News Sentiment-based Inflation Indicator

Nandeesh Bhatrai[a], Karmanya Gupta[a], Sunil Paul[b] and Clint P. George[a,c]

[a] School of Mathematics and Computer Science, Indian Institute of Technology Goa, Farmagudi, 403401, Goa, India
[b] School of Humanities and Social Sciences, Indian Institute of Technology Goa, Farmagudi, 403401, Goa, India
[c] School of Interdisciplinary Life Sciences, Indian Institute of Technology Goa, Farmagudi, 403401, Goa, India

ARTICLE INFO

Keywords:
Inflation Sentiment Index
News Analytics
Natural Language Processing
BM25
TF–IDF
Sentence Embeddings
Large Language Models
RoBERTa
Cross-Encoder Models
Cosine Similarity
RMSE Evaluation
Economic Forecasting

ABSTRACT

This study investigates whether news-based sentiment can serve as a reliable proxy for inflation dynamics in India. Using a decade of news articles (2014-2024), we construct multiple inflation sentiment indices spanning traditional lexical methods, frequency-based document scoring, semantic embedding similarity, and large language model (LLM)-based contextual scoring. The methodology evaluates TF-IDF, BM25 and BM25+ relevance metrics, their polarity-corrected negated variants, the Loughran-McDonald financial lexicon, and VADER sentiment analysis. To incorporate semantic context, we use sentence-embedding similarity between articles and inflation-related query prompts, applying threshold-based filtering to exclude semantically irrelevant text. Furthermore, we compute sentiment using two families of LLM models: a RoBERTa Twitter sentiment classifier (with both weighted and unweighted variants), and a cross-encoder LLM that jointly scores article-query pairs. For each method, we generate a monthly sentiment index and compare it with CPI inflation using RMSE and visual time-series alignment. Results show a clear performance hierarchy: traditional lexical methods capture broad fluctuations but exhibit substantial noise; embedding-based filtering offers smoother and more relevant indices; Twitter-trained LLMs perform poorly due to domain mismatch; and cross-encoder LLM scoring achieves the lowest RMSE and strongest temporal correspondence with CPI. The findings underscore the significance of comprehensive contextual modeling and semantic filtering in deriving economic indicators from textual data.

1. Introduction and Overview

Inflation forecasting traditionally relies on macroeconomic variables, but recent work suggests that textual information-such as news coverage-contains forward-looking signals that are not fully captured by numerical indicators. Economic news reflects market expectations, policy shifts, supply-side disruptions, and consumer sentiment, making it a potentially powerful supplementary source for inflation monitoring. However, extracting meaningful inflation-related sentiment from large text corpora remains a methodological challenge. Lexical approaches overlook context, frequency-based scoring can misinterpret polarity, and general-purpose sentiment models often fail to capture domain-specific signals in economic language.

This work systematically evaluates a wide spectrum of text-analysis techniques to construct news-based inflation sentiment indices for India from 2014 to 2024. We begin with traditional methods, including TF–IDF, BM25, BM25+, their negated variants, and financial lexicons, to establish a baseline performance. Next, we incorporate semantic understanding through sentence-embedding similarity, using cosine thresholds to filter articles that are contextually related to inflation. Building upon this, we integrate large language models in two forms: (i) a RoBERTa Twitter sentiment classifier with weighted and unweighted aggregation schemes, and (ii) a cross-encoder LLM that jointly processes article–query pairs to produce fine-grained contextual sentiment scores. Each method yields a monthly sentiment index, which we compare against the normalized CPI series using RMSE and time-series plots to assess trend alignment, turning-point detection, and overall dynamic correspondence.

By evaluating models that range from lexical frequency scoring to deep contextual LLMs, this study provides a comprehensive analysis of how different levels of linguistic representation influence the construction of economic sentiment indicators. The results demonstrate that deeper semantic modeling-especially cross-encoder LLMs-offers substantial advantages over traditional techniques, suggesting a promising path forward for integrating textual data into macroeconomic monitoring frameworks.

ORCID(s): 0000-0003-3630-9811 (C.P. George)

2. Literature Review and Related Work

A growing body of research investigates how textual data can be used to measure sentiment and improve macroeconomic forecasting. Early work demonstrates that sentiment extracted from news can predict key economic outcomes, including consumer sentiment, output, and financial market behavior. Shapiro, Sudhof and Wilson (2022) develops domain-specific lexicons and machine-learning-enhanced sentiment models tailored to economic news, showing substantial predictive gains over off-the-shelf methods and highlighting the importance of domain alignment in sentiment scoring. Similarly, Seki, Ikuta and Matsubayashi (2022) constructs a business sentiment index using deep learning on newspaper text and finds strong correlations with survey-based indicators, reinforcing the value of large-scale textual signals for economic monitoring.

Recent work directly explores the relationship between news sentiment and inflation. Eugster and Uhl (2024) shows that sentiment derived from nearly 730,000 news articles significantly improves U.S. inflation forecasts across multiple horizons, outperforming standard benchmarks such as random walk and autoregressive models. In the Indian context, Pratap and Ranjan (2021) constructs commodity-specific sentiment indices from news coverage of tomato, onion, and potato (TOP) prices and demonstrates that incorporating these indicators improves short-term food inflation forecasting relative to models using only price data or weather variables. More recent studies integrate large language models (LLMs) into similar frameworks. Allard, Teiletche and Zinebi (2024) proposes InflaBERT, a BERT-based model fine-tuned on inflation-related news to generate a monthly sentiment index (NEWS). They show that incorporating this index marginally improves the Cleveland Fed's inflation nowcasting model, particularly during high-volatility periods such as COVID-19, highlighting the promise and practical challenges of deploying LLM-based sentiment systems for real-time macroeconomic monitoring

The sentiment analysis techniques used in these studies vary widely, ranging from lexicon-based models, such as VADER, which incorporates rules for intensifiers and negation (Hutto and Gilbert (2015)), to bag-of-words frequency models, embeddings, and transformer-based large language models. Collectively, the literature indicates that richer linguistic representations tend to yield more accurate economic signals, particularly when sentiment is closely tied to economic concepts (e.g., inflation, policy uncertainty, supply shocks).

Our work contributes to this literature by systematically comparing a broad spectrum of sentiment extraction techniques, TF-IDF, BM25, domain lexicons, VADER, embedding similarity, RoBERTa-base Twitter LLM, and cross-encoder LLM, in constructing a monthly inflation sentiment index for India using a decade of news articles. Unlike prior Indian studies that focused on food inflation or commodity-specific shocks, we aim to develop a general inflation sentiment index and benchmark its alignment with CPI dynamics across various methods. This situates our work at the intersection of textual sentiment modeling and macroeconomic indicator construction, extending both methodological breadth and empirical scope. In doing so, our study complements recent work on LLM-driven inflation sentiment and provides one of the first comprehensive evaluations of classic NLP methods versus modern transformer-based scoring for constructing inflation sentiment indicators.

3. Methodology

In this report, we describe the methodology used to construct a financial sentiment-based economic index by analyzing large-scale financial news data from The Economic Times. The overall process involves (i) scraping and constructing a corpus of financial articles, (ii) preprocessing and filtering textual data, and (iii) computing sentiment scores through multiple approaches, including lexicon-based, statistical techniques, and LLM sentiment extraction. The methodology is organized into several subparts, detailed below.

3.1. Data Acquisition

We construct a comprehensive corpus of financial news articles by programmatically scraping data from publicly available online archives. Our initial sources include "The Economic Times", "Financial Express", and "The Times of India". Using Python-based web scraping scripts, we send HTTP requests to article URLs and retrieve the corresponding HTML content. We then parse this content to extract relevant fields such as the article title, publication date, and main body text, while removing advertisements, navigation elements, and other non-informative components.

