# Enhancing Inflation Nowcasting with LLM
*Sentiment Analysis on News*

**Marc-Antoine Allard**[*], **Paul Teiletche**[*], **Adam Zinebi**[*]

[*]EPFL, Lausanne
{firstname.lastname}@epfl.ch

## Abstract

**This study explores integrating large language models (LLMs) into classic inflation nowcasting frameworks, particularly in light of high inflation volatility periods such as the COVID-19 pandemic. We propose InflaBERT, a BERT-based LLM fine-tuned to predict inflation-related sentiment in news. We use this model to produce NEWS, an index capturing the monthly sentiment of the news regarding inflation. Incorporating our expectation index into the Cleveland Fed's model, which is only based on macroeconomic autoregressive processes, shows a marginal improvement in nowcast accuracy during the pandemic. This highlights the potential of combining sentiment analysis with traditional economic indicators, suggesting further research to refine these methodologies for better real-time inflation monitoring.**

**The source code is available at https://github.com/paultltc/InflaBERT.**

## 1 Introduction

One of the significant economic consequences of the COVID-19 pandemic has been a surge in inflation worldwide. The pandemic caused massive disruptions in global supply chains, a rise in food prices, and the built-up of inflation expectations by economic agents. Inflation is commonly defined as the annualized percent change in consumer prices. According to the World Bank, world inflation surged from 2% in 2019, before the pandemic outbreak, to 8% in 2022[1], reaching levels unseen for most countries over nearly three decades. Since then, inflation has receded but remains significantly higher than before the pandemic.

Controlling inflation is one of the main objectives of major central banks such as the Federal Reserve of the United States, the European Central Bank, and the Swiss National Bank. Following the inflation shock, these central banks have proceeded with significant interest rate increases that significantly impacted the economy, notably real estate. It is a critical dimension of central banks' work to monitor real-time inflation dynamics. Over the recent years, central banks have notably developed inflation "nowcasting" models. "Nowcasting" stands for the contraction of "Now" and "forecasting". Originating in meteorology, it refers to "the prediction of the very recent past, the present, and the very near future state of an economic indicator" (Wikipedia). In economics, nowcasting has been originally applied to economic growth (Bańbura et al., 2013), but models of inflation nowcasting have been recently developed.

One such popular model has been developed by Knotek and Zaman (Knotek and Zaman, 2017, 2024) and is forming the basis of the inflation nowcasting model of the Federal Reserve Bank of Cleveland. The model combines through an Ordinary Least Squares (OLS) regression forecasts of three main inflation components (core inflation, food prices, energy) based on autoregressive processes and high-frequency data such as daily oil prices. The authors show that the forecasts of the nowcasting model outperform those of professional economists (as captured by the Survey of Professional Forecasters – SPF below).

More recently, nowcasting models are trying to benefit from machine learning and artificial intelligence development. These methodologies allow researchers to efficiently process a more extensive base of indicators than traditional economic statistics, including internet search trends, social media sentiment, or credit card transaction data, which can all be relevant to tracking the economy and inflation in real-time. Recent examples include Angelico et al. (Angelico et al., 2022), de Bandt et al. (Olivier et al., 2023), and Beck et al. (Beck et al., 2024). The advent of Large Language Mod-

[1] See the World Bank indicator here.

els (LLMs) also offers the opportunity to enhance the inflation nowcasting models, as recently shown by Faria e Castro and Leibovici (Faria-e Castro and Leibovici, 2023) with Google AI's PaLM.

This project investigates whether these recent methodologies can improve traditional inflation nowcasting models. We model news inflation sentiment by introducing `InflaBERT`, a `BERT`-based (Devlin et al., 2018) LLM fine-tuned to predict inflation-related news sentiment. To train this model we use an extensive news database extracted from the `FNSPID` dataset (Dong et al., 2024). We then propose NEWS, an index replicating monthly news sentiment regarding inflation. Finally, we test whether considering this index can improve the predictions of the Cleveland Fed nowcasting model. We focus on the COVID period (2020-2023) to analyze how these models would have allowed us to better track the strong moves and volatility in inflation during that period.

## 2 Related literature

Our paper contributes to the literature on improving classic inflation nowcasting techniques through the use of ML and AI techniques to capture expectations. The integration of AI in economic forecasting, specifically inflation prediction, is a burgeoning field that leverages advancements in machine learning and natural language processing. Recent studies, such as the one conducted by Faria-e-Castro and Leibovici (2024) at the Federal Reserve Bank of St. Louis (Faria-e Castro and Leibovici, 2023), explore the potential of large language models (LLMs) like Google's PaLM to generate accurate inflation forecasts. This research demonstrates that LLMs can produce conditional inflation forecasts with lower mean-squared errors compared to traditional methods such as the Survey of Professional Forecasters (SPF) over multiple years and forecast horizons. The findings suggest that AI-based models can offer an effective and cost-efficient alternative to traditional expert and survey-based forecasting methods. This study aligns with previous works like those by Bybee (Bybee, 2023), who used GPT-3.5 to simulate economic expectations, and extends the understanding of how LLMs can be utilized for macroeconomic and financial predictions.

