. Instruction-Tuned LLM Training

The two large language models, GEMMA-2-9B-IT and PHI-3.5 MINI INSTRUCT, are fine-tuned using a supervised instruction-following setup. Training examples are formatted as instruction-completion pairs, where the model is prompted to annotate a full Italian economic news article with a sentiment score. The instruction explicitly defines sentiment as tone rather than factual content and constrains the output format to a valid JSON object, ensuring structured and machine-readable predictions. (see Appendix A).

Fine-tuning is performed using Low-Rank Adaptation (LoRA) [8], which enables efficient specialization of large models by updating a limited set of low-rank matrices while keeping the backbone parameters frozen. LoRA adapters are applied to the attention and feed-forward projection layers, automatically inferred from the model architecture. This strategy substantially reduces memory usage and training cost, allowing fine-tuning on standard hardware while preserving the representational capacity of the base models.

The subset of articles annotated by all human annotators is reserved exclusively for evaluation, serving as a common benchmark for assessing agreement between human annotators themselves and between models and human annotators. These articles are not used during training or validation.

The remaining articles, each annotated by a single annotator, are used for model training. For instruction-tuned large language models fine-tuned with LoRA, all available training articles (750 documents) are used for supervised fine-tuning, without an explicit validation split. Model selection and early stopping are based on loss computed on the held-out evaluation set.

For the Longformer Encoder-Decoder model, the 750 training articles are further split into training and validation sets using an 85/15 split. The validation set is used for early stopping and hyperparameter selection, while the final evaluation is conducted on the 50 multi-annotated articles. This split reflects the different training paradigms of the two approaches and ensures a fair comparison on a shared evaluation set.

4. Preliminary Evaluation Results

Results are primarily reported using agreement-based metrics, which are well suited to capturing alignment with human judgments, while traditional error and accuracy measures are reported as complementary indicators of model performance [31].

4.1. Dataset

Human annotation was conducted by seven annotators. To assess the reliability of the annotation scheme, a subset of 50 articles was annotated by all annotators, enabling the computation of inter-annotator agreement. The remaining articles were annotated by a single annotator and subsequently used for model training, balancing the need for agreement assessment with dataset scalability. For modeling purposes, sentiment labels are treated using label encoding that preserves the ordinal ordering of the scale.

The results, visible in the first line of Table 3, highlight the intrinsic difficulty of the sentiment annotation task. Agreement across all annotators is relatively low for sentiment, with $\kappa = 0.24$, reflecting the subjective and nuanced nature of tone-based sentiment assessment. When considering pairwise agreement, the highest concordance is observed for selected annotator pairs, with $\kappa = 0.41$.

Following Fleiss [32], κ values are interpreted relative to chance agreement, with larger values indicating stronger agreement. For qualitative interpretation, we additionally refer to commonly used benchmarks proposed by Landis and Koch [36], according to which values between 0.21 and 0.40

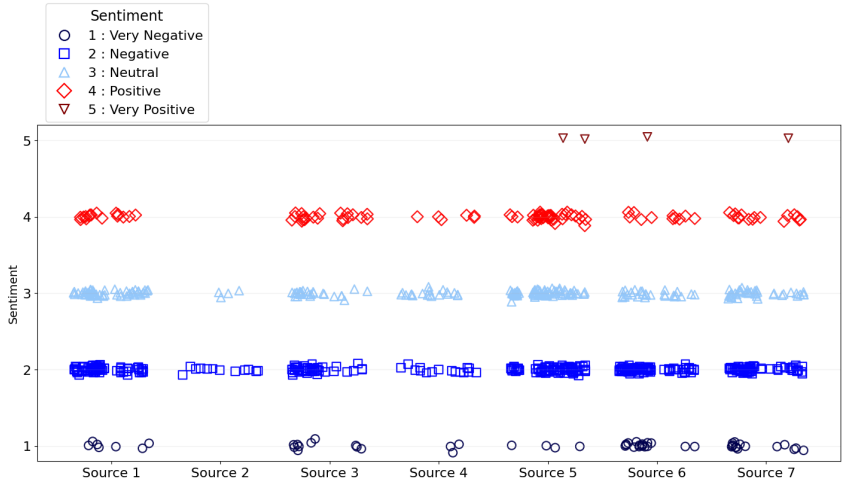

Figure 2: Sentiment distribution across sources of human annotation. Each dot represents an individual observation, with colors indicating discrete sentiment levels. Categories on the x-axis correspond to anonymized sources, with observations ordered chronologically within each category. The y-axis reports sentiment scores, allowing comparison of sentiment dispersion and central tendency across sources over time.