To ensure that the dataset remains focused on inflation-related content, we create a bag of approximately 250 inflation-associated keywords, including terms such as "inflation", "deflation", "economy", "GDP", "imports",

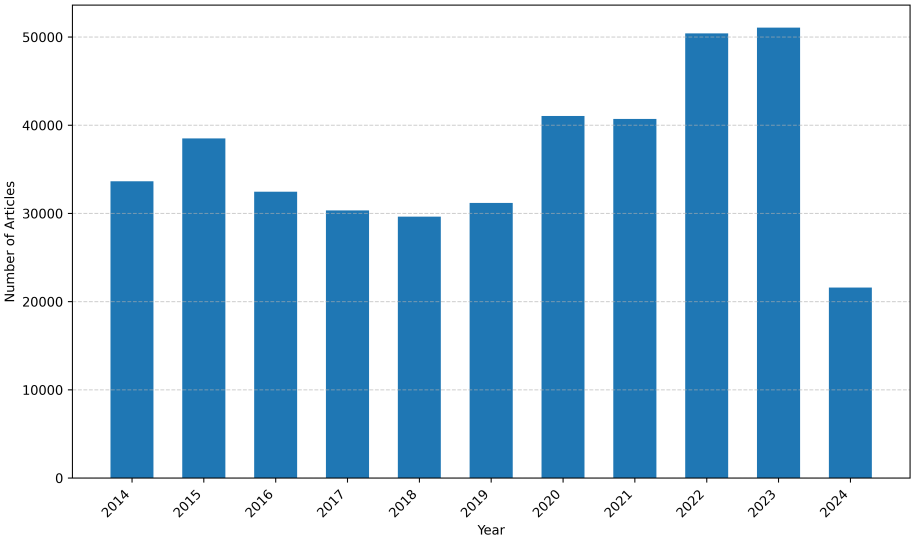

Figure 1: Timed trend of number of articles (Economic Times)

"exports", and "markets". We retain only those articles whose headlines contain at least one of these keywords, thereby filtering out unrelated financial or general news.

After experimenting with multiple sources, we focus exclusively on The Economic Times, as it provides a consistent and extensive archive that aligns with our target ten-year period from 2014 to 2024. This selection ensures both historical continuity and topical relevance across different phases of India's economic cycle. The final dataset contains approximately 400,000 high-quality financial news articles, each labeled with metadata including publication year, month, title, and cleaned article text.

3.2. Data Pre-processing

The raw articles contain several noisy elements such as advertisements, author bylines, watermarks, and other extraneous HTML artifacts. To ensure textual uniformity and relevance, we implement a multi-step preprocessing pipeline. We first remove all HTML tags, JavaScript fragments, and watermarks from the articles, and then discard empty or incomplete entries. We clean the remaining text by standardizing white spaces, correcting encoding inconsistencies, and stripping non-alphanumeric characters except for essential punctuation.

Next, we tokenize each article, dividing the text into smaller units called tokens, and convert all tokens to lowercase for consistency. We remove common stop words such as "the", "is", "of", "in", and "and" to reduce noise. Finally, we apply lemmatization using the NLTK library to reduce words to their root forms, ensuring that different inflections of a word are treated uniformly.

The resulting corpus contains clean and normalized text suitable for downstream sentiment analysis. Each record consists of five primary fields: Year, Month, Day, Title, and Text, representing the temporal and textual components of the dataset.

3.3. Task A - Cosine Similarity Using TF-IDF

The TF-IDF (Term Frequency-Inverse Document Frequency) model is a foundational technique in quantitative text analysis that weighs the importance of words within a document relative to a corpus. By operating within a Vector Space Model (VSM), TF-IDF transforms textual data into numerical vectors, enabling the calculation of Cosine Similarity to quantify the thematic alignment between news articles and a target concept.

We employ the TF-IDF approach to overcome the limitations of simple keyword counting. In raw frequency counts, common words can dominate the signal, potentially obscuring the specific thematic content of an article. TF-IDF mitigates this by penalizing terms that appear too frequently across the entire corpus while boosting the weight of terms that are discriminative to specific documents. Furthermore, using Cosine Similarity rather than Euclidean distance

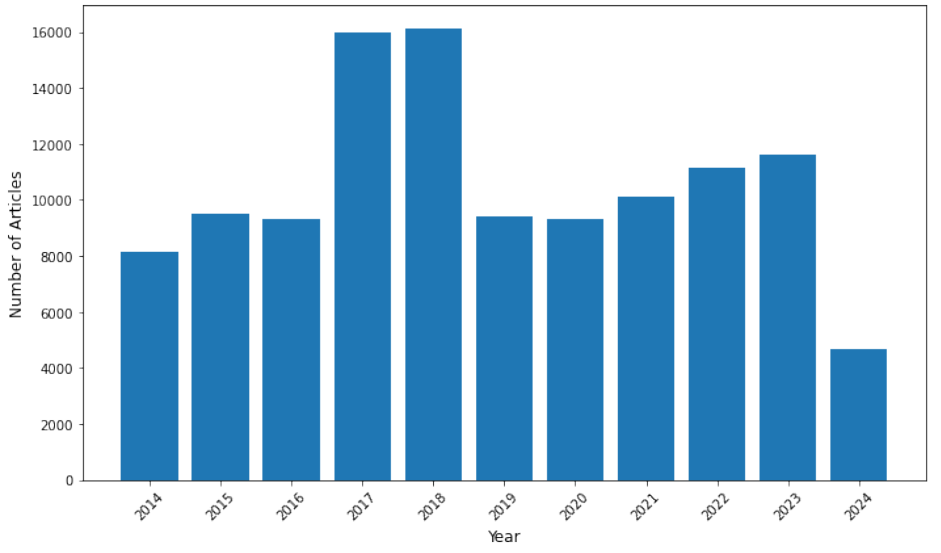

Figure 2: Timed trend of number of articles (TOI)

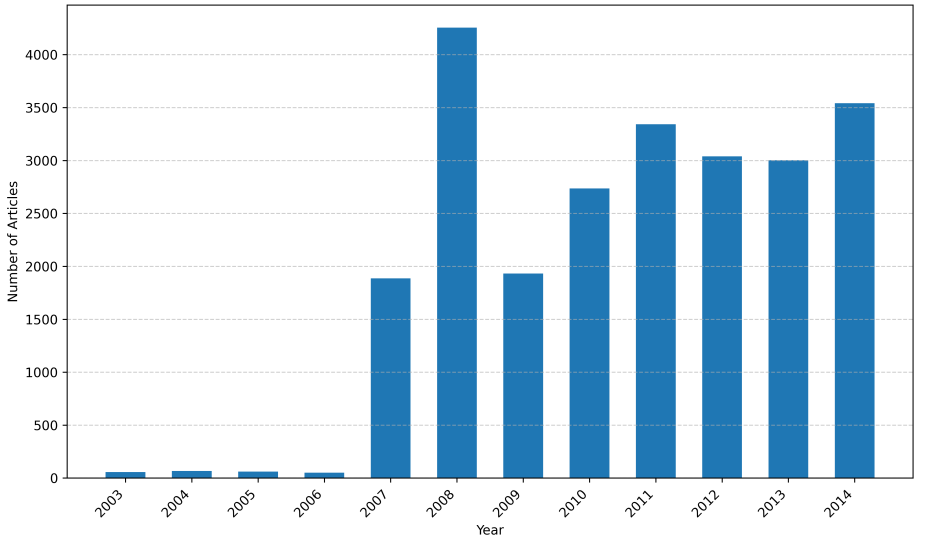

Figure 3: Timed trend of number of articles (Financial Express)

ensures that the relevance score is based on the orientation of the vector (the thematic angle) rather than its magnitude, making the metric robust to variations in article length.

Our methodology proceeds in several stages. First, we define a specialized vocabulary comprising inflation-related terms (e.g., "price," "rates," "economy"). We preprocess the news corpus using the NLTK library to tokenize text and apply the WordNet Lemmatizer, ensuring that variations of words map to their root forms. Using the TfidfVectorizer, we construct a TF-IDF vector for each article, restricted specifically to this vocabulary space.

To assess relevance, we construct a reference "inflation query" vector consisting of ones, representing an ideal document where all inflation keywords are present with equal weight. We then compute the Cosine Similarity between each article's TF-IDF vector and this reference query vector. This yields a similarity score for every article, indicating its semantic proximity to the concept of inflation. Finally, we aggregate these scores to a monthly frequency and z-normalize the resulting series to produce the final TF-IDF-based inflation sentiment index.

bank	banks	billion	budget	cent	commodity
consumption	cost	crore	credit	crude	demand
deficit	deposits	diesel	economy	economic	energy
exports	fiscal	food	fund	funds	futures
gdp	gst	growth	import	imports	index
inflation	interest	investors	loan	loans	market
million	monetary	oil	percent	petrol	price
prices	production	rate	rates	retail	rice
rbi	rupee	sebi	steel	stocks	subsidy
sugar	supply	tax	tonne	utilitie	wheat

Table 1
Keyword list used as query in BM25 based similarity scoring.