Despite the promising capabilities of large language models (LLMs) in inflation forecasting, several limitations and challenges persist. Notably, no specific training was conducted on inflation data; instead, instruction tuning was employed. Instructing the LLM to neglect future data is not entirely robust because the model inherently uses weights that were trained on more recent data. This reliance makes it difficult to rigorously test the validity of the model's conditional forecasts as truly out-of-sample. Additionally, the inherent "black-box" nature of LLMs complicates understanding the underlying mechanics driving their predictions, posing challenges for transparency and interpretability in economic forecasting.

## 3 News Inflation Sentiment Analysis

Improving inflation nowcasting by incorporating population expectations seems reasonable when considering the string correlation between human perceptions and realized inflation. For instance, data from sources like Google Trends illustrate this phenomenon, showing that people tend to search more for the term "inflation" during high inflation periods (refer to Figure 1).

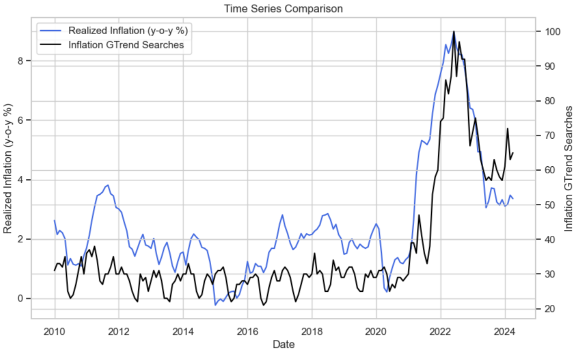
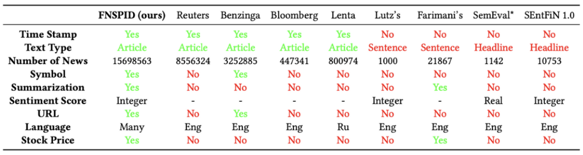


Figure 1: Number of google searches for ‘inflation’ over time

Recognizing the valuable insights embedded in human behavior and perceptions, we chose news articles as our primary data source. News are comprehensive and highly relevant, reflecting the opinions and signals that traders consider alongside mathematical models and prior information. Lastly, news is readily accessible, which is a significant advantage compared to the costly acquisition of social data, such as Tweets.

### 3.1 Data Sources

Among the targeted available datasets, we select the Financial News Sentiment and Price Information Dataset (FNSPID). The FNSPID dataset boasts a strong foundation due to its high information content (see figure 2). It contains over 15.7 million time-aligned financial news records spanning 1999 to 2023. Additionally, the dataset is notable as it includes preprocessed article summaries, which condense the text and make it more concise. This reduces the complexity for machine learning models, enabling faster and easier training compared to using long articles with intricate dependencies that are difficult to interpret.



| | FNSPID (ours) | Reuters | Benzinga | Bloomberg | Lenta | Lutz's | Farimani's | SemEval* | SEntFiN 1.0 |
|---|---|---|---|---|---|---|---|---|---|
| **Time Stamp** | Yes | Yes | Yes | Yes | Yes | No | No | No | No |
| **Text Type** | Article | Article | Article | Article | Article | Sentence | Sentence | Headline | Headline |
| **Number of News** | 15698563 | 8556324 | 3252885 | 447341 | 800974 | 1000 | 21867 | 1142 | 10753 |
| **Symbol** | Yes | No | Yes | No | No | No | No | No | No |
| **Summarization** | Yes | No | No | No | No | No | Yes | No | No |
| **Sentiment Score** | Integer | - | - | - | - | Integer | - | Real | Integer |
| **URL** | Yes | No | Yes | No | No | No | No | No | No |
| **Language** | Many | Eng | Eng | Eng | Ru | Eng | Eng | Eng | Eng |
| **Stock Price** | Yes | No | No | No | No | No | Yes | No | No |

Figure 2: Comparison of existing datasets for Time Series Financial Analysis

However, leveraging this valuable data source comes with several challenges. First, processing millions of news articles within our time and resources constraints seems unpractical, imposing down-sampling of the data. Then, this dataset lacks a critical element for training sentiment models: sentiment labels associated with each news article.

## 3.2 Data Preparation

To conduct a sentiment analysis in order to build an effective and expressive Sentiment index, a precise and **labeled** dataset is needed.

We address these challenges by outlining the preprocessing pipeline we apply to the dataset to obtain training, evaluation, and testing samples. We will also discuss the labeling solution we chose.

### Preprocessing

We restrict ourselves to a random sample of 3% (approximately 500,000 articles) from the FNSPID dataset to address storage and computational constraints. This sample spans the years 1999 to the present day. To facilitate tokenization and expedite model training, we further filter this sample by excluding articles with summaries exceeding 100 words. This results in a substantial dataset of general financial news.

