CARMA 2026

8th Int. Conf. on Advanced Research Methods and Analytics
1-3 July 2026, Valencia, Spain
DOI: https://doi.org/10.4995/CARMA2026.2026.22007

'Pretend You Are a Central Bank Economist': Exploring GenAI-Based News Surveillance using Azure OpenAI

Mary Kryslette Bunyi[1]
[1]Department of Economic Research, Bangko Sentral ng Pilipinas, Philippines

How to cite: Bunyi, M. K. (2026). 'Pretend You Are a Central Bank Economist': Exploring GenAI-Based News Surveillance using Azure OpenAI. In: 8th International Conference on Advanced Research Methods and Analytics (CARMA 2026). Valencia, 1-3 July 2026. https://doi.org/10.4995/CARMA2026.2026.22007

Abstract

Central banks must monitor developments and sentiment across multiple sectors of the economy. While traditional news sentiment indices complement official statistics, they often lack context and nuance that generative AI (GenAI) can provide. This paper develops a GenAI-based macroeconomic surveillance framework to analyze news articles. It evaluates the reliability of large language models (LLMs) by benchmarking outputs against human judgment as well as dictionary-based and machine learning approaches. It also examines the usefulness of an LLM ensemble for validation. Results suggest that GenAI exhibits the highest agreement with human judgment and performs well on clear issues but struggles with neutral and contentious cases. The index also correlates moderately with existing sentiment measures. Finally, findings suggest that a single-LLM pipeline exhibits general reliability, so ensemble validation is best reserved for specialized, complex cases. Overall, this framework illustrates the reliability and practical value of GenAI for central bank news analytics.

Keywords: *Generative AI; central banking; news surveillance; sentiment analysis.*

1. Introduction

Central banks routinely monitor news to assess economic developments, risks, and sentiment across sectors. As tracking high-volume, diverse news coverage can be resource-intensive and time-consuming, news-based indices have thus emerged as a widely used tool for economic surveillance (e.g., Arcin et al., 2024; Nguyen & La Cava, 2020; Shapiro et al., 2020). Nonetheless, while useful for summarizing sentiment into a single indicator, these indices are limited in their ability to support targeted analysis and uncover narratives underlying sentiment.

Recent advances in GenAI have addressed the limitations of traditional news-based indices. Prior studies have applied GenAI to summarize news, quantify uncertainty and sentiment, and perform classification tasks (Ghomi & Hurtado, 2026; Kwon et al., 2025; Lee, 2026).

Article in Press

This work is licensed under a Creative Commons License CC BY-NC-SA 4.0
Editorial Universitat Politècnica de València

This paper contributes to the literature by leveraging GenAI to produce topic-specific sentiment assessments, distinguish between past and forward-looking coverage, and assess relevance to monetary policy. It also introduces a dual-layer validation framework—LLM-as-a-judge and LLM-as-an-arbiter—to audit outputs. Using two institutional datasets, the paper benchmarks LLM performance against human-labeled data and the Bangko Sentral ng Pilipinas (BSP)'s existing dictionary- and machine learning (ML)-based sentiment indices (Arcin et al., 2024). To the author's knowledge, this is the first study to assess the contribution of dual-layer LLM validation in this context.

The paper pursues three objectives: (i) to assess the reliability of LLM-generated sentiment analyses against human judgment; (ii) to compare LLM performance against dictionary- and ML-based approaches, and (iii) to evaluate whether ensemble validation adds meaningful value over a single-LLM pipeline.

2. Related Literature

Various institutions have developed news-based sentiment indices to supplement traditional macroeconomic indicators. At the BSP, research has applied genetic algorithm to construct a Philippine-specific lexicon and FinBERT for sentiment analysis (Arcin et al., 2024). Similar dictionary and ML-based approaches underpin indices developed by the Federal Reserve Bank of San Francisco, International Monetary Fund, and Reserve Bank of Australia (Shapiro et al., 2020; Huang et al., 2019; Nguyen & La Cava, 2020).

However, these approaches do not consider the full context as they judge sentiment based on word-level signals. Sentiment may also be diluted by less relevant content, such as firm disclosures and earnings reports, which tend to skew sentiment positively.