3.4. Task B - Best Matching 25 (BM25) Based Similarity Scoring

BM25 is a widely used ranking algorithm in information retrieval systems, designed to estimate the relevance of a document to a given query. It operates under a bag-of-words framework and evaluates relevance based on the presence and frequency of query terms, independent of their position within the text. BM25 improves upon TF-IDF by modelling term-frequency saturation and adjusting for document-length differences, allowing it to capture relevance more effectively in collections where articles vary widely in length and keyword emphasis.

We utilize the BM25 family of retrieval functions to assess the degree to which each news article aligns with the concept of inflation. Using the Python library rank-bm25, we implement three variants: Okapi BM25, BM25L, and BM25+ as described in Trotman, Puurula and Burgess (2014). These variants address the known limitations of the original model: Okapi BM25 tends to over-penalize longer documents due to its document-length normalization term. In contrast, BM25L and BM25+ introduce smoothing adjustments that mitigate this bias, providing more stable similarity estimates across heterogeneous article lengths.

We construct a bag-of-words query comprising approximately sixty inflation-related terms (e.g., inflation, deficit, economy, GDP, imports, exports). Each article is scored against this query using the BM25 variants, and the resulting scores are aggregated at a monthly frequency to produce a temporal relevance signal. Finally, we z-normalize the monthly averages and use the normalized series as the BM25-based inflation sentiment index. We observe that BM25L and BM25+ yield nearly identical results in our setting; therefore, we proceed with using BM25+ for subsequent analysis.

3.5. Task C - Sentiment Analysis Using Loughran-McDonald Lexicon

We employ a dictionary-based approach to quantify the polarity of economic news articles, utilizing the Loughran-McDonald Master Dictionary as described in Pratap and Ranjan (2021). Unlike general-purpose sentiment analysis, this method relies on a specialized financial lexicon to calculate a sentiment score S for each document, defined as the normalized sum of weighted term polarities:

$$S = \frac{\sum_{i=1}^{n} w_i}{N} \tag{1}$$

where w_i represents the sentiment weight of the i-th word and N represents the total number of words in the document.

Standard sentiment dictionaries are often unsuited for financial texts, as they frequently misclassify neutral business terminology (e.g., “cost,” “liability,” “tax”) as negative. The Loughran-McDonald lexicon addresses this by providing a domain-specific list of positive and negative terms curated specifically for financial narratives. Furthermore, simple keyword matching often fails to capture the nuances of language; for instance, the phrase "not good" contains a positive word but conveys a negative meaning. To address this, our approach incorporates contextual valence shifters, specifically negators and amplifiers, to ensure the sentiment score accurately reflects the text’s intended tone.

The sentiment calculation proceeds in a multi-stage process:

1. **Preprocessing:** First, we preprocess the text by tokenizing it into sentences and words, removing numbers, and converting all text to lowercase. We then filter the tokens to remove standard English stopwords while retaining essential context words.

2. **Base Weight Assignment:** For each remaining word, we assign a base weight (w_{base}). Terms found in the LM positive list are assigned $+1$, while terms in the negative list are assigned -1. Words not present in the lexicon are assigned a weight of 0.
3. **Contextual Adjustment (Look-back Mechanism):** We implement a "look-back" mechanism to adjust these weights based on the preceding context:
    - **Negation:** If a sentiment word is preceded by a negator (e.g., "not," "no," "never," "hardly"), the sign of the weight is flipped (e.g., $+1$ becomes -1).
    - **Amplification:** If a sentiment word is preceded by an amplifier (e.g., "very," "extremely," "highly"), the magnitude of the weight is doubled (e.g., $+1$ becomes $+2$, or -1 becomes -2).

Finally, the individual word weights are summed and divided by the total word count N of the article to produce a normalized document-level sentiment score. These scores are then aggregated by month to construct the final sentiment index.

3.6. Task D - Lexicon-Based Sentiment Analysis (VADER)

This task incorporates lexicon-based sentiment analysis, motivated by the need for simple, interpretable, and domain-adaptable sentiment extraction methods. Unlike model-based approaches, lexicon methods provide transparent scoring rules and offer a baseline for evaluating the effectiveness of more advanced techniques. They are particularly useful for economic text where certain keywords carry consistent polarities across contexts.

We employ the Valence Aware Dictionary and Sentiment Reasoner (VADER), a rule-based sentiment model widely applied to social media, news articles, and short-form text. VADER assigns polarity using predefined sentiment-weighted word lists and contextual heuristics, including negation handling, intensifiers, and punctuation-based emphasis. Each article receives a *compound sentiment score* in the range [1, 1], summarizing its overall polarity (-1 being most negative and $+1$ being most positive).

Although originally optimized for social media language, VADER remains a valuable baseline for economic text because of its transparent scoring mechanism and strong performance on general-purpose sentiment tasks. At the same time, its simplicity exposes limitations: VADER may misinterpret specialized financial terminology and lacks sensitivity to domain-specific semantic nuances. These limitations motivate the subsequent use of embedding-based and LLM-based models that better capture contextual and economic meaning..

The z-normalized monthly average sentiment across all articles forms a `financial sentiment index`. We compare this index with CPI inflation to evaluate its alignment with real economic trends.

3.7. Task E - Sentence Similarity Using Text Embeddings

A text embedding represents a text as a numerical vector that captures its semantic meaning. Modern embedding models derive these representations from the hidden layers of large transformer architectures trained on vast text corpora. We use embedding-based similarity because traditional keyword or frequency-based methods often fail to capture the nuanced and contextual ways in which inflation is discussed in news media. Text embeddings enable the mapping of semantically similar phrases such as "price pressures," "cost escalation," or "monetary tightening" to the inflation context, even when the exact keywords do not appear. This enables a more robust and context-aware measure of inflation-related sentiment. The method also performs well on long, unstructured news articles, making it suitable for our large-scale corpus.

For our study, we use the Qwen3-Embedding-0.6B model since the Qwen3 Embedding model series is the latest proprietary model of the Qwen family, specifically designed for text embedding and ranking tasks. We access the model through the LangChain Python library, using the `embed_document()` function to embed news articles and the `embed_query()` function to embed our inflation-related query.

We pass each article through the model to obtain its corresponding embedding vector. We also construct a context prompt that encapsulates the broader theme of inflation and convert it into an embedding vector using the same model. We then compute the cosine similarity between each article's embedding and the context prompt's embedding. The resulting similarity scores reflect how closely each article aligns with the inflation context. Finally, we z-normalize these scores and use the normalized series as the embedding-based inflation sentiment index.

3.8. Task F - LLM-based approaches: RoBERTa based sentiment pipeline and Cross-Encoder ranking pipeline

This subsection documents two distinct LLM-based methodologies evaluated for producing news-derived inflation indicators. The first approach uses a pre-trained RoBERTa classifier to extract per-article sentiment probabilities. The second approach uses a cross-encoder ranking model to produce per-article relevance/reasoning scores with respect to a curated inflation prompt. Both approaches are implemented with two aggregation variants (unweighted and weighted) and are benchmarked against official CPI using normalized RMSE. The slides and experimental tables referenced below summarize model choices, inference settings, thresholding strategy, and RMSE comparisons.

3.8.1. Approach A - RoBERTa based sentiment pipeline (classification-based sentiment)

Model and outputs: We apply the pre-trained RoBERTa classifier `cardiffnlp/twitter-roberta-base-sentiment` to every article. For each document the model returns a softmax-normalized probability vector

$$\big(p(\text{neg}),\ p(\text{neu}),\ p(\text{pos})\big),$$

which we treat as continuous sentiment features rather than coarse labels. Articles are processed in chunks and truncated to 1024 tokens to conform with model input constraints.

Unweighted variant (Twitter LLM - unweighted). In the unweighted variant we first optionally apply a cosine-similarity filter (retain articles with cosine $> \tau$) and then aggregate sentiment probabilities at monthly frequency without further scaling by relevance. Concretely, for month m the monthly sentiment indicator S_m^{unw} is computed as

$$S_m^{\text{unw}} = \frac{1}{N_m} \sum_{i \in \mathcal{D}_m} s_i,$$

where $\mathcal{D}_m$ is the set of articles in month m that pass the cosine threshold (if filtering is applied), N_m is the count, and s_i is the chosen scalar sentiment measure for article i.