Our next step involves isolating and selecting news specifically related to inflation. We employ a heuristic method based on lexicon selection. This means we only retain articles containing at least one word from our predefined inflation lexicon (which is based on fundamental components of this latest): *"Inflation", "Gasoline prices", "Food prices", "Deflation", "Consumer price index", "CPI", "Core CPI".*

This filtering process reduces the dataset from 500,000 articles to over 74,000 articles, while still maintaining a good representation.

Focusing on historical trends of inflation rates, we analyze news articles from 2010 to the present day. This specific timeframe is particularly interesting because it encompasses two distinct phases: a period of low inflation and a period of high inflation. To address the lack of sentiment labels in the dataset, we limit ourselves to a sample of 20 articles per month, creating a manageable dataset for manual labeling.

### Labeling

Labeling remains the cornerstone of our project. Inaccurate labels or labels that misrepresent the true sentiment of the news articles in our training set will significantly hinder our ability to develop a high-performing model with strong predictive power. While a simple sentiment classification (positive, negative, or neutral) might seem like a straightforward approach for our initial selection of 2,167 inflation-related news articles, it wouldn't provide the level of granularity we need.

To achieve a more informative labeling system, we have opted for the following classification scheme (details to follow):

- **1** if the article expresses inflation will go up.
- **-1** if the article expresses inflation will go down.
- **0** if the article sentiment about inflation is neutral.

While human labeling by financial and economic specialists would have undoubtedly been the ideal approach, budgetary and time constraints inherent to student projects removes this method from our options. However, the recent release of OpenAI's highly performant GPT-4 model has coincided with a surge in Large Language Model (LLM) usage for labeling tasks, evident on platforms like Hugging Face. Recognizing this trend, we have opted to leverage the GPT-4 Turbo API for labeling our news articles using a custom prompt (refer to Appendix A).

The labeling method proves its effectiveness. As shown in Figure 3, most articles are classified as

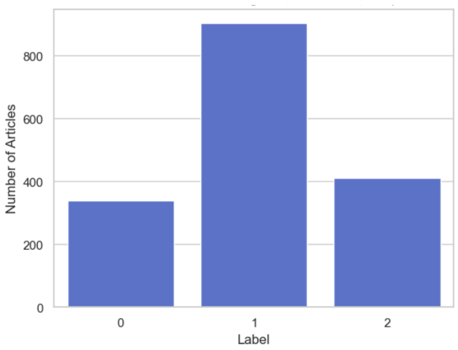


Figure 3: Sentiment Distribution: 0 is negative, 1 neutral, 2 postive

expressing a neutral sentiment towards inflation. This aligns with our expectations, as many articles likely mention inflation incidentally without explicitly conveying a positive or negative opinion.

However, despite this well-labeled dataset, a challenge persists. Cost constraints prevent further labeling with the expensive GPT-4. Additionally, the class imbalance (predominantly neutral sentiment) combined with the limited dataset size may hinder the effectiveness of fine-tuning a large language model (LLM). There's a high risk that the LLM won't have enough training data to learn the specific nuances of this task. To address the class imbalance, we opt to augment our training set. This process involves increasing the data points in underrepresented classes to achieve a more balanced distribution. We achieve this by leveraging an open-source pre-trained paraphraser, specifically the <*humarin/chatgpt_paraphraser_on_T5_base*> model (Vladimir Vorobev, 2023). This tool allows us to generate paraphrases of existing neutral-sentiment articles, effectively creating synthetic data points for the positive and negative sentiment classes.

Through this data augmentation technique, we expand our training set to a final size of 6,000 labeled news articles.

## 3.3 Model

We conduct a comparative analysis of two prominent machine learning paradigms to identify the most effective models for this sentiment classification task: traditional machine learning models and deep learning models, with a particular focus on transformers, a powerful deep learning architecture.

### Traditional Machine Learning Techniques

For each model type, we evaluate two variants: one trained from scratch using Continuous Bag-of-Words (CBOW) or word2vec word embeddings, and another leveraging pre-trained GloVe embeddings (glove.6B.100d).

- **Logistic Regression:** A foundational statistical method widely used for binary classification tasks. While relatively simple, it can achieve strong performance when paired with effective optimization algorithms and regularization techniques (like L1 or L2 regularization).

- **Random Forest:** An ensemble learning method that combines predictions from multiple decision trees. This approach offers several advantages, including interpretability and robustness to high-dimensional and complex data.

- **Support Vector Machine (SVM):** A classic machine learning technique that excels at finding a clear separation between classes in high-dimensional spaces. It can be particularly effective when combined with non-linear kernel functions to handle complex data patterns. However, SVMs can be computationally expensive for very large datasets.

- **XGBoost:** An advanced ensemble learning method based on gradient boosting. It trains multiple decision trees sequentially, where each tree corrects the errors of the previous one. This approach leads to highly accurate models, but XGBoost can be prone to overfitting if not carefully tuned (Chen and Guestrin, 2016).