Recent advances in GenAI offer a way to address these limitations. For instance, a Bank of Korea study supplements keyword matching and semantic clustering with GenAI to summarize the contents of identified clusters (Lee, 2026). Meanwhile, Bank for International Settlements research uses LLM prompting to rate directional sentiment related to growth and inflation and classify the dominant demand and supply drivers within each article (Kwon et al., 2025). More recently, a Banco de España paper constructed topic-specific uncertainty indicators using semantic retrieval to filter topic-specific news articles and LLM prompting to quantify the level of uncertainty in each article (Ghomi & Hurtado, 2026).

This paper adds to existing approaches in three ways. First, it leverages GenAI beyond narrow classification tasks and tests the extent of what an LLM can do beyond sentiment scoring and summarization, supplementing these with temporal delineation, policy relevance assessment, and explanation generation. Second, it benchmarks LLM outputs against human-labeled data

Article in Press

and existing institutional indices. Third, it introduces and evaluates dual-layer validation guardrails (LLM-as-a-judge and LLM-as-an-arbiter).

3. Methodology

3.1. Data

The analysis uses 37,597 finance and business news articles published from 1 February 2025 to 30 April 2026, covering fifteen months of reporting from five major Philippine broadsheets.[1] The data are drawn from the BSP's repository of web-scraped articles used in constructing existing news-based indicators.

For evaluation, LLM-generated sentiment scores are compared with those from BSP's dictionary- and ML-based methods. In addition, outputs are benchmarked against 3,012 human-labeled sentences used as training data in Arcin et al. (2024).

3.2. LLM Analyst Framework

The system processes articles from a user-uploaded CSV file of news articles. The LLM ingests each article's headline, publication date, and full text. Given GenAI's capabilities, preprocessing is limited to basic encoding and whitespace normalization

Each article is analyzed in a single LLM call (Azure OpenAI GPT-5.4-mini). The model generates an article summary, sentiment label and score, and an importance rating for monetary policy. Considering the importance of outlook in monetary policy formulation, the LLM is also asked to distinguish between past and forward-looking information.

The framework also produces topic-specific analyses across a range of topics aligned with BSP surveillance priorities, including monetary policy, inflation, economic activity, exchange rate, equities, and rice-related developments.

Considering computational costs, the LLM conducts a second-pass topic-specific analysis only for articles exceeding an importance threshold of 0.8. This filters out less relevant content, such as company-specific disclosures and highly sector-specialized developments, among others. The threshold narrows the in-depth analysis to 16.6 percent of articles.

3.3. Ensemble Validation

The framework investigates the value-added of incorporating an ensemble LLM validation mechanism to reduce hallucinations. A second LLM (GPT-4.1) evaluates whether the LLM analyst's outputs are substantiated by the source text, retrieving relevant excerpts from the

[1] The news outlets are Business Mirror, Business World, Manila Bulletin, Manila Standard, and Philippine Daily Inquirer.

Article in Press

source article where necessary. When disagreements arise, a third LLM (GPT-5.4) arbitrates and judges the competing outputs without knowing their source. It rates each candidate's accuracy separately and decides which has a stronger argument, or abstains when the articles are roughly equally correct or are both false or unsupported.

This three-stage design provides a structured accuracy check and aims to mitigate model bias. Given computational costs, ensemble validation is applied to a random sample of 1,000 articles.

4. Results

4.1. Agreement with Human Judgment

The LLM outperforms both dictionary- and ML-based sentiment methods in terms of agreement with human-labeled data (Table 1). The LLM appears to perform well across clear and policy-relevant issues, with errors appearing to cluster around neutral, contentious, and lower-importance sentences (Table 2).

Table 1. Performance Metrics of Sentiment Indices. Source: Author's estimates; Arcin et al. (2024).

Model	Accuracy	Macro F1
NSI-D (Arcin et al., 2024)	0.66	0.64
NSI-ML (Arcin et al., 2024)	0.66	0.63
LLM Analyst	0.73	0.71

Table 2. LLM Analyst Accuracy and Macro F1 by Subsample. Source: Author's estimates.

Subsample	n	Accuracy	Macro F1
By human label			
Negative	1,176	0.82	-
Neutral	880	0.52	-
Positive	956	0.80	-
By human agreement			
Unanimous	1,802	0.85	0.83
Disagreement	1,210	0.55	0.54
By importance quartile			
Lowest quartile (0.00–0.10)	1,004	0.69	0.68
Highest quartile (0.43–0.98)	745	0.80	0.65[2]
By absolute sentiment score			
Weak (< 0.2)	694	0.66	0.27
Strong (≥ 0.4)	625	0.88	0.63

[2] The macro F1 score for highest-importance sentences was pulled down by the neutral class (precision of 0.38 and recall of 0.20).