Weighted variant (Twitter LLM - weighted). In the weighted variant, each article's sentiment is scaled by its continuous cosine similarity before monthly aggregation. The per-article weighted score is

$$w_i = \mathbf{1}\{\text{cosine}_i > \tau\} \times \text{cosine}_i \times s_i,$$

and the monthly weighted index is

$$S_m^{\text{w}} = \sum_{i \in \mathcal{D}_m} w_i.$$

This formulation integrates topical relevance (cosine) and sentiment intensity (LLM probabilities) so that sentiment coming from articles more semantically aligned with inflation topics has proportionally greater influence. We report weighted-RMSE values for a grid of thresholds τ (including $\tau \in \{-1, 0.0, 0.1, 0.2, 0.3, 0.4\}$) for both positive-only and combined Positive+Negative sentiment formulations.

Implementation notes and choices. The RoBERTa model is not fine-tuned on news text in the reported experiments, and is used off-the-shelf and therefore may suffer from a domain mismatch (tweet-trained vs news articles). Sentiment is therefore interpreted as a probabilistic surface-level polarity estimate and not as a causal statement about inflation pressure. The tweets-trained RoBERTa is nonetheless convenient due to its calibrated softmax outputs and fast inference at corpus scale.

3.8.2. Approach B - Cross-Encoder LLM scoring pipeline

Model and outputs. The cross-encoder model (`cross-encoder/ms-marco-MiniLM-L6-v2`) is applied as a high-precision text-ranking LLM. The model scores article-query pairs where the query is a carefully curated inflation prompt. The output is a scalar relevance score that captures the model's judgment about how strongly an article contributes to the inflation narrative. This cross-encoder is applied selectively (often only to articles that exceed a modest cosine similarity threshold), combining the efficiency of embedding-based filtering with the higher precision of LLM scoring.

Unweighted variant (Cross-Encoder - unweighted). In the cross-encoder unweighted pipeline, cosine similarity is again used as an initial filter, and the cross-encoder score itself is aggregated across each month without additional multiplicative weighting by cosine. That is, if r_i denotes the cross-encoder relevance score for article i, the monthly unweighted cross-encoder index is:

$$R_m^{\text{unw}} = \frac{1}{N_m} \sum_{i \in D_m} r_i.$$

This variant leverages the cross-encoder's fine-grained ranking but treats each retained article's contribution equally in the aggregation step (after normalization).

Weighted variant (Cross-Encoder - weighted). The weighted cross-encoder pipeline combines the cross-encoder score with cosine similarity multiplicatively, producing a per-article contribution

$$w_i = \mathbf{1}\{\text{cosine}_i > \tau\} \times \text{cosine}_i \times r_i,$$

and the monthly index is $R_m^{\text{w}} = \sum_{i \in D_m} w_i$. This approach retains the cross-encoder's high-precision relevance judgments while still rewarding articles with stronger semantic alignment via the cosine multiplier.

Operational trade-offs. The cross-encoder is more computationally expensive per article than the RoBERTa classifier; the slides therefore recommend using cosine-based pre-filtering (embeddings) to restrict cross-encoder inference to candidate articles (improving efficiency while preserving precision). The exact inflation prompt used to elicit cross-encoder judgments is curated and kept constant across experiments to ensure comparability.

3.8.3. Thresholding, normalization, smoothing, and final alignment

Across both pipelines we sweep the cosine threshold τ to study the trade-off between recall and precision. Articles below the chosen τ are either discarded entirely (when acting as a filter) or zeroed out in the weighted product formulation. After monthly aggregation (sum or mean depending on variant), the resulting series are standardized via z-score normalization. A centered rolling window is applied to reduce high-frequency noise prior to visual comparison with CPI. The aligned series are then compared to official CPI using RMSE computed after standardization.

3.8.4. Interpretation and methodological conclusions

The two LLM-based strategies implement different conceptual approaches: the Twitter-RoBERTa pipeline emphasizes polarity extraction (probabilistic sentiment features) that can be aggregated either directly or after relevance-based scaling. The cross-encoder pipeline emphasizes relevance- and judgment-based scoring (a ranking-style LLM) that is intended to measure each article's contribution to an inflation narrative. Empirically, combining embedding-based filtering with cross-encoder scoring and cosine weighting delivered the strongest RMSE performance against CPI in the reported experiments, suggesting that high-precision LLM judgments gated by semantic relevance yield a more informative monthly indicator than raw or relevance-scaled classifier probabilities alone.

4. Experimental Analysis

4.1. Evaluation Framework

To assess how closely each sentiment-based index tracks the official CPI inflation series, we evaluate all methods using the Root Mean Squared Error (RMSE) between the generated index and the normalized CPI. RMSE captures the average magnitude of deviation and is well-suited for comparing continuous temporal indicators, as it penalizes larger errors more heavily. All models are evaluated consistently over the full 2014-2024 period to ensure comparability across methods.

To contextualize the evaluation of our sentiment-based indices, we first examine the behavior of India's CPI inflation over the 2014–2024 period. Figure 4 shows the inflation series annotated with major macroeconomic events that coincide with notable spikes and dips in CPI. These events provide a reference frame for interpreting how well the constructed sentiment indices capture real economic fluctuations.

Method	RMSE
TF–IDF Sentiment	1.4648
BM25	1.5326
BM25+	1.1601
Negated BM25	0.9208
Negated BM25+	1.5168
Loughran–McDonald Lexicon	1.2853
VADER (Positive)	1.3344
VADER (Negative)	1.2372
VADER (Compound)	1.2199

Table 2
RMSE comparison for TF–IDF, BM25 variants, negated BM25 variants, Loughran–McDonald lexicon scores, and VADER sentiment scores.

Method	-1	0	0.1	0.2	0.3	0.4
Embeddings	1.2444	1.2601	1.2379	1.1831	**1.0833**	1.1003
Twitter LLM (Weighted Pos)	**1.2574**	1.2663	1.2637	1.2873	1.3628	1.4050
Twitter LLM (Unweighted Pos)	1.2862	1.2800	**1.2619**	1.2820	1.3720	1.4120
Twitter LLM (Weighted Pos+Neg)	1.2688	1.2727	1.2410	1.2246	1.2354	**1.0439**
Twitter LLM (Unweighted Pos+Neg)	1.3090	1.3052	1.3240	1.2801	1.3276	**1.1475**
Cross-Encoder LLM (Weighted)	1.0174	0.9751	**0.9416**	0.9584	1.0696	1.0747
Cross-Encoder LLM (Unweighted)	0.9397	0.9269	**0.8921**	0.9568	1.8676	1.0814

Table 3
RMSE comparison across threshold-based embedding similarity, RoBERTa-based LLM scoring, and cross-encoder LLM scoring (values rounded to four decimal places) for different thresholds.

4.2. RMSE results

To quantitatively evaluate the performance of all approaches, we compute the RMSE between the generated sentiment index and the normalized CPI series. Table 2 reports RMSE values for TF-IDF, BM25 variants, their negated counterparts, the Loughran-McDonald lexicon, and VADER sentiment scores. Table 3 presents the threshold-dependent results for embedding-based similarity, RoBERTa-based LLM scoring, and cross-encoder LLM scoring. Here, the threshold refers to a cosine-similarity cutoff applied to the embedding model, such that only articles whose embedding similarity to the inflation context exceeds this value are retained. The two tables together provide a comprehensive overview of how each method aligns with CPI across different parameter settings.

4.3. Time-Series Visual Comparison

While RMSE provides a quantitative measure of alignment with CPI, it does not fully capture the temporal behavior of each index. To complement the numerical results, we present time-series plots comparing the *best-performing variant* of each method against the normalized CPI series. These graphs illustrate how effectively each approach tracks major inflation trends, turning points, and periods of volatility over the 2014–2024 horizon.