### Deep Learning Techniques (Transformers Models)

- **BERT-base:** BERT base version containing 100 million parameters (Devlin et al., 2018). Its pre-training data encompasses the concatenation of the Toronto Book Corpus and English Wikipedia. As all others Transformers models BERT use self-attention layer enabling it to encode and align to long term dependencies in the input(Vaswani et al., 2023).

- **FinBERT:** A pre-trained NLP model to analyze sentiment of financial text (Araci, 2019).

It is built by further training the BERT language model in the finance domain, using a large financial corpus and thereby fine-tuning it for financial sentiment classification. Financial PhraseBank is used for fine-tuning(Malo et al., 2014).

- **Fin-distill-RoBERTa:** A distilled version of the RoBERTa-base model (Sanh et al., 2019). It follows the same training procedure as DistilBERT. The model has 6 layers, 768 dimension and 12 heads, totalizing 82M parameters. It was also fine-tuned on the on the Financial Phrasebank dataset(Malo et al., 2014).

### 3.4 Experiment

For the selection and training procedure, we first split our dataset into training and testing sets based on timeframe rather than a fixed ratio. The training split contains news articles from January 1st, 2010 to December 31st, 2019 (low inflation period), and the test split contains articles from January 1st, 2020 to December 31st, 2023 (high inflation period). As mentioned earlier, out-of-sample testing is crucial for our case.

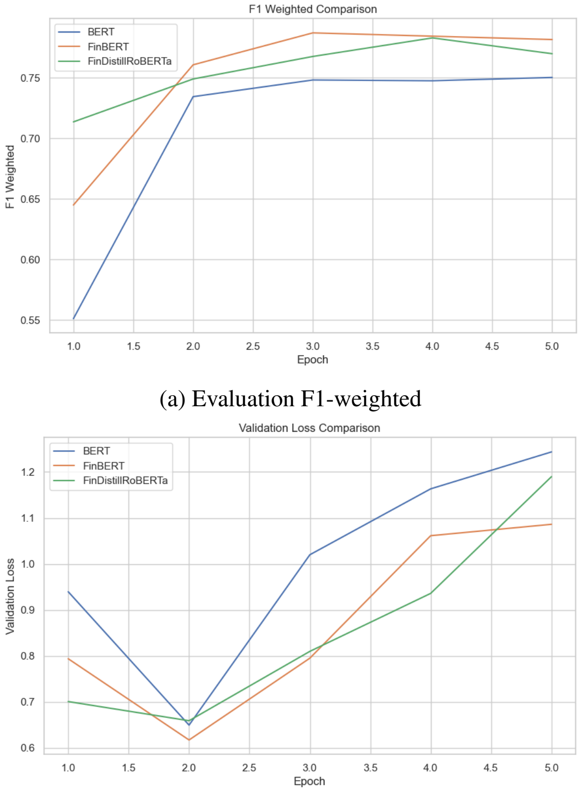
We begin by selecting the best performing configuration from the traditional machine learning models (refer to Appendix B to see all the hyperparameters values tested) . We achieve this through a grid search with cross-validation using a 3-fold k-fold procedure. A train-evaluation split with a 0.1 ratio is applied to the initial training set.

Next, using the same split, we fine-tune our transformer models to assess and improve their effectiveness in sentiment analysis tasks. We employ the transformers library (**?**) to fine-tune the three large language models (LLMs) for our specific task. This process involves conducting 5 training epochs. The choice of five epochs is a balance between achieving realistic training within a reasonable timeframe and considering resource constraints. Given the relatively small dataset size, employing more epochs could lead to overfitting. For this benchmarking step, we utilize the AdamW optimizer with a fixed learning rate of 2e-5 and a batch size of 16.

### 3.5 Results

The benchmarking of the traditional machine learning models highlights the superiority to XGBoost in a similar and rigid test setup (see Table 5).

For the transformer models, we identify the best epoch checkpoints based on the evaluation loss during the fine-tuning process (see Figure 4). Subsequently, we use these checkpoint models on the test set alongside the two best XGBoost configurations to determine the overall best performing model (see results Table 1).



(a) Evaluation F1-weighted

(b) Evaluation loss

Figure 4: Training Analytics: Transformers Models

As mentioned earlier, XGBoost is powerful, but both versions exhibit overfitting on strongly positive and negative sentiments (see Figure X for details, where applicable). However, in terms of overall performance on detecting all three sentiment classes, the best performing model is Fin-distill-RoBERTa, achieving a strong F1-weighted score of **0.570034** on the out-of-sample test set.

This new high-performing fine-tuned sentiment analysis model, named InflaBERT, will be the model used in our research. InflaBERT can be found on our Hugging Face page: https://huggingface.co/MAPAi/InflaBERT.