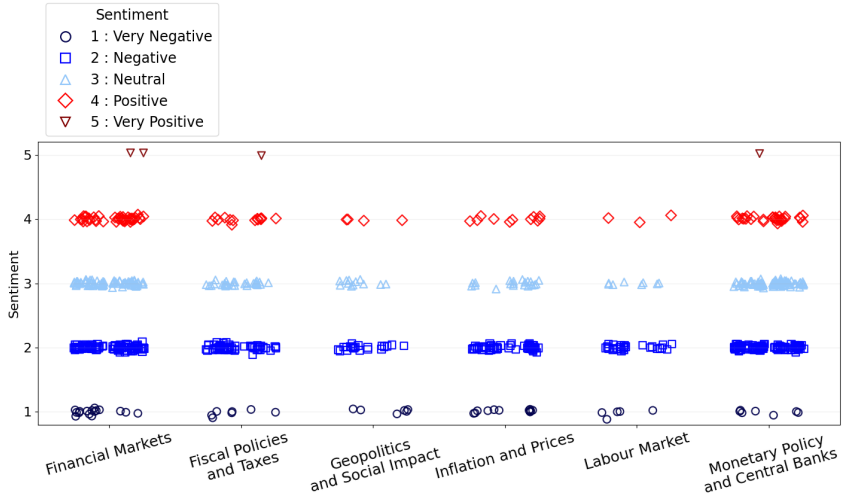

Figure 3: Sentiment distribution across topics of human annotation. Categories on the x-axis correspond to topics, with observations ordered chronologically within each category. The y-axis reports sentiment scores, allowing comparison of sentiment dispersion and central tendency across topics over time.

correspond to fair agreement, values between 0.41 and 0.60 to moderate agreement, and values above 0.60 to substantial agreement or perfect agreement.

Each article in our dataset is also annotated with exactly one topic label corresponding to the most salient economic theme discussed. The predefined topic categories, Monetary Policy and Central Banks, Financial Markets, Inflation and Prices, Fiscal Policies and Taxes, Labor Market, and Geopolitics and Social Impact, were defined in consultation with economic experts to reflect standard thematic distinctions in economic analysis.

It is interesting to see in Figure 2 and Figure 3 that expert sentiment annotations exhibit substantial dispersion across both economic topics and information sources, despite being measured on a discrete five-point scale. While most annotations cluster around negative, neutral, and positive categories, the presence of variation within each topic and source highlights the inherently subjective nature of sentiment assessment in economic text. Moreover, the similarity of the sentiment distributions across topics and sources suggests that heterogeneity in judgments is not driven by a single thematic dimension or outlet, but rather reflects systematic differences in interpretation among annotators.

Loss	MSE	MAE	Mean pred.	Std.
MSE	0.364	0.509	2.59	0.00
CE	0.365	0.507	2.87	0.01
CE (soft)	0.365	0.508	2.86	0.01
Distribution	0.411	0.517	2.88	0.01
CORAL	0.537	0.627	3.00	0.00

Table 1
Sentiment prediction performance of the LED model under different loss functions, evaluated on the multi-annotated test set. Lower MSE and MAE indicate better performance. Mean and standard deviation summarize the distribution of predicted sentiment scores, highlighting the collapse toward a single sentiment level across losses.

Model	MSE
Gemma (base)	0.3396
Gemma + LoRA	**0.2741**
Phi (base)	0.4327
Phi + LoRA	**0.3568**

Table 2
Performance of instruction-tuned LLMs before and after LoRA fine-tuning. Sentiment is evaluated using MSE against the mean human score.

4.2. Performance evaluation

4.2.1. Longformer Encoder–Decoder

Table 1 summarizes the performance of the Longformer Encoder–Decoder (LED) model under different loss functions and regularization settings. When evaluated using aggregate error metrics alone, several configurations achieve sentiment prediction errors comparable to those reported for instruction-tuned large language models. In particular, mean squared error values cluster around 0.36 for cross-entropy, soft cross-entropy, and bounded regression losses, indicating a reasonable average fit to the aggregated human sentiment scores.