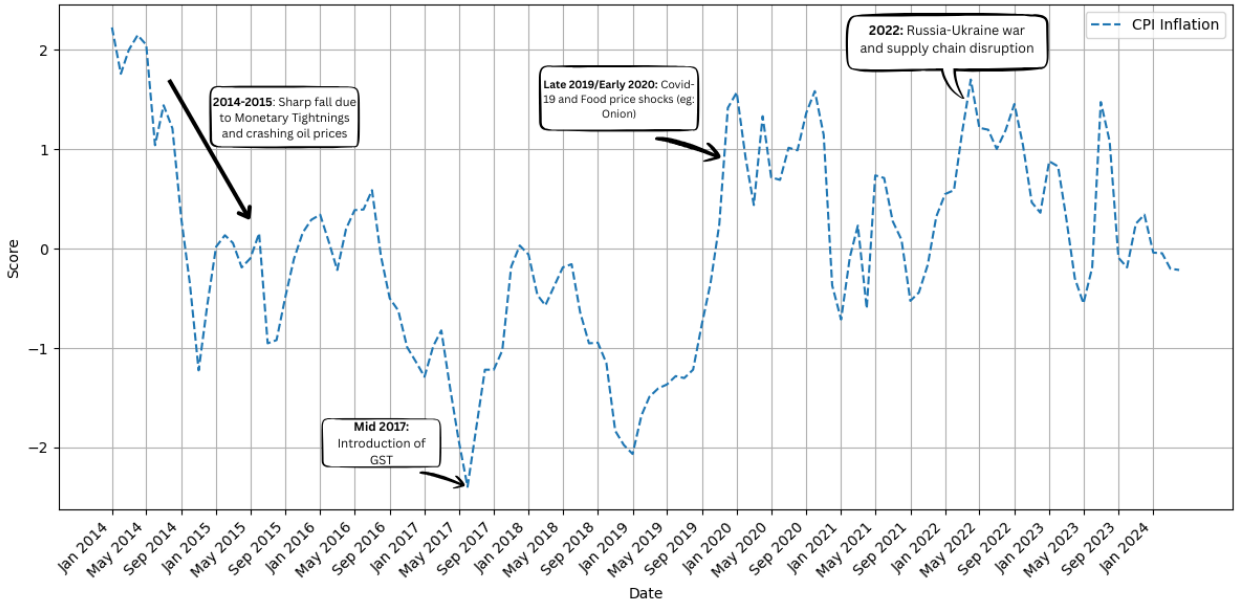

Figure 4: Annotated CPI inflation series for India (2014-2024), highlighting major macroeconomic shocks. These events correspond to sharp movements in CPI and serve as contextual anchors for later analysis.

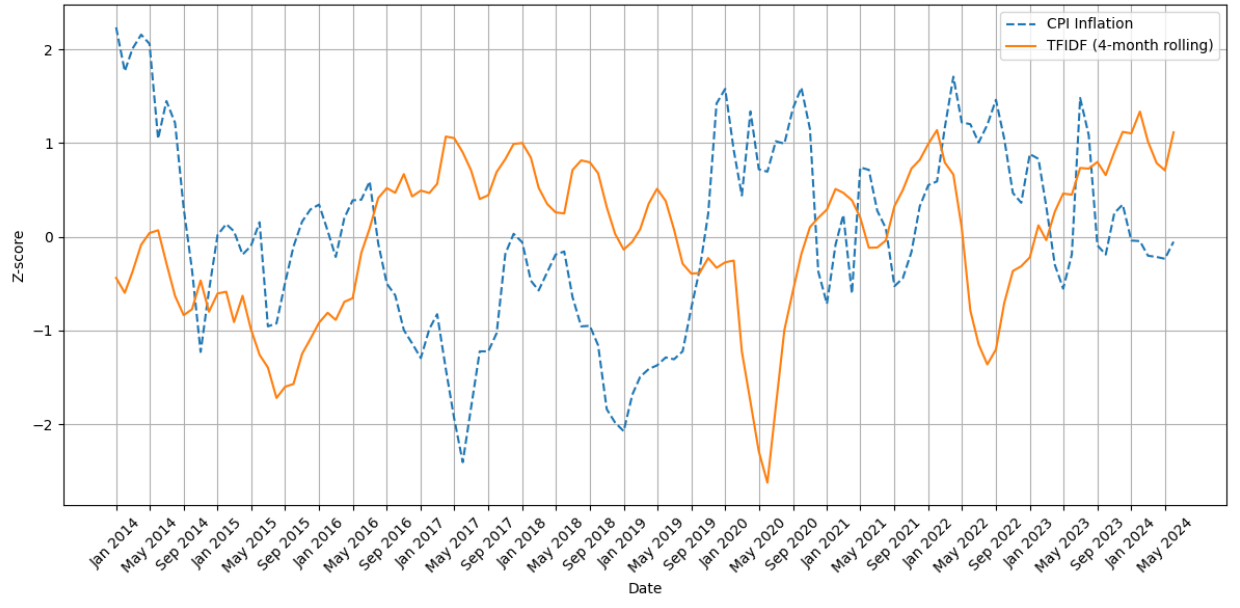

Figure 5: Time-series comparison of the TF-IDF-based inflation sentiment index and the normalized CPI inflation. The TF-IDF index captures broad fluctuations but shows noticeable noise and weaker alignment during major turning points.

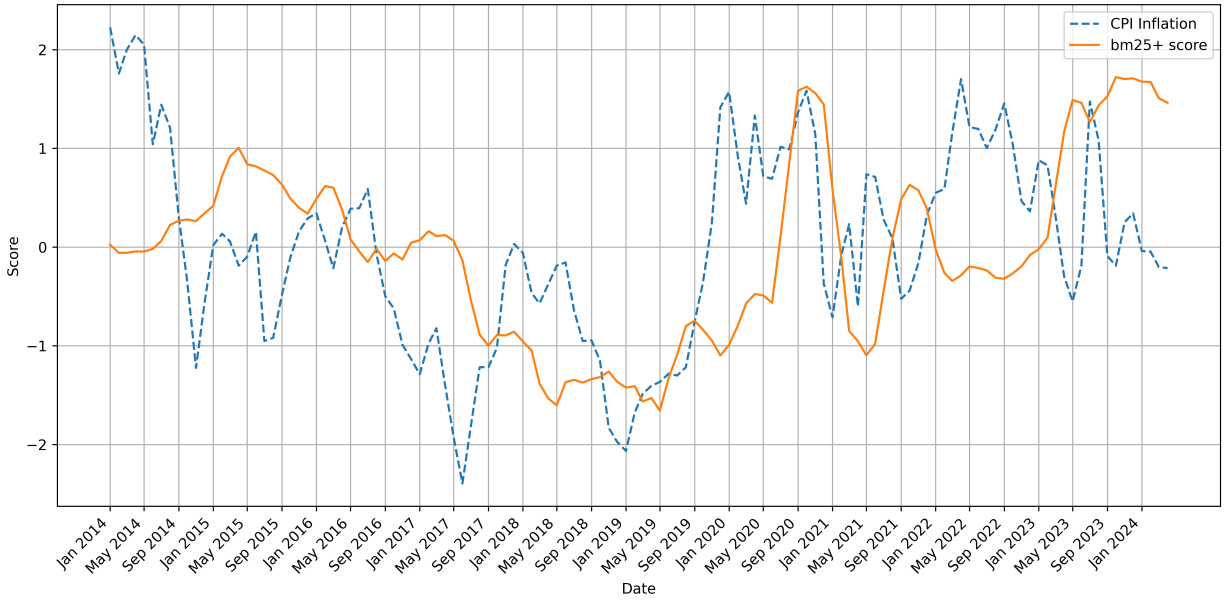

Figure 6: Time-series comparison of the BM25+ index and the normalized CPI inflation. BM25+ offers smoother dynamics and improved trend matching compared to TF-IDF, reflecting more stable term weighting and a better capture of inflation-related vocabulary.

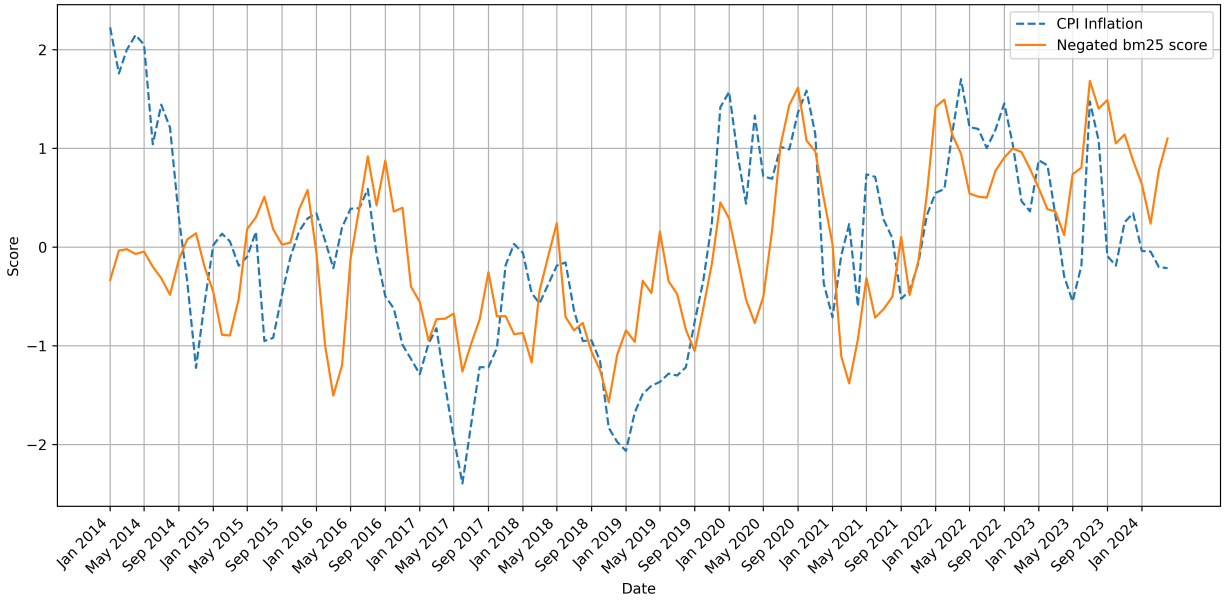

Figure 7: Time-series comparison of the negated BM25 index and the normalized CPI inflation. Negation improves alignment by correcting the polarity mismatch inherent in relevance scoring, resulting in closer tracking of inflation cycles.

