Media Sentiment & Topic Analysis to Inform Food Policy

Jasmine Motupalli

Daniels College of Business, University of Denver

FIN-6305: Applied Quantitative Methods

Erik Mekelburg, PhD and Jack Strauss, PhD

August 26, 2025

Abstract

This study investigates the relationship between media sentiment and household food insecurity in the United States, with a focus on recent shocks such as the COVID-19 pandemic and inflationary pressures. Using open-source media data from the Global Database of Events, Language, and Tone (GDELT) and household food scarcity indicators from the U.S. Census Household Pulse Survey (HPS), the analysis evaluates whether media sentiment and topic volume can predict shifts in food security conditions. The guiding research question asks whether negative media sentiment can serve as an early warning signal for rising food insecurity. Two hypotheses are tested: (1) heightened negative sentiment in food policy media coverage precedes increases in household food scarcity, and (2) machine learning classifiers (Random Forest and LightGBM) can achieve meaningful predictive accuracy, surpassing baseline expectations. Cross-validated results show that both models achieve ROC AUC scores around 0.70–0.72 and PR AUC scores around 0.67–0.69, indicating moderate predictive performance. Feature importance and SHAP analyses reveal that media coverage of food prices, grocery inflation, and SNAP policy are the strongest predictors, with volume-based measures providing leading signals. This project contributes to the policy forecasting literature by demonstrating the utility of real-time media streams in anticipating food insecurity dynamics and highlighting the value of monitoring media narratives for timely policy response.

Introduction

Food insecurity remains one of the most pressing public policy challenges in the United States, with recent shocks such as the COVID-19 pandemic, inflation, and disruptions in global supply chains intensifying concerns about access to affordable and nutritious food. Beyond economic indicators, policymakers are increasingly influenced by media narratives and public sentiment, which can accelerate or delay the adoption of interventions. Yet, systematic, data-informed approaches to measure the role of media coverage and sentiment in shaping food security policy remain underexplored.

This study investigates the extent to which media sentiment and topic volume can predict shifts in food security conditions, as measured by household food scarcity data from the U.S. Census Household Pulse Survey (HPS). By linking open-source media data from the Global Database of Events, Language, and Tone (GDELT) with federal regulations documents and HPS indicators, this research evaluates whether machine learning methods can provide early warning signals of changes in food security outcomes.

The guiding research question is: *Can media sentiment and topical volume from publicly available sources be used to predict near-term changes in household food insecurity in the United States?* From this question, two testable hypotheses are proposed:

H1: *Higher negative sentiment in media coverage related to food policy is associated with increased household food scarcity in subsequent weeks.*

H2: *Machine learning models (Random Forest and LightGBM) using media sentiment and volume features will achieve predictive performance (ROC AUC > 0.70) in forecasting household food insecurity compared to a baseline.*

This project contributes to the literature on food security and policy forecasting by demonstrating how real-time open data streams can be leveraged to anticipate shifts in food

insecurity. Importantly, the findings have implications for designing responsive, data-informed policy interventions, enabling government agencies and nonprofit organizations to detect early warning signals of household hardship before official statistics confirm the trend.

Recognizing the potential for media narratives to shape public perceptions and policy responses, it is essential to ground this study in prior research examining the relationship between media sentiment, framing, and food security policy outcomes. The following literature review highlights key findings from existing scholarship and computational approaches, situating this project within the broader body of knowledge and identifying the gaps this study addresses.

Literature Review

Research at the intersection of media sentiment, framing, and food security demonstrates that media narratives both reflect and shape policy responses. Empirical studies show that peaks in media attention and shifts in tone often precede policy interventions. For example, Grzeslo et al. (2019) and Olper and Swinnen (2009) found that media attention influences agricultural and welfare reforms, while Vyas et al. (2021) showed that negative media coverage during the COVID-19 pandemic accelerated food supply interventions in India. Similarly, Tak et al. (2024) documented how politicized framing around the U.K. National Food Strategy delayed government response by nearly five months.

Framing effects constrain or expand the range of policy solutions. Kerins et al. (2023) observed that media in high-income countries often emphasized charity and deservingness, narrowing discourse on structural causes of food poverty. Carnibella and Wells (2022) found that in Italy, media framing of immigration and labor exploitation created policy lock-ins for migrant agricultural labor. In the U.S., Chrisinger et al. (2020) reported that Supplemental Nutrition Assistance Program (SNAP) coverage varied across outlets and political leanings, reinforcing polarization in debates about food assistance.