At the same time, the joint inspection of error metrics and distributional statistics in Table 1 reveals a systematic limitation across all LED configurations. Regardless of the loss function employed, predicted sentiment values collapse toward a single level, with mean predictions tightly concentrated and near-zero standard deviation across the evaluation set. This pronounced regression-to-the-mean behavior suggests that low aggregate error can be achieved even when the model fails to capture meaningful variation in sentiment, highlighting the limits of relying solely on MSE- or MAE-based evaluation in this setting.

This behavior indicates that low aggregate error alone is not sufficient to assess the quality of sentiment predictions in this setting. While minimizing squared or absolute error favors predictions near the central tendency of the label distribution, such solutions fail to capture document-specific tonal variation. The observed collapse is consistent with the relatively low inter-annotator agreement for sentiment and highlights a limitation of supervised encoder-decoder architectures when applied to ordinal sentiment annotation with limited and noisy data.

4.2.2. Instruction-Tuned Large Language Models

Table 2 reports the performance of the instruction-tuned large language models, Gemma-2-9B-IT and Phi-3.5 Mini Instruct, comparing the base checkpoints with their LoRA-adapted versions. Sentiment prediction performance is measured using mean squared error (MSE) computed with respect to the mean sentiment value assigned by human annotators.

LoRA fine-tuning yields consistent improvements for both models. For Gemma-2-9B-IT, sentiment prediction error is reduced by approximately 19%, while Phi-3.5 Mini Instruct achieves an 18%

Task	Agreement	Mean κ	Min–Max
Sentiment	Human–Human	**0.24**	0.06–0.41
Sentiment	Gemma Base	0.2079	0.0270–0.3108
Sentiment	Gemma + LoRA	**0.2838**	0.1453–0.3827
Sentiment	Phi Base	0.2420	0.1342–0.3549
Sentiment	Phi + LoRA	0.2207	0.0344–0.3121

Table 3
Pairwise Cohen's κ agreement for sentiment annotation. We report human–human agreement (mean and range across annotator pairs) and model–human agreement for Gemma and Phi base models and their LoRA-adapted variants. Bold values indicate the highest agreement achieved among models.

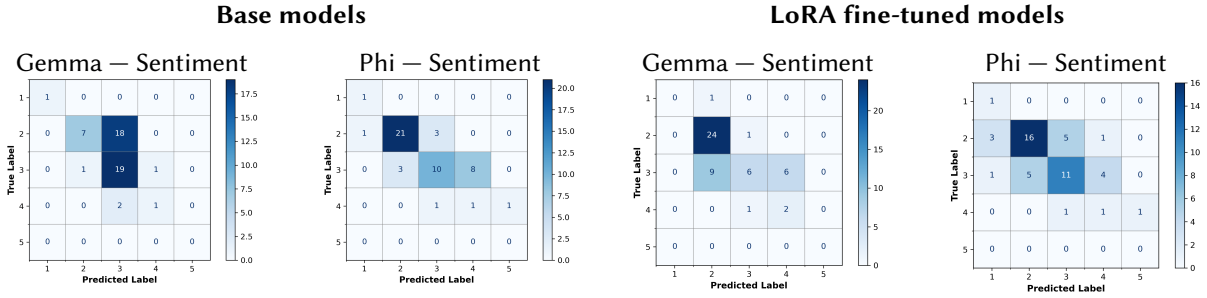

Figure 4: Confusion matrices for sentiment classification on the evaluation set. Left column: base instruction-tuned models; Right column: LoRA fine-tuned models.

reduction in MSE. These results demonstrate that parameter-efficient fine-tuning effectively aligns instruction-tuned large language models with the annotation guidelines and the linguistic characteristics of Italian economic news.

To better characterize the error patterns of the models, we also report confusion matrices (Figure 4). The labels predicted by the model are compared with labels derived from those assigned by humans: the sentiment target is the average human score rounded to the nearest class.

To further assess alignment with human judgments, we compute pairwise Cohen's κ [32] between model predictions and human annotations for both Gemma-2-9B-IT and Phi-3.5 Mini Instruct (Table 3). For Gemma, LoRA fine-tuning substantially improves agreement with annotators for sentiment classification.

In contrast, Phi-3.5 Mini Instruct exhibits limited or negative effects from LoRA fine-tuning: sentiment agreement improves only marginally.