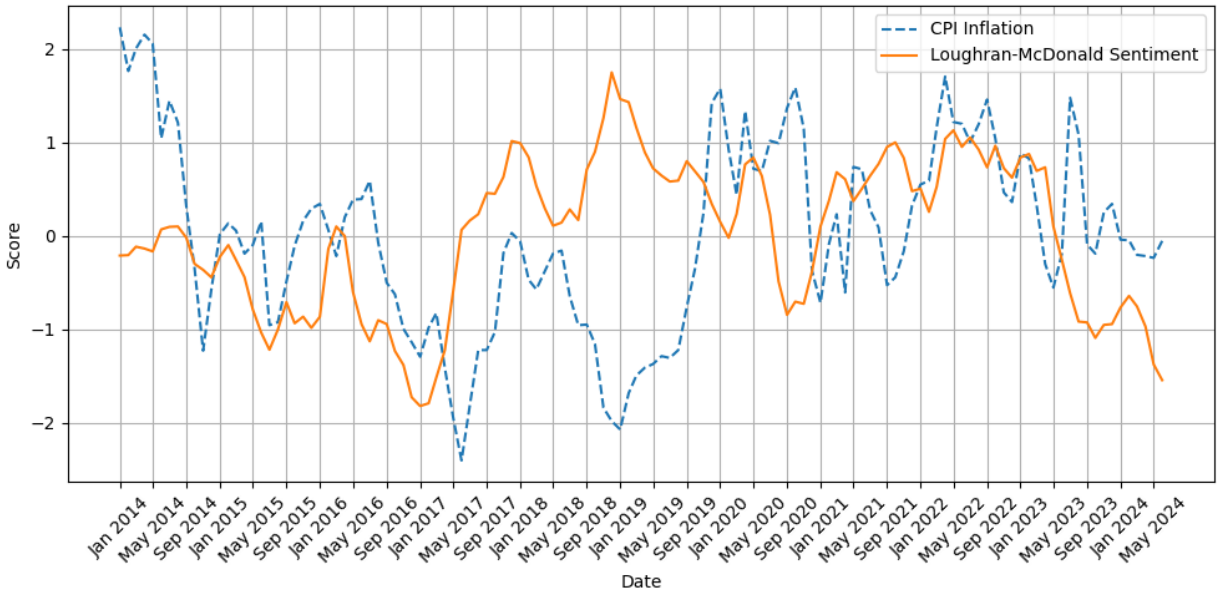

Figure 8: Time-series comparison of the Loughran–McDonald lexicon-based sentiment index and the normalized CPI inflation. The figure illustrates the co-movement between monthly sentiment extracted from financial news and realized inflation dynamics, highlighting periods where sentiment shifts precede changes in CPI.

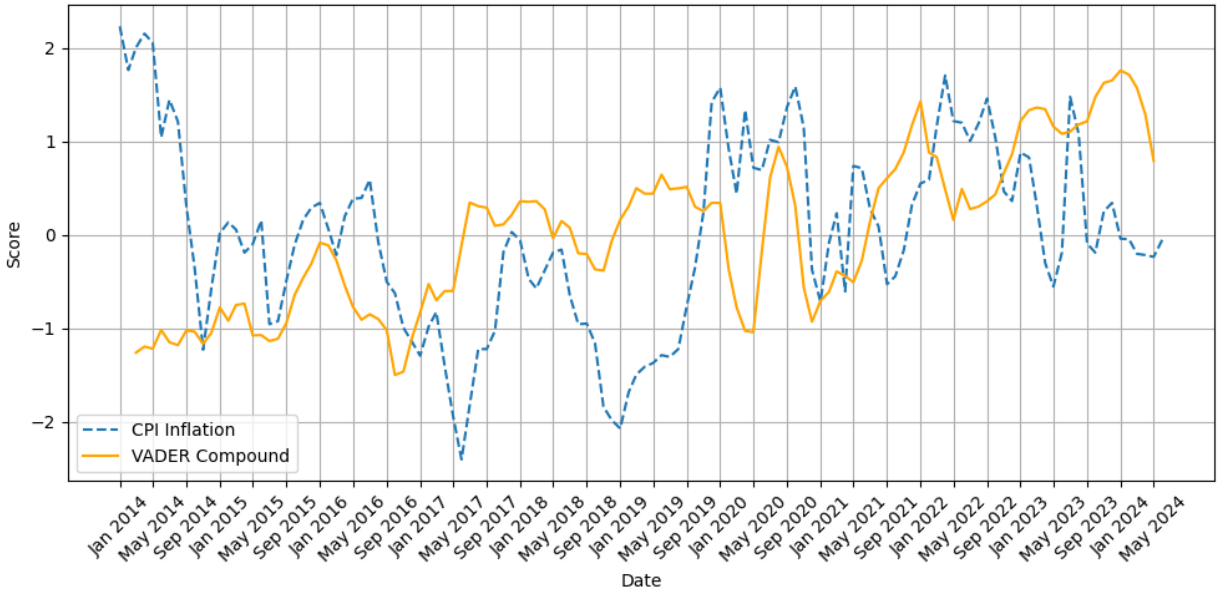

Figure 9: Time-series comparison of the VADER-compound sentiment index and CPI inflation. VADER exhibits high variability and weaker inflation-related structure, reflecting the limitations of general-purpose sentiment lexicons for economic text.

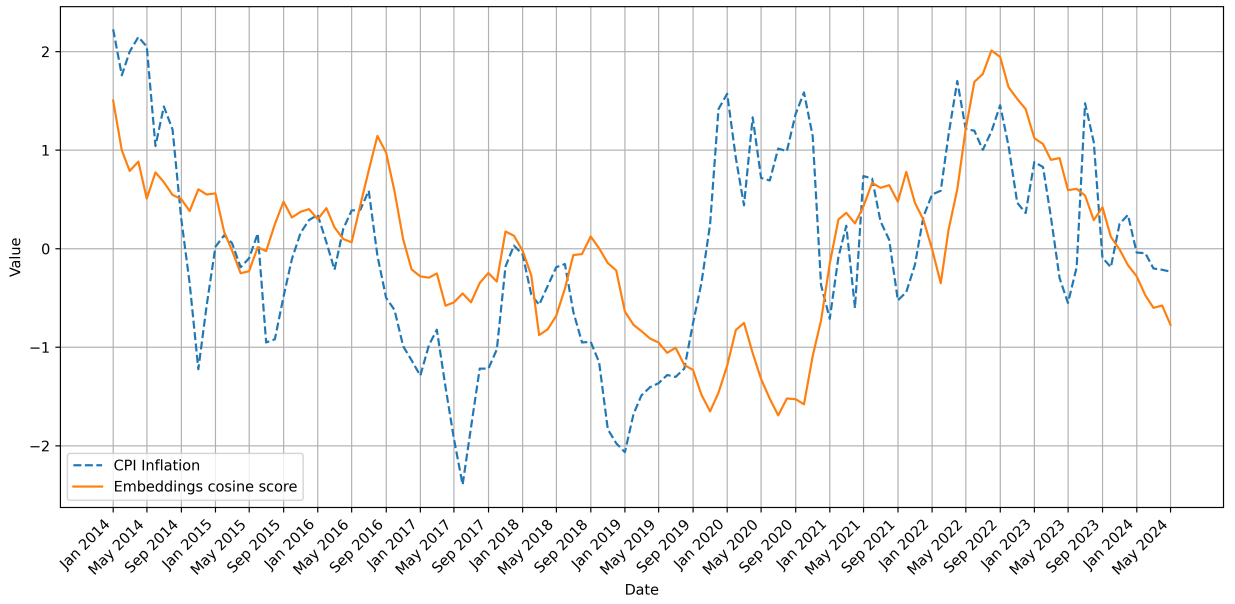

Figure 10: Time-series comparison of the embedding-based inflation sentiment index (best-performing threshold) and CPI. Semantic filtering through embeddings enhances trend capture and smoothness, yielding closer alignment with CPI than lexical methods.