## 4 Inflation Nowcaster

### 4.1 NEWS Index

Using `InflaBERT` (derived in section 3), we construct an index replicating the news sentiment regarding inflation.

| Model | F1 Label 0 | F1 Label 1 | F1 Label 2 | F1 Weighted |
|---|---|---|---|---|
| Fin-distill-RoBERTa | **0.486842** | 0.527132 | **0.677419** | **0.570034** |
| FinBERT | 0.378378 | 0.522059 | 0.586667 | 0.516634 |
| BERT-case | 0.416000 | **0.560261** | 0.625000 | 0.554771 |
| XGBoost W2V | 0.449869 | 0.111111 | 0.645598 | 0.451693 |
| XGBoost GloVe | 0.487519 | 0.139130 | 0.615819 | 0.487519 |

Table 1: Table : F1 Scores on the Test Set (Out Sample)

To accomplish this, we first need to select a score function that transforms our sentiment probabilities into individual scores, which we can then aggregate for each news item within a period. Initially, we consider the argmax score – $score(n) = \text{argmax}\{\text{InflaBERT}(n)\}$ – which is a straightforward and intuitive method that assigns scores based on the most probable label. However, this method is highly sharp (see figure 5a) and does not account for the model's uncertainty. Therefore, we also consider an alternative, the polarity score:

$$score(n) = \mathbb{E}[\text{InflaBERT}(n)]$$

Where $\mathbb{E}[\text{InflaBERT}(n)]$ is the model's expectation regarding a news $n$ vis-à-vis its labels $l$, i.e:

$$\mathbb{E}[\text{InflaBERT}(n)] = \sum_{l} l \cdot p_l^{\text{InflaBERT}}(n)$$

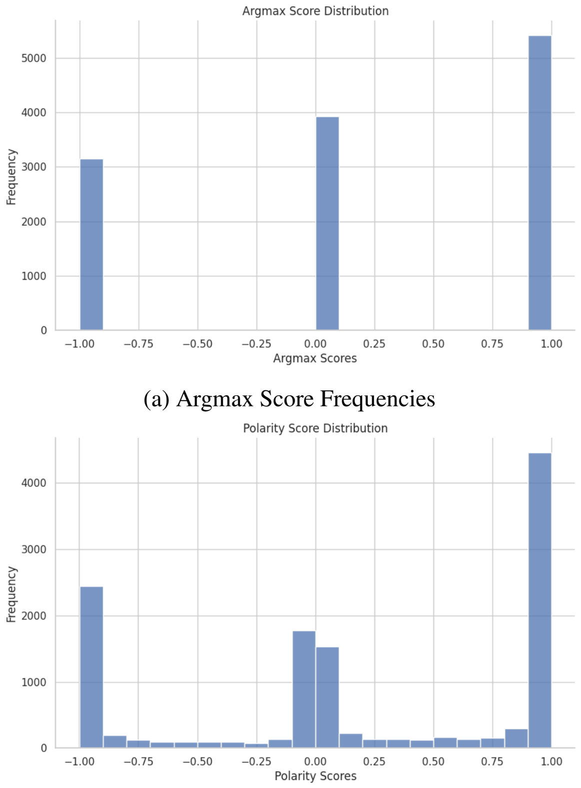
The polarity score has a significant advantage as it accounts for the model's uncertainty. When `InflaBERT` is uncertain between labels, it produces a weighted average of the labels based on their respective certainties. This results in a smoother scoring function (see Figure 5b).

Then, we apply our scoring function to each news of our dataset and aggregate them monthly using cumulative expectation:

$$P_t^{\text{NEWS}} = \sum_{i=0}^{t} \mathbb{E}_{N_i}[score(N_i)]$$

In practice, we don't have access to the exact monthly news inflation sentiment – $\mathbb{E}_{N_i}[polarity(N_i)]$ –, so we estimate it using the sample mean, that we know from the strong Law of Large Numbers (Chung, 2008) converge asymptotically to the population monthly news inflation sentiment:

$$\frac{1}{|N_i|} \sum_{n \in N_i} score(n)$$



(a) Argmax Score Frequencies

(b) Polarity Score Frequencies

Figure 5: Different Scores Frequencies

This leads to `NEWS`, an index replicating the news sentiment regarding inflation, as shown in figure 6.

Even though the `NEWS` index captures the high inflation period during COVID-19 quite well, it appears overly aggressive (predicting periods of high deflation and high inflation) during stable periods of realized inflation. This indicates that further work is needed to stabilize the index, such as smoothing the scoring function even more, training on more data, or using a more complex model.

### 4.2 NEWS For Inflation Nowcasting

We aim to enhance state-of-the-art inflation nowcasters by incorporating our `NEWS` index. To achieve this, we modify the Cleveland Federal Re-

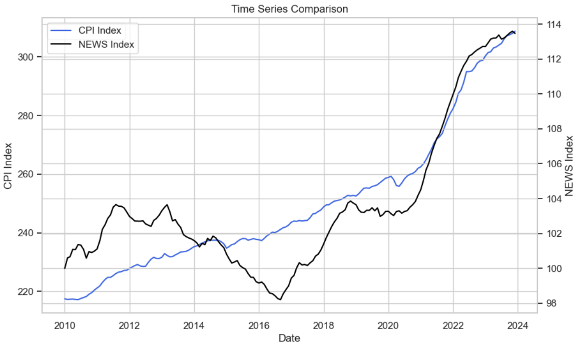


Figure 6: The NEWS index over the 2010-2023 period

serve Bank model (Knotek and Zaman, 2017) by adding a term that reflects the influence of our NEWS index.