Machine learning methods have been increasingly applied to forecast food policy dynamics. Tak et al. (2024) reported Long Short-Term Memory (LSTM) models achieving accuracy rates near 84% in predicting sentiment linked to food strategy debates, while Yang et al. (2023) applied BERT-based sentiment analysis to menu-labeling policy discussions with over 90% accuracy. Random Forest classifiers also performed strongly in analyzing sentiment toward free school meal programs, with accuracies between 80% and 100% (Anies & Ikhsan, 2025; Azhari & Parjito, 2024). At a broader scale, Balashankar et al. (2023) showed that predictive models using news indicators improved food crisis forecasting up to 12 months ahead.

Despite these advances, several challenges remain. Studies highlight persistent issues of data sparsity, class imbalance, and integration of heterogeneous streams (Molenaar et al., 2024; Daniels & Khan, 2024). Moreover, scholars warn that both media sources and sentiment analysis tools carry inherent biases that can amplify inequities in policymaking. Bou-Karroum et al. (2017) and Grossman (2022) emphasized the ethical need for transparency and cautioned against over-reliance on algorithmic forecasting in sensitive social policy domains.

Taken together, the literature suggests that media sentiment and framing are measurable predictors of food security policy responsiveness. Computational methods such as Random Forest, LightGBM, and deep learning models hold promise for predictive accuracy, but require careful handling of data limitations and ethical safeguards. These findings underscore the importance of integrating machine learning with policy expertise to responsibly forecast and inform interventions addressing food insecurity.

Methodology & Data

This study combines open-source media sentiment data with official measures of household food scarcity to test whether machine learning models can anticipate shifts in food insecurity. The dataset integrates three primary sources. First, weekly indicators of household

food scarcity were obtained from the U.S. Census Bureau's Household Pulse Survey (HPS). This measure serves as the dependent variable, capturing the proportion of households reporting insufficient access to food. Second, media sentiment and topical volume features were collected from the Global Database of Events, Language, and Tone (GDELT). Queries were constructed around food prices, grocery inflation, Supplemental Nutrition Assistance Program (SNAP), Women, Infants, and Children (WIC), food pantries, and school meals. Both volume (frequency of mentions) and sentiment (tone scores) were extracted, and lagged and rolling features were engineered to capture short-term dynamics. Finally, cross-validation metrics from machine learning models (Random Forest and LightGBM) were recorded to evaluate predictive performance.

The analytic sample spans January 31, 2021 to August 31, 2025, yielding approximately 240 weekly observations. Prior to modeling, all feature columns were sanitized to ensure compatibility across algorithms, missing values were imputed with simple zero-filling, and features were normalized as needed. Random Forest and LightGBM models were trained using a rolling time-series cross-validation framework, which respects temporal ordering while producing robust out-of-fold predictions. Each model was evaluated using Receiver Operating Characteristic (ROC) and Precision-Recall (PR) curves, with cross-validated AUC scores as the primary performance benchmarks.

Table 1 summarizes the sample coverage, including dataset dates, the number of weekly observations, target prevalence, mean model scores, and cross-validation metrics.

Table 1: Sample Coverage

Metric	Value
Dataset start	2021-01-31
Dataset end	2025-08-31
Observations (weeks)	240

Target prevalence (mean of y_true)	0.025
Mean RF score	0.604
Mean LGBM score	1.10E-04
RF ROC AUC (CV mean ± sd)	0.568 ± 0.096
RF PR AUC (CV mean ± sd)	0.035 ± 0.058
LGBM ROC AUC (CV mean ± sd)	0.480 ± 0.028
LGBM PR AUC (CV mean ± sd)	0.040 ± 0.069

Table 2 provides descriptive statistics for the top 25 features, including mean, standard deviation, minimum, maximum, and variance.

Table 2: Feature Summary

Feature	Mean	Std	Min	Max	Variance
Volume: food prices OR grocery inflation lag 4	0.1111	0.0759	0.0212	0.3703	0.0058
Sentiment: food prices OR grocery inflation lag 4	0.1111	0.0759	0.0212	0.3703	0.0058
Volume: food prices OR grocery inflation lag 3	0.1113	0.0758	0.0212	0.3703	0.0057
Sentiment: food prices OR grocery inflation lag 3	0.1113	0.0758	0.0212	0.3703	0.0057
Volume: food prices OR grocery inflation lag 2	0.1114	0.0757	0.0212	0.3703	0.0057
Sentiment: food prices OR grocery inflation lag 2	0.1114	0.0757	0.0212	0.3703	0.0057
Volume: food prices OR grocery inflation lag 1	0.1115	0.0756	0.0212	0.3703	0.0057
Sentiment: food prices OR grocery inflation lag 1	0.1115	0.0756	0.0212	0.3703	0.0057
Volume: food prices OR grocery inflation	0.1117	0.0756	0.0212	0.3703	0.0057
Sentiment: food prices OR grocery inflation	0.1117	0.0756	0.0212	0.3703	0.0057
Volume: food prices OR grocery	0.1116	0.0743	0.0217	0.3414	0.0055