Human–human agreement provides important context for interpreting these results. Topic annotation shows high consistency across annotators ($\kappa \approx 0.49$), whereas sentiment agreement is substantially lower ($\kappa \approx 0.24$), reflecting the inherently subjective nature of tone-based sentiment annotation. Within this context, sentiment performance for all models remains bounded by annotator disagreement.

5. Discussions

This study investigates whether sentiment annotations in Italian economic news can be reliably replicated by transformer-based models, with a particular focus on comparing long-context encoder-decoder architectures and instruction-tuned large language models adapted through parameter-efficient fine-tuning.

The Longformer Encoder-Decoder model demonstrates the ability to process entire news articles without truncation and achieves competitive aggregate error metrics for sentiment prediction under several loss functions. However, a detailed analysis reveals a systematic regression-to-the-mean behavior across all training configurations. This indicates that optimizing for aggregate error alone is insufficient to capture the nuanced and subjective nature of economic sentiment, particularly when inter-annotator

agreement is limited and the training data are relatively small.

In contrast, instruction-tuned large language models exhibit greater flexibility in replicating human annotation behavior. LoRA fine-tuning substantially improves sentiment accuracy for Gemma-2-9B-IT. These results suggest that instruction-following pretraining, combined with parameter-efficient adaptation, provides a strong inductive bias for modeling evaluative tone in economic texts. At the same time, the more heterogeneous behavior observed for Phi-3.5 Mini Instruct highlights that the effectiveness of LoRA depends on model capacity and alignment with the task-specific taxonomy.

A central aspect of this work is the use of human–human agreement as a benchmark for evaluating model performance. Sentiment annotation is inherently subjective and model outputs should therefore be interpreted relative to this level of human consistency rather than in absolute terms. The results also highlight the importance of high-quality human-annotated data and well-defined annotation guidelines: in their absence, even powerful model architectures may converge to trivial or weakly informative solutions.

Future work includes applying the proposed framework to larger corpora and to additional languages and economic contexts in order to provide a more effective test of *generalizability*, exploring LoRA fine-tuning across a broader range of instruction-tuned models with different architectures and parameter scales, and investigating uncertainty-aware annotation schemes, such as soft labels, distributions, or confidence scores, that could better capture the subjective nature of sentiment annotation. Further research should also evaluate commercial large language models and, considering both zero-shot and few-shot settings, compare proprietary systems with open-source fine-tuned models to clarify trade-offs in terms of transparency, cost, and annotation quality. Finally, integrating these methodologies into forecasting models of macroeconomic target variables may enhance their predictive performance, particularly by incorporating timely sentiment-based signals alongside traditional indicators.

Acknowledgments

We would like to sincerely thank all the expert annotators who contributed to the creation of the dataset. Their domain expertise, careful and consistent annotations, and the numerous fruitful discussions throughout the annotation process were essential to ensuring the quality, reliability, and relevance of the data.

Declaration on Generative AI

During the preparation of this work, the author(s) used GPT 5.2 in order to: grammar, spelling check and help in the correction of sentence structure. After using these tool, the author reviewed and edited the content as needed and takes full responsibility for the publication's content.

References

[1] D. Giannone, L. Reichlin, D. Small, Nowcasting: The real-time informational content of macroeconomic data, Journal of Monetary Economics 55 (2008) 665–676. URL: https://www.sciencedirect.com/science/article/pii/S0304393208000652. doi:https://doi.org/10.1016/j.jmoneco.2008.05.010.

[2] L. Barbaglia, S. Consoli, S. Manzan, Forecasting with economic news, Journal of Business & Economic Statistics 41 (2023) 708–719. URL: https://doi.org/10.1080/07350015.2022.2060988. doi:10.1080/07350015.2022.2060988(online).

[3] J. H. Stock, M. W. Watson, Macroeconomic forecasting using diffusion indexes, Journal of Business & Economic Statistics 20 (2002) 147–162.

[4] A. H. Shapiro, M. Sudhof, D. J. Wilson, Measuring news sentiment, Journal of Econometrics 228 (2022) 221–243. URL: https://www.sciencedirect.com/science/article/pii/S0304407620303535. doi:https://doi.org/10.1016/j.jeconom.2020.07.053.