The Cleveland model is an OLS regression that relates the percentage changes in main inflation components – Core CPI, Food CPI, and seasonally adjusted Gasoline CPI – to overall inflation. Specifically, let $\pi_t^{\text{i}} = 100 * (P_t^{\text{i}}/P_{t-12}^{\text{i}} - 1)$ represent the percentage change of index $i$, the Cleveland model is then defined as follows:

$$\pi_t^{\text{CPI,FED}} = \beta_0 + \beta_1\pi_t^{\text{Core CPI}} + \beta_2\pi_t^{\text{Food CPI}} + \beta_3\pi_t^{\text{Gasoline}} + e_t \quad (1)$$

Adding the NEWS index to the Cleveland model described in (1), we obtain our new model (see model (2)).

$$\pi_t^{\text{CPI,FED+NEWS}} = \pi_t^{\text{CPI,FED}} + \beta_4\pi_t^{\text{NEWS}} + e'_t \quad (2)$$

### 4.2.1 Experiment

To assess the enhancement our novel index can provide, we compare our two models from COVID-19 (January 2020 to December 2023). This analysis help to determine how effectively the new index could have tracked the movements in inflation during that highly volatile time. Notably, `InflaBERT` is out-of-sample for this period, ensuring that our news sentiment index remains unbiased by previously observed news.

The models are trained on data from January 2015 to December 2019, covering the 60 preceding periods as suggested by the baseline authors (Knotek and Zaman, 2017).

To compute the prediction for a given month, we need the values of all the indices at that time. However, the values for Core CPI, Food CPI, and Gasoline for month $t$ are released only around the 15th day of the following month $(t+1)$. Therefore, we replace the exact values in the nowcast with a prediction obtained by computing the moving average of the indices over the past year (12 periods), i.e: $\hat{\pi}_t^{\text{i}} = \frac{1}{12}\sum_{k=1}^{12}\pi_{t-k}^{\text{i}}$. The prediction of inflation at a month $t$ can thus be computed around the 15th day of the month $t$ and is formulated as follows:

$$\hat{\pi}_t^{\text{CPI}} = \beta_0 + \beta_1\hat{\pi}_t^{\text{Core CPI}} + \beta_2\hat{\pi}_t^{\text{Food CPI}} + \beta_3\hat{\pi}_t^{\text{Gasoline}}$$

For the model that includes the NEWS index, assuming we have access to the news from the first 15 days of month $t$, we can compute the exact value of $\pi_t^{\text{NEWS}}$ and use it to nowcast.

We evaluate our nowcaster models on annualized month-over-month inflation predictions (3) using the methodology proposed by Knotek (Knotek II and Zaman, 2023), employing Root Mean Square Error (RMSE) and the Giacomini-White test (Giacomini and White, 2006) to quantify the significance of our improvements.

$$\hat{\pi}_t^{\text{CPI,Annualized}} = 100 * [(\frac{\hat{\pi}_t^{\text{CPI}}}{100} + 1)^{12} - 1] \quad (3)$$

### 4.2.2 Results

Table 2 presents the OLS regression results, indicating that the coefficient related to the NEWS index is statistically significant at the 10% level.

Figure 7 displays the time series of the annualized models' forecasts. Adding the NEWS index seems to slightly enchance the performance during high-time inflation. Table 3 shows the RMSE and the significance of these forecasts compared to realized inflation. We observe a slight improvement in the forecasts (a reduction of 0.02% in RMSE), although this improvement is not statistically significant.

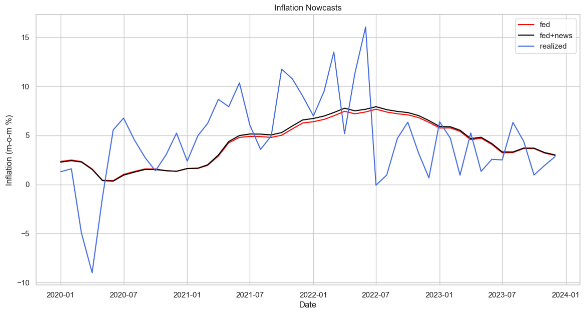


Figure 7: Annualized Inflation Nowcasts

| | *Dependent variable: CPI* | |
|---|---|---|
| | fed | fed+news |
| | (1) | (2) |
| const | 0.021 | 0.017 |
| | (0.026) | (0.025) |
| pi-CCPI | 0.616$^{***}$ | 0.634$^{***}$ |
| | (0.135) | (0.133) |
| pi-FCPI | 0.186$^{***}$ | 0.176$^{***}$ |
| | (0.064) | (0.063) |
| pi-Gasoline | 0.035$^{***}$ | 0.034$^{***}$ |
| | (0.002) | (0.002) |
| pi-NEWS | | 0.149$^{*}$ |
| | | (0.081) |
| Observations | 60 | 60 |
| $R^2$ | 0.892 | 0.898 |
| F Statistic | 153.612$^{***}$ | 120.877$^{***}$ |
| *Note:* | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 | |

Table 2: Regression Results

| | **RMSE** |
|---|---|
| FED | 0.0409 |
| | (–) |
| FED+NEWS | 0.0407 |
| | (0.19) |
| *Note:* | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 |

Table 3: Inflation Forecasts of the Models. In parenthesis are the p-values of the Giacomini-White test.