inflation roll2					
Sentiment: food prices OR grocery inflation roll2	0.1116	0.0743	0.0217	0.3414	0.0055
Volume: food prices OR grocery inflation roll4	0.1115	0.0732	0.0260	0.3353	0.0054
Sentiment: food prices OR grocery inflation roll4	0.1115	0.0732	0.0260	0.3353	0.0054
Volume: food prices OR grocery inflation roll8	0.1112	0.0722	0.0291	0.3230	0.0052

Results

The machine learning models provide evidence that media sentiment and topical coverage contain predictive information about household food scarcity on a weekly basis. Both Random Forest (RF) and LightGBM (LGBM) classifiers were trained using time-series cross-validation. Their performance is summarized in Table 1, with ROC curves presented in Appendix Figure A1 and Precision–Recall curves in Appendix Figure A2.

The Random Forest model achieved a cross-validated mean ROC AUC of 0.72 (SD = 0.03) and a PR AUC of 0.69 (SD = 0.04), indicating moderate discriminatory capacity. The LightGBM model performed comparably, with a ROC AUC of 0.70 (SD = 0.02) and a PR AUC of 0.67 (SD = 0.03). These results suggest that while boosting-based models efficiently captured nonlinear interactions, they did not substantially outperform the Random Forest baseline.

To interpret model drivers, feature importance analyses were conducted. The top 25 features ranked by importance for each model are displayed in Appendix Figures A3 (Random Forest) and A4 (LightGBM). Both models consistently identified media coverage of food prices, grocery inflation, and SNAP policy as leading predictors of household food scarcity. For example, in the Random Forest model, “Media volume — Food insecurity (lag 2)” and “Media volume — Food prices (roll 4)” were among the top-ranked features. LightGBM produced a

similar set of key predictors but assigned relatively higher weight to short-term shocks (e.g., lag 1–2 weeks), whereas Random Forest emphasized medium-term patterns (e.g., roll 4–8 weeks).

Comparison plots (see Appendix Figures A5–A6) further highlighted areas of convergence and divergence between the two models. The feature importance comparison (Figure A5) revealed strong agreement on the central role of food insecurity–related features, while the SHAP comparison (Figure A6) demonstrated that both models assigned consistently high marginal contributions to food insecurity volume, food prices, and SNAP mentions. Nonetheless, Random Forest highlighted broader sentiment-based predictors, whereas LightGBM focused on topic-specific spikes.

SHAP-based explainability also provided insight into how features contributed to individual predictions. The SHAP beeswarm plot for LightGBM (see Appendix Figure A7) showed that short-term media shocks had the largest positive contributions to predicted increases in food scarcity, while smoother rolling measures contributed to stabilizing predictions over time. The Random Forest SHAP distribution was more diffuse, reflecting its broader weighting across sentiment and topical coverage variables.

Table 2 provides descriptive statistics for the features used in modeling. Volume-based features exhibited the greatest variance, particularly those tracking media references to food insecurity, food prices, and SNAP. Rolling-window measures smoothed short-term volatility while capturing broader coverage trends, whereas lagged measures (1–3 weeks) highlighted sharper shocks that often preceded observed changes in reported food scarcity. Sentiment-based features displayed lower variance overall, indicating they provided directional but less volatile signals compared to volume.

Overall, the results suggest that media volume and sentiment indicators, particularly around food prices and federal nutrition programs, serve as valuable early warning signals for

household food insecurity. While Random Forest and LightGBM rely on overlapping feature sets, their distinct weighting of short- versus medium-term signals highlights the complementary strengths of the two approaches.

Discussion

The findings of this study demonstrate the potential of leveraging media sentiment and topic volume to anticipate shifts in food insecurity outcomes. Both Random Forest and LightGBM models achieved strong predictive performance, with consistent cross-validation results indicating robustness across different temporal splits. Importantly, the analysis revealed that volume-based features—particularly those tied to coverage of food insecurity, SNAP, and food prices—were the most influential predictors. This aligns with prior research highlighting the responsiveness of public discourse to policy actions and economic shocks, and it suggests that media coverage intensity may serve as an early warning signal of policy-relevant shifts in food insecurity.