Article in Press

The model appears more aggressive in tagging directional sentiment, correctly identifying only half of human-labeled neutral sentences. Accuracy improves with clearer signals: it reaches 85.2 percent when human raters are unanimous, but declines to 54.6 under inter-rater disagreement. Performance also increases with sentence importance and sentiment strength. Near-neutral sentiment yields 66.4 percent accuracy, while strong sentiment achieves 87.7 percent accuracy.

Overall, these results suggest that the LLM is most reliable on clear polarities and policy-relevant issues. One possible explanation for its reduced effectivity in neutral and ambiguous cases is that sentiment classification at the sentence-level may lack the broader context that a human expert would already know but an LLM might still need. Reading full articles might improve the LLM's ability to properly detect sentiment.

4.2. Comparison with Existing Indices

For business news articles from February 2025 to April 2026, the LLM exhibits broadly similar patterns to the ML-based index in both direction and dispersion of sentiment scores, although it tends to show higher agreement with the dictionary-based measure (Table 3).

Table 3. Comparison of LLM, ML, and Dictionary Sentiment Scores. Source: Author's estimates.

Metric	LLM	ML	Dictionary
Mean	0.09	0.11	0.01
Std. Dev.	0.31	0.28	0.03
Range	[-0.95, 0.80]	[-1.00, 1.00]	[-0.19, 0.17]
Spearman ρ vs. NSI-ML	0.46	-	0.44
Spearman ρ vs. NSI-D	0.55	0.44	-
Sign agreement vs. NSI-ML	0.63	-	0.62
Sign agreement vs. NSI-D	0.71	0.62	-

The dictionary-based index appears more conservative, with article scores closer to neutral. By contrast, the LLM exhibits larger dispersion, suggesting stronger differentiation across articles and potentially higher sensitivity to shifts and movements in sentiment. Nonetheless, the LLM remains generally consistent with the existing methods, showing moderate correlation with both and notably exhibiting higher sign agreement with each method than they do with each other.

Table 3 indicates that sentiment indices are generally positive-leaning, reflecting the prevalence of favorable business news such as firm disclosures. However, dissecting by importance reveals a clear divergence between low- and high-relevance articles. Low-importance news (≤0.2) are predominantly positive, with 65.1 percent tagged positive, 13.7 percent as negative, and a mean sentiment of +0.14. In contrast, high-importance articles (≥0.8) exhibit more negative sentiment, with only 42.1 percent positive, 56.2 percent negative, and a mean sentiment of -0.11. This suggests that monetary policy-relevant news over the sample period have tended to cover macroeconomic risks and challenges.

Article in Press

4.3. Ensemble Validation

Results suggest high LLM analyst output reliability, with 85.9 percent of the sampled articles judged as supported by the LLM validator. Disagreements are typically isolated to a few fields (top 3 are summary of past developments, outlook sentiment score, and monetary policy importance score) rather than the entire output, with almost half (47.5 percent) involving a single field. The arbiter also favors the analyst over the validator in a majority of cases (56.4 percent).

The average article processing time increases from 4.5 to 16.0 seconds with the addition of the LLM validator and arbiter. The additional LLM calls also raise computational costs. Considering the strong performance of the single-LLM pipeline, ensemble validation appears more suitable for complicated cases requiring more nuance or in-depth reasoning.

4.4. Thematic Analysis of LLM Outputs

A key advantage of the GenAI framework is its ability to produce sector-level sentiment. Figure 1 shows that overall sentiment may have been generally stable and slightly positive, but underlying monetary policy-relevant topics have actually exhibited more pronounced and heterogeneous movements since February 2025.

Notably, in March 2026, sentiment declined sharply across sectors following a global energy shock resulting from the war in Iran. The largest decline was recorded in oil, with immediate spillovers to equities, FX, and inflation.

Beyond sentiment scoring, the LLM's ability to generate summaries of sector-specific developments would also enable analysts to quickly learn about the drivers of shifts in sentiment and compare developments across time.