## 5 Conclusion

Overall, the ongoing development of AI and ML technologies holds significant promise for more accurate and efficient economic nowcasting, thereby aiding central banks and policymakers in their critical decision-making processes. In particular, the direct use of LLMs demonstrates promising advancements in economic forecasting methodologies (as shown by Faria-e Castro (Faria-e Castro and Leibovici, 2023) and Bybee (Bybee, 2023)). This work, through the development of InflaBERT—a BERT-based model fine-tuned for predicting inflation-related news sentiment—has shown that leveraging advanced machine learning techniques can enhance traditional models like the Cleveland Fed's nowcasting model. The creation of the NEWS index from InflaBERT's sentiment analysis offered a novel dimension to nowcasting, particularly during the volatile economic period of COVID-19.

The results indicated that the inclusion of the NEWS index provided a marginal improvement in forecast accuracy, as measured by a slight reduction in RMSE. Although the improvement was not statistically significant, it suggests potential for further refinement. This indicates that machine learning and sentiment analysis can complement traditional economic models, providing a more nuanced understanding of real-time inflation dynamics.

Future research could aim to address the observed limitations by improving the accuracy of the NEWS sentiment index through a larger labeled dataset and utilizing more advanced models, such as the latest open-source generative models (like the recent LLaMa-3 family (AI@Meta, 2024)). Additionionally, exploring non-linear nowcasters, such as LSTM models which have demonstrated stronger performance in similar tasks (Siami-Namini et al., 2018; Cao et al., 2019), could better capture the complexities of economic indicators.

## References

AI@Meta. 2024. Llama 3 model card.

Cristina Angelico, Juri Marcucci, Marcello Miccoli, and Filippo Quarta. 2022. Can we measure inflation expectations using twitter? *Journal of Econometrics*, 228(2):259–277.

Dogu Araci. 2019. Finbert: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*.

Marta Bańbura, Domenico Giannone, Michele Modugno, and Lucrezia Reichlin. 2013. Now-casting and the real-time data flow. In *Handbook of economic forecasting*, volume 2, pages 195–237. Elsevier.

Günter W Beck, Kai Carstensen, Jan-Oliver Menz, Richard Schnorrenberger, and Elisabeth Wieland. 2024. Nowcasting consumer price inflation using high-frequency scanner data: Evidence from germany.

Leland Bybee. 2023. Surveying generative ai's economic expectations. *arXiv preprint arXiv:2305.02823*.

Jian Cao, Zhi Li, and Jian Li. 2019. Financial time series forecasting model based on ceemdan and lstm. *Physica A: Statistical mechanics and its applications*, 519:127–139.

Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16. ACM.

KL Chung. 2008. The strong law of large numbers. *Selected Works of Kai Lai Chung*, pages 145–156.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Zihan Dong, Xinyu Fan, and Zhiyuan Peng. 2024. Fnspid: A comprehensive financial news dataset in time series. *arXiv preprint arXiv:2402.06698*.

Miguel Faria-e Castro and Fernando Leibovici. 2023. Artificial intelligence and inflation forecasts. Technical report.

Raffaella Giacomini and Halbert White. 2006. Tests of conditional predictive ability. *Econometrica*, 74(6):1545–1578.

Edward S Knotek and Saeed Zaman. 2017. Nowcasting us headline and core inflation. *Journal of Money, Credit and Banking*, 49(5):931–968.

Edward S Knotek and Saeed Zaman. 2024. Nowcasting inflation.

Edward S Knotek II and Saeed Zaman. 2023. A real-time assessment of inflation nowcasting at the cleveland fed. *Economic Commentary*, (2023-06).

P. Malo, A. Sinha, P. Korhonen, J. Wallenius, and P. Takala. 2014. Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology*, 65.

De Bandt Olivier, Bricongne Jean-Charles, Denes Julien, Dhenin Alexandre, De Gaye Annabelle, and Robert Pierre-Antoine. 2023. Using the Press to Construct a New Indicator of Inflation Perceptions in France. Technical report.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Sima Siami-Namini, Neda Tavakoli, and Akbar Siami Namin. 2018. A comparison of arima and lstm in forecasting time series. In *2018 17th IEEE international conference on machine learning and applications (ICMLA)*, pages 1394–1401. IEEE.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need.