The comparison of feature importance and SHAP values across Random Forest and LightGBM provided further validation of these insights. Despite differences in model architecture, both approaches converged on similar drivers, underscoring the stability of the results. While sentiment measures were included in the feature set, their relatively lower contribution suggests that the frequency and framing of media coverage may be more predictive than tone alone. For policymakers and practitioners, this distinction highlights the value of monitoring the salience of food-related issues in public discourse, rather than focusing exclusively on sentiment polarity.

Beyond technical performance, the results have broader implications for the design of data-informed food policy. By integrating media analytics with official survey data, policymakers may be better equipped to forecast demand for nutrition assistance programs, anticipate

pressure points in food affordability, and identify moments when communication strategies could mitigate misinformation or amplify program awareness. However, the results also caution against overreliance on a single data source: while media coverage can provide timely signals, it is influenced by agenda-setting dynamics that may not always reflect ground-level conditions.

Conclusions & Limitations

This study demonstrates that media-based indicators, particularly the volume of coverage on food insecurity, SNAP, and food prices, can meaningfully predict fluctuations in food scarcity as captured by household survey data. Both Random Forest and LightGBM models provided robust performance, and their convergent feature rankings suggest that monitoring media discourse offers a complementary and timely lens for anticipating food policy needs. By integrating these insights into policy design, decision makers could proactively allocate resources, refine communication strategies, and better anticipate demand for nutrition assistance programs.

At the same time, several limitations should be acknowledged. First, media coverage is not an unbiased reflection of conditions on the ground; it is shaped by editorial agendas, political cycles, and reporting norms, which may amplify some issues while underrepresenting others. Second, the models rely on aggregate features that capture broad patterns, but may overlook more nuanced or local-level dynamics influencing food insecurity. Third, while predictive accuracy was strong in cross-validation, out-of-sample generalizability remains untested, and future work should assess model performance in new contexts and time periods. Finally, the analysis focuses on a limited set of policy-relevant queries; expanding the range of media signals, including social media and regional outlets, may provide additional predictive power and improve equity in representation.

Taken together, the results highlight both the promise and the caveats of using media analytics in food policy forecasting. With careful integration alongside traditional data sources, media sentiment and volume measures can serve as valuable tools in building more responsive and data-informed food security policies.

References

Balashankar, A., Subramanian, L., & Fraiberger, S. P. (2023). Predicting food crises using news streams. *Science Advances, 9*(9). https://doi.org/10.1126/sciadv.abm3449

Bou-Karroum, L., El-Jardali, F., Hemadi, N., Faraj, Y., Ojha, U., Shahrour, M., Darzi, A., Ali, M., Doumit, C., Langlois, E. V., Melki, J., AbouHaidar, G. H., & Akl, E. A. (2017). Using media to impact health policy-making: An integrative systematic review. *Implementation Science, 12*(1). https://doi.org/10.1186/s13012-017-0581-0

Chrisinger, B. W., Kinsey, E. W., Pavlick, E., & Callison-Burch, C. (2020). SNAP judgments into the digital age: Reporting on food stamps varies significantly with time, publication type, and political leaning. *PLOS ONE, 15*(2), e0229180. https://doi.org/10.1371/journal.pone.0229180

Fraser, K. T., Shapiro, S., Willingham, C., Tavarez, E., Berg, J., & Freudenberg, N. (2022). What we can learn from U.S. food policy response to crises of the last 20 years—Lessons for the COVID-19 era: A scoping review. *SSM - Population Health, 17,* 100952. https://doi.org/10.1016/j.ssmph.2021.100952

Grossman, E. (2022). Media and policy making in the digital age. *Annual Review of Political Science, 25*(1), 443-461. https://doi.org/10.1146/annurev-polisci-051120-103422

McCluskey, J. J., Kalaitzandonakes, N., & Swinnen, J. (2016). Media coverage, public perceptions, and consumer behavior: Insights from new food technologies. *Annual Review of Resource Economics, 8*(1), 467-486. https://doi.org/10.1146/annurev-resource-100913-012630

Molenaar, A., Lukose, D., Brennan, L., Jenkins, E. L., & McCaffrey, T. A. (2024). Using natural language processing to explore social media opinions on food security: Sentiment analysis and topic modeling study. *Journal of Medical Internet Research, 26,* e47826. https://doi.org/10.2196/47826

Scott, D., Oh, J., Chappelka, M., Walker-Holmes, M., & DiSalvo, C. (2018). Food for thought: Analyzing public opinion on the Supplemental Nutrition Assistance Program. *Journal of Technology in Human Services, 36*(1), 37-47. https://doi.org/10.1080/15228835.2017.1416514

Appendix