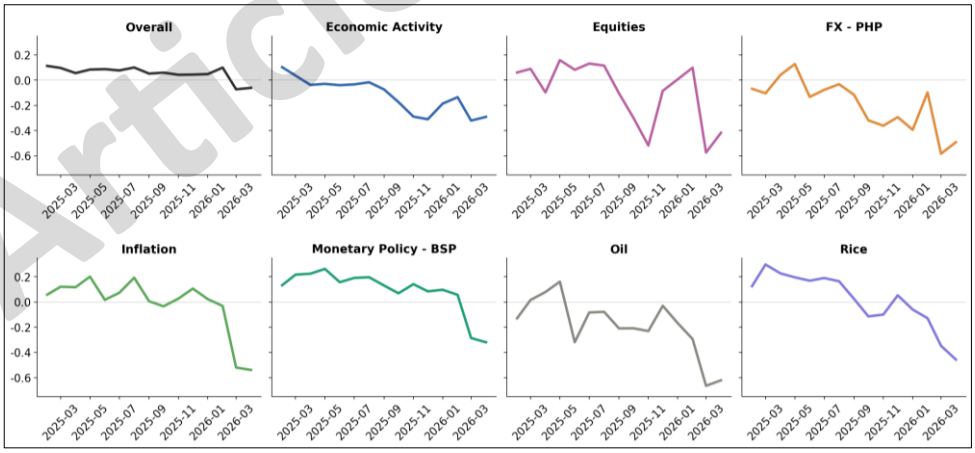


Figure 1. Monthly importance-weighted sentiment across topics. Source: Author's estimates.

Article in Press

5. Conclusion

This paper develops a GenAI-based system for macroeconomic news surveillance and evaluates its performance against human judgment and existing dictionary- and ML-based sentiment indices, as well as within the framework of an LLM-based ensemble validation.

Results show that the system outperforms traditional methods in aligning with human judgment. Errors cluster around generally more ambiguous cases, suggesting that the LLM is most reliable in matters most important for policy and highlighting the value of filtering high-signal content.

The GenAI-based index is moderately correlated with BSP's operational sentiment measures and tends to agree more with each method than they do with each other. It also exhibits greater sensitivity in scoring sentiment, which may enhance the detection of emerging shifts in news sentiment and improve the agility of economic surveillance.

Ensemble validation results further support the reliability of the LLM analyst. Considering the strong performance of the single-model pipeline, ensemble validation may be reserved for more contentious and complicated analyses requiring more nuance and varied perspectives.

Finally, this exercise highlights a key advantage of a GenAI sentiment index over traditional sentiment indices – its ability to dissect sentiment into sectors of interest with minimal nudging, supervision, and oversight.

Overall, the findings demonstrate the value of GenAI as a complementary tool for economic surveillance and monetary policy formulation. The LLM framework can support more timely, granular, and interpretable assessments, with potential portability to other textual data, topics of interest, and analytical focus.

References

Arcin, A. C. T., Esclanda-Lo, C. G., Ong, C. A. S., & Reyes, R. S. (2024). Constructing high-frequency and thematic economic sentiment indicators from online news articles. *Bangko Sentral ng Pilipinas (BSP) Economic Newsletter 24-02*. https://www.bsp.gov.ph/Media_And_Research/Publications/EN24-02.pdf

Ghomi, M., & Hurtado, S. (2026). RAUI: Uncertainty indicators built with artificial intelligence. *Banco de España Working Paper No. 2609.* https://doi.org/10.53479/42605

Huang, C., Simpson, S., Ulybina, D., & Roitman, A. (2019). News-based sentiment indicators. *International Monetary Fund (IMF) Working Paper No. 19/273*. https://www.imf.org/en/publications/wp/issues/2019/12/06/news-based-sentiment-indicators-48740

Kwon, B., Park, T., Rungcharoenkitkul, P., & Smets, F. (2025). Parsing the pulse: Decomposing macroeconomic sentiment with LLMs. *Bank for International Settlements (BIS) Working Paper No. 1294*. https://www.bis.org/publ/work1294.pdf

Article in Press

Lee, Y. (2026). Extracting economic issues from news data: With the help of generative AI. In Bank for International Settlements (Ed.), *Data science in central banking: Exploring generative AI* (IFC Bulletin No. 67). https://www.bis.org/ifc/publ/ifcb67_12.pdf

Nguyen, K., & La Cava, G. (2020). News sentiment and the economy. *Reserve Bank of Australia (RBA) Bulletin - June 2020*, 18–25. https://www.rba.gov.au/publications/bulletin/2020/jun/pdf/news-sentiment-and-the-economy.pdf

Shapiro, A. H., Sudhof, M., & Wilson, D. (2020). Measuring news sentiment. *Federal Reserve Bank of San Francisco Working Paper 2017-01*. https://doi.org/10.24148/wp2017-01

Article in Press