Maxim Kuznetsov Vladimir Vorobev. 2023. A paraphrasing model based on chatgpt paraphrases.

# Appendix

## A GPT Labeling Prompting Strategy

This section presents the prompting strategy used to query GPT-4 Turbo API when labeling the dataset.

> Custom Prompting Strategy
>
> Here is a financial news article related to inflation. Please read the article and determine whether it expresses a sentiment that inflation will go up, go down, or if it makes no clear expression. Use any financial knowledge you have.
>
> Respond with a **SINGLE** number (no explanation needed):
>
> - **1** if the article expresses inflation will go up.
> - **-1** if the article expresses inflation will go down.
> - **0** if the article sentiment about inflation is neutral.
>
> Here's the article:
> {*ARTICLE*}

## B Models Cross-Validation

In this section, we discuss the cross-validation results employed to ascertain the most effective model for analyzing sentiment concerning inflation in news articles. Table 4 displays the values of the hyperparameters tested, while Table 5 showcases the test F1-Scores from the best cross-validation setups.

## C Full Regression Results

This section presents additional models result. The models are a combination of the existing Cleveland model regressors with the NEWS index. Table 6 present the models regression results.

| **Logistic Regression** | |
|---|---|
| *Hyperparameters* | *Values* |
| penalty | elasticnet |
| C | 0.01, 0.1, 0.2, 0.5, 1, 5 |
| l1_ratio | 0.0, 0.1, 0.5, 0.7, 1.0 |
| **Random Forest** | |
| *Hyperparameters* | *Values* |
| n_estimators | 10, 50, 100, 200 |
| max_depth | None, 10, 20, 30 |
| **SVM** | |
| *Hyperparameters* | *Values* |
| C | 0.1, 1, 10 |
| kernel | linear, poly, rbf, sigmoid |
| **XGBOOST** | |
| *Hyperparameters* | *Values* |
| max_depth | 5, 10, 100 |
| learning_rate | 0.01, 0.1, 0.2 |
| n_estimators | 20, 100, 200 |

Table 4: Hyperparameters tested.

| **Model x Embedding** | **Hyperparameters** | **F1-Score (Weighted)** |
|---|---|---|
| Logistic Regression + **W2V** | {**'C'**=5, **'l1_ratio'**=1.0} | 0.6051 |
| Random Forest + **W2V** | {**'max_depth'**:20, **'n_estimators'**:200} | 0.7522 |
| SVM + **W2V** | {**'C'**:10, **'kernel'**: 'rbf'} | 0.7571 |
| XGBoost + **W2V** | {**'learning_rate'**:0.2, **'max_depth'**:100, **'n_estimators'**: 200} | **0.7912** |
| Logistic Regression + **GloVe** | {**'C'**:5, **'l1_ratio'**: 0.5,**'penalty'**: 'elasticnet'} | 0.5670 |
| Random Forest + **GloVe** | {**'max_depth'**:None, **'n_estimators'**: 200} | 0.6876 |
| SVM + **GloVe** | {**'C'**:10, **'kernel'**: 'poly'} | 0.5993 |
| XGBoost + **GloVe** | {**'learning_rate'**:0.1, **'max_depth'**:10, **'n_estimators'**: 200} | **0.7347** |

Table 5: Comparison of different models and embedding techniques using GridSearch with 3-fold Cross Validation

| | *Dependent variable: CPI* | | | | |
|---|---|---|---|---|---|
| | fed | news | fed+news | fed-gas+news | ccpi+news |
| | (1) | (2) | (3) | (4) | (5) |
| const | 0.021 | $0.144^{***}$ | 0.017 | 0.012 | 0.003 |
| | (0.026) | (0.024) | (0.025) | (0.073) | (0.071) |
| pi-CCPI | $0.616^{***}$ | | $0.634^{***}$ | $0.814^{**}$ | $0.810^{**}$ |
| | (0.135) | | (0.133) | (0.383) | (0.381) |
| pi-FCPI | $0.186^{***}$ | | $0.176^{***}$ | -0.105 | |
| | (0.064) | | (0.063) | (0.177) | |
| pi-Gasoline | $0.035^{***}$ | | $0.034^{***}$ | | |
| | (0.002) | | (0.002) | | |
| pi-NEWS | | $0.419^{*}$ | $0.149^{*}$ | $0.456^{*}$ | $0.448^{*}$ |
| | | (0.237) | (0.081) | (0.232) | (0.230) |
| Observations | 60 | 60 | 60 | 60 | 60 |
| $R^2$ | 0.892 | 0.051 | 0.898 | 0.126 | 0.121 |
| Adjusted $R^2$ | 0.886 | 0.035 | 0.890 | 0.080 | 0.090 |
| Residual Std. Error | 0.064 | 0.185 | 0.062 | 0.181 | 0.180 |
| F Statistic | $153.612^{***}$ | $3.131^{*}$ | $120.877^{***}$ | $2.702^{*}$ | $3.921^{**}$ |
| *Note:* | | | | | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 |

Table 6: Regression Results