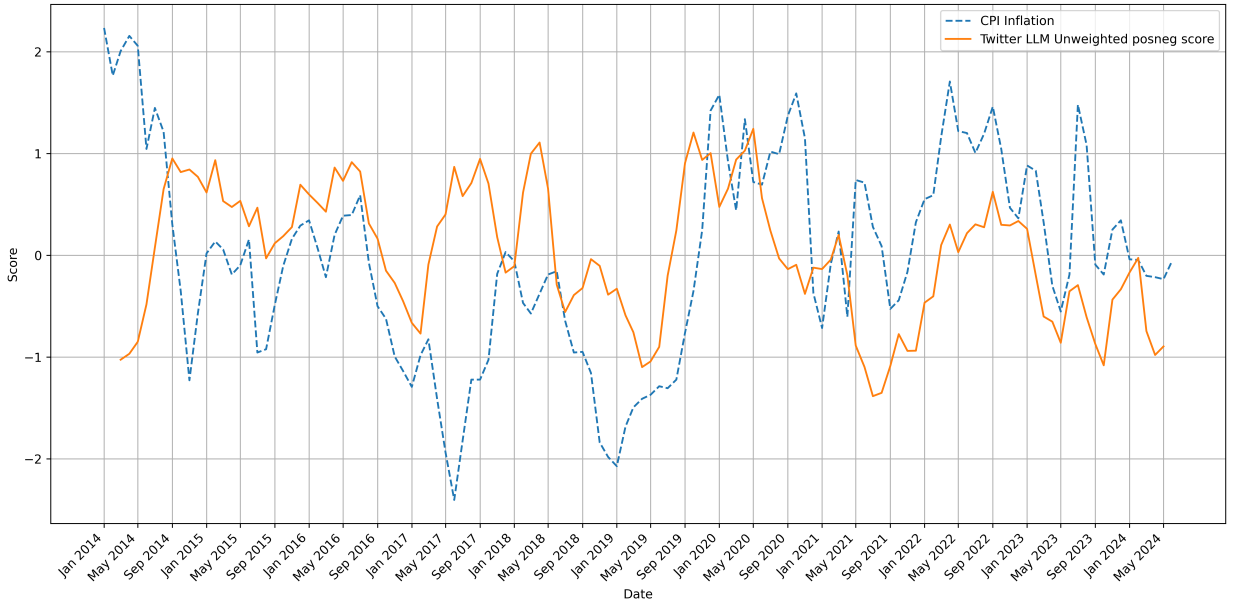

Figure 11: Time-series comparison of the unweighted Pos+Neg Twitter RoBERTa sentiment index and the normalized CPI. The unweighted variant shows limited alignment with CPI trends and exhibits substantial noise, reflecting the difficulty of applying social-media-trained sentiment models to formal economic news text.

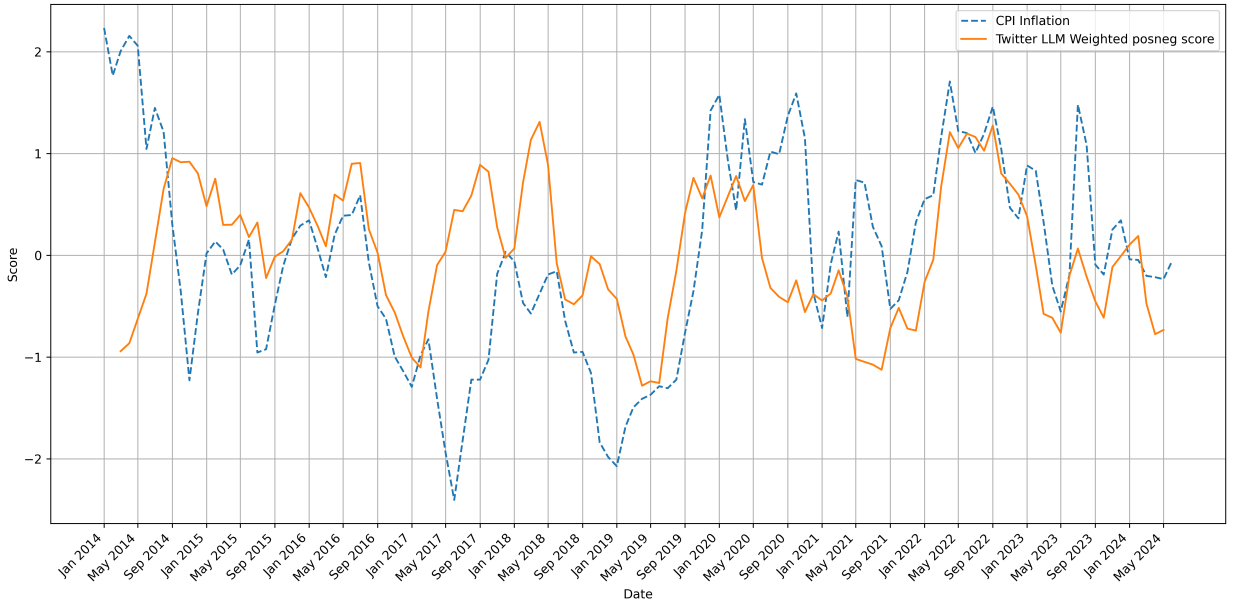

Figure 12: Time-series comparison of the weighted Pos+Neg Twitter RoBERTa sentiment index and CPI. The weighted variant performs similarly poorly, with weak trend correspondence and high volatility, indicating that the Twitter-trained sentiment model does not effectively capture inflation-related sentiment in news articles.

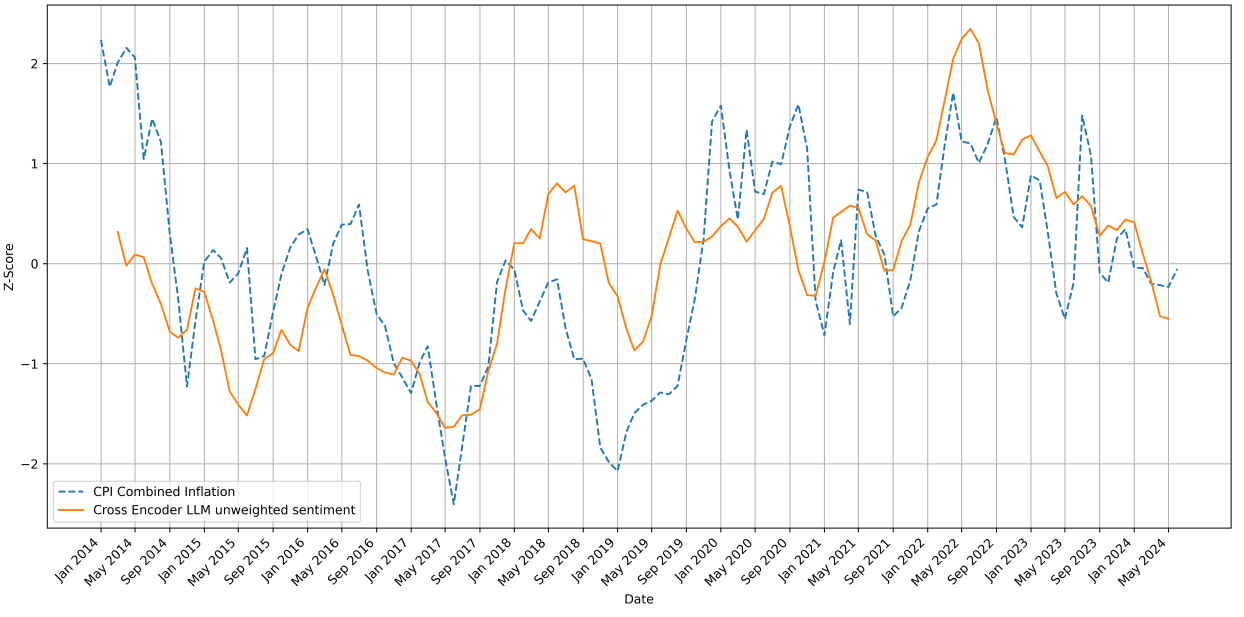

Figure 13: Time-series comparison of the unweighted cross-encoder LLM-based inflation sentiment index and CPI. The unweighted approach yields the lowest RMSE among all methods, closely tracking CPI movements while preserving the model's fine-grained contextual understanding of inflation-related language.