[5] S. R. Baker, N. Bloom, S. J. Davis, Measuring economic policy uncertainty, The Quarterly Journal of Economics 131 (2016) 1593–1636.
[6] R. Thoppilan, et Al, Lamda: Language models for dialog applications, CoRR abs/2201.08239 (2022). URL: https://arxiv.org/abs/2201.08239. arXiv:2201.08239.
[7] J. W. Rae, et Al, Scaling language models: Methods, analysis & insights from training gopher, CoRR abs/2112.11446 (2021). URL: https://arxiv.org/abs/2112.11446. arXiv:2112.11446.
[8] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, Lora: Low-rank adaptation of large language models, in: Proceedings of the International Conference on Learning Representations (ICLR), 2022.
[9] I. Beltagy, M. E. Peters, A. Cohan, Longformer: The long-document transformer, CoRR abs/2004.05150 (2020). URL: https://arxiv.org/abs/2004.05150. arXiv:2004.05150.
[10] S. A. Sharpe, N. R. Sinha, C. A. Hollrah, The power of narrative sentiment in economic forecasts, International Journal of Forecasting 39 (2023) 1097–1121. URL: https://www.sciencedirect.com/science/article/pii/S0169207022000590. doi:https://doi.org/10.1016/j.ijforecast.2022.04.008.
[11] M. Gentzkow, B. Kelly, M. Taddy, Text as data, Journal of Economic Literature 57 (2019) 535–74. URL: https://www.aeaweb.org/articles?id=10.1257/jel.20181020. doi:10.1257/jel.20181020.
[12] P. C. Tetlock, Giving content to investor sentiment: The role of media in the stock market, The Journal of Finance 62 (2007) 1139–1168.
[13] P. C. Tetlock, M. Saar-Tsechansky, S. Macskassy, More than words: Quantifying language to measure firms' fundamentals, The Journal of Finance 63 (2008) 1437–1467.
[14] B. Seo, Econometric forecasting using ubiquitous news text: Text-enhanced factor model, International Journal of Forecasting 41 (2025) 1055–1072. doi:10.1016/j.ijforecast.2024.11.001.
[15] L. N. Ferreira, Forecasting with VAR-teXt and DFM-teXt models: Exploring the predictive power of central bank communication, Technical Report 559, Banco Central do Brasil, Working Paper Series, 2021.
[16] T. Loughran, B. McDonald, When is a liability not a liability? textual analysis, dictionaries, and 10-ks, The Journal of Finance 66 (2011) 35–65.
[17] D. Garcia, Sentiment during recessions, The Journal of Finance 68 (2013) 1267–1300.
[18] P. J. Stone, D. C. Dunphy, M. S. Smith, The General Inquirer: A Computer Approach to Content Analysis, MIT Press, 1966.
[19] B. Pang, L. Lee, Opinion mining and sentiment analysis, Foundations and Trends in Information Retrieval 2 (2008) 1–135.
[20] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, M. Stede, Lexicon-based methods for sentiment analysis, Computational Linguistics 37 (2011) 267–307.
[21] L. Young, S. Soroka, Affective news: The automated coding of sentiment in political texts, Political Communication 29 (2012) 205–231.
[22] R. Catelli, S. Pelosi, M. Esposito, Lexicon-based vs. bert-based sentiment analysis: A comparative study in italian, Electronics 11 (2022). URL: https://www.mdpi.com/2079-9292/11/3/374. doi:10.3390/electronics11030374.
[23] E. Öhman, The validity of lexicon-based sentiment analysis in interdisciplinary research, in: M. Hämäläinen, K. Alnajjar, N. Partanen, J. Rueter (Eds.), Proceedings of the Workshop on Natural Language Processing for Digital Humanities, NLP Association of India (NLPAI), NIT Silchar, India, 2021, pp. 7–12. URL: https://aclanthology.org/2021.nlp4dh-1.2/.
[24] W. Zhang, Y. Deng, B. Liu, S. J. Pan, L. Bing, Sentiment analysis in the era of large language models: A reality check, 2023. URL: https://arxiv.org/abs/2305.15005. arXiv:2305.15005.
[25] C. Shen, L. Cheng, X.-P. Nguyen, Y. You, L. Bing, Large language models are not yet human-level evaluators for abstractive summarization, in: H. Bouamor, J. Pino, K. Bali (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2023, Association for Computational Linguistics, Singapore, 2023, pp. 4215–4233. URL: https://aclanthology.org/2023.findings-emnlp.278/. doi:10.18653/v1/2023.findings-emnlp.278.
[26] X. He, et Al, AnnoLLM: Making large language models to be better crowdsourced annota-

tors, in: Y. Yang, A. Davani, A. Sil, A. Kumar (Eds.), Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track), Association for Computational Linguistics, Mexico City, Mexico, 2024, pp. 165–190. URL: https://aclanthology.org/2024.naacl-industry.15/. doi:10.18653/v1/2024.naacl-industry.15.

[27] Z. Kasner, V. Zouhar, P. Schmidtová, I. Kartáč, K. Onderková, O. Plátek, D. Gkatzia, S. Mahamood, O. Dušek, S. Balloccu, Llms as span annotators: A comparative study of llms and humans, 2025. URL: https://arxiv.org/abs/2504.08697. arXiv:2504.08697.

[28] G. H. Chen, S. Chen, Z. Liu, F. Jiang, B. Wang, Humans or LLMs as the judge? a study on judgement bias, in: Y. Al-Onaizan, M. Bansal, Y.-N. Chen (Eds.), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 8301–8327. URL: https://aclanthology.org/2024.emnlp-main.474/. doi:10.18653/v1/2024.emnlp-main.474.

[29] A. Elangovan, L. Liu, L. Xu, S. B. Bodapati, D. Roth, ConSiDERS-the-human evaluation framework: Rethinking human evaluation for generative large language models, in: L.-W. Ku, A. Martins, V. Srikumar (Eds.), Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 1137–1160. URL: https://aclanthology.org/2024.acl-long.63/. doi:10.18653/v1/2024.acl-long.63.

[30] J. Ni, Y. Fan, V. Zouhar, D. Rooein, A. Hoyle, M. Sachan, M. Leippold, D. Hovy, E. Ash, Can reasoning help large language models capture human annotator disagreement?, 2026. URL: https://arxiv.org/abs/2506.19467. arXiv:2506.19467.

[31] N. Calderon, R. Reichart, R. Dror, The alternative annotator test for LLM-as-a-judge: How to statistically justify replacing human annotators with LLMs, in: W. Che, J. Nabende, E. Shutova, M. T. Pilehvar (Eds.), Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Vienna, Austria, 2025, pp. 16051–16081. URL: https://aclanthology.org/2025.acl-long.782/. doi:10.18653/v1/2025.acl-long.782.

[32] J. L. Fleiss, Measuring nominal scale agreement among many raters, Psychological Bulletin 76 (1971) 378–382. doi:10.1037/h0031619.

[33] G. Team, Gemma: Open models based on gemini research and technology, arXiv preprint arXiv:2403.08295 (2024).

[34] M. Abdin, et al., Phi-3 technical report: A highly capable language model locally on your phone, arXiv preprint arXiv:2404.14219 (2024).

[35] B. Magnini, R. Zanoli, M. Resta, M. Cimmino, P. Albano, M. Madeddu, V. Patti, Evalita-llm: Benchmarking large language models on italian, 2025. URL: https://arxiv.org/abs/2502.02289. arXiv:2502.02289.

[36] J. R. Landis, G. G. Koch, The measurement of observer agreement for categorical data, Biometrics 33 (1977) 159–174.

A. Prompt

Note: The prompt shown here is the English-translated version of the original Italian prompt used in the experiments.

Listing 1: Instruction prompt used for LLM fine-tuning

```
SYSTEM PROMPT
-------------
You are an expert annotator of Italian economic news articles.

Your task is to analyze an article and produce the following annotation:

Sentiment: assign a score from 1 to 5 based exclusively on the overall linguistic
    tone (not on the factual content).
   - 1 = very negative
   - 2 = negative
   - 3 = neutral
   - 4 = positive
   - 5 = very positive

Rules:
- Sentiment must reflect only the communicative tone.
- If the tone is intermediate between two adjacent categories, use a value with a
    0.5 increment.
- Do not provide explanations or any additional text.

Response format (mandatory):
{"Sentiment": <number between 1 and 5>}

No other text before or after the JSON.

USER PROMPT
-----------
Analyze the following article and produce the required annotation.

ARTICLE:
[Article text]

Respond with JSON only.
```