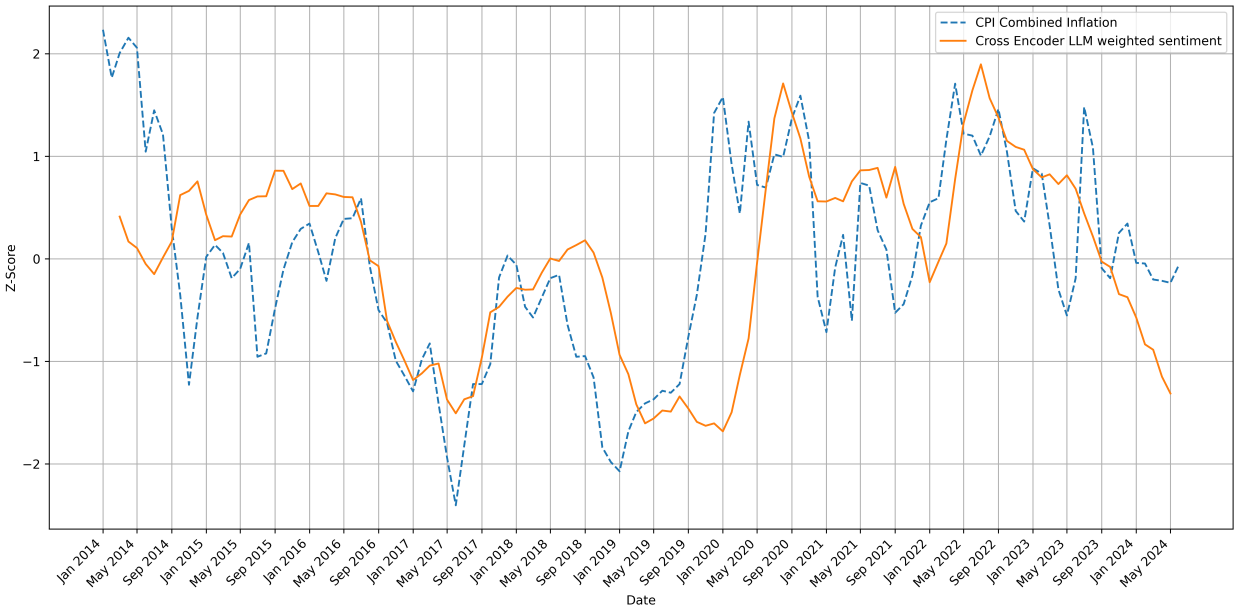

Figure 14: Time-series comparison of the weighted cross-encoder LLM-based inflation sentiment index and the normalized CPI. Weighting sentiment probabilities produces a smoother and more responsive signal, resulting in strong alignment with CPI trends and improved detection of major inflation turning points.

Across all methods, we observe a clear hierarchy in performance. Traditional lexical and frequency-based techniques, including TF-IDF, BM25, Loughran-McDonald Lexicon, and VADER, capture broad fluctuations in inflation sentiment but exhibit substantial noise and limited temporal alignment with the CPI. Negated BM25 improves performance by correcting polarity mismatches inherent in relevance scoring. Embedding-based similarity achieves lower RMSE and smoother temporal behavior, demonstrating the advantage of semantic filtering. The RoBERTa Twitter sentiment model performs poorly, reflecting domain mismatch between social-media text and economic news. The strongest results arise from the cross-encoder LLM, whose contextual scoring yields the lowest RMSE and the closest visual correspondence with CPI, effectively capturing turning points and inflation cycles. Overall, deeper semantic modeling provides the most reliable inflation sentiment indicators.

5. Conclusion

Across all evaluated methods, our results demonstrate that incorporating relevance signals-rather than relying solely on raw sentiment-substantially improves the predictive quality of news-based indicators. While unweighted LLM sentiment provides a strong baseline due to its ability to capture nuanced textual cues, it is sensitive to irrelevant or weakly related articles. Methods that combine both sentiment and relevance, such as cosine-filtered LLM scoring and cross-encoder–based relevance estimation, consistently achieve lower RMSE and produce smoother, more interpretable temporal trends.

Weighted approaches show clear advantages when relevance estimates are reliable, particularly in settings where sentiment needs to be amplified or suppressed based on the economic importance of an article. The hybrid cosine-LLM pipeline emerges as one of the strongest methods in practice: cosine similarity serves as an effective low-cost filter, and LLM sentiment provides fine-grained polarity assessment. Cross-encoder relevance scoring further enhances performance by jointly modelling text and inflation keywords, though its computational cost is higher.

Overall, the study confirms that well-designed filtering and weighting strategies are crucial for extracting meaningful inflation signals from large textual datasets. The comparison across methods also reveals the trade-off between interpretability, cost, and predictive accuracy, offering practical guidance for future sentiment-based economic forecasting systems.

6. Future Work

Several directions can strengthen and extend this research:

1. `Richer Economic Feature Integration:` Combine sentiment signals with traditional macroeconomic variables such as commodity prices, monetary policy indicators, and market indices using multivariate forecasting models.
2. `Advanced Temporal Models:` Employ models such as VAR, LSTMs, Transformers, or state-space approaches to capture long-range dependencies and nonlinear interactions.
3. `Finer-Grained Article Classification:` Introduce topic models or supervised relevance classifiers trained specifically on inflation-related content to improve filtering beyond cosine thresholds.
4. `Event-Driven Weighting Mechanisms:` Dynamically adjust weights around major policy announcements, geopolitical shocks, or supply-chain disruptions to capture periods of heightened economic sensitivity.
5. `Cross-Country Generalization:` Extend the pipeline to multi-country datasets to evaluate robustness across different economic structures and media ecosystems.
6. `Real-Time Deployment Pipeline:` Implement automated scraping, filtering, scoring, and inflation nowcasting in a real-time system to assess practical utility.
7. `Human-in-the-Loop Evaluation:` Incorporate expert feedback for validating relevance labels and calibrating sentiment scores, potentially further reducing noise.

7. Individual Contribution

The project tasks were divided among the team members as follows. For data acquisition, the news archives were distributed across sources: The Economic Times was extracted by Nandeesh, The Times of India by Sahil, and The

Financial Express by Karmanya. Each member independently carried out the preprocessing for their respective datasets, including cleaning, filtering, and structuring the articles for downstream analysis.

In the methodology development, Sahil implemented the TF–IDF similarity model and the Loughran–McDonald lexicon-based approach. The BM25 family of retrieval techniques and the sentence-embedding-based similarity models were implemented by Karmanya. The VADER sentiment model and all LLM-based scoring approaches were developed by Nandeesh.

Finally, the experimental evaluation, interpretation of results, and preparation of visualizations were jointly conducted by all team members, with collaborative effort in refining the findings and integrating them into the final report.

A. Supplementary Materials

The sourcecode, datasets, and additional details for this work are available here

- https://github.com/iitgoa-ml/ml-economics.
- https://drive.google.com/drive/folders/1yM4LT7UoClEatfoEb9N2Wlmndth09LJw?usp=sharing

References

Allard, M.A., Teiletche, P., Zinebi, A., 2024. Enhancing inflation nowcasting with llm: Sentiment analysis on news. arXiv preprint arXiv:2410.20198 .

Eugster, P., Uhl, M.W., 2024. Forecasting inflation using sentiment. Economics Letters 236, 111575.

Hutto, C., Gilbert, E., 2015. Vader: A parsimonious rule-based model for sentiment analysis of social media text.

Pratap, B., Ranjan, A., 2021. Forecasting food inflation using news-based sentiment indicators. RBI-Occasional Papers 42.

Seki, K., Ikuta, Y., Matsubayashi, Y., 2022. News-based business sentiment and its properties as an economic index. Information processing & management 59, 102795.

Shapiro, A.H., Sudhof, M., Wilson, D.J., 2022. Measuring news sentiment. Journal of econometrics 228, 221–243.

Trotman, A., Puurula, A., Burgess, B., 2014. Improvements to bm25 and language models examined, in: Proceedings of the 19th Australasian Document Computing Symposium, pp. 58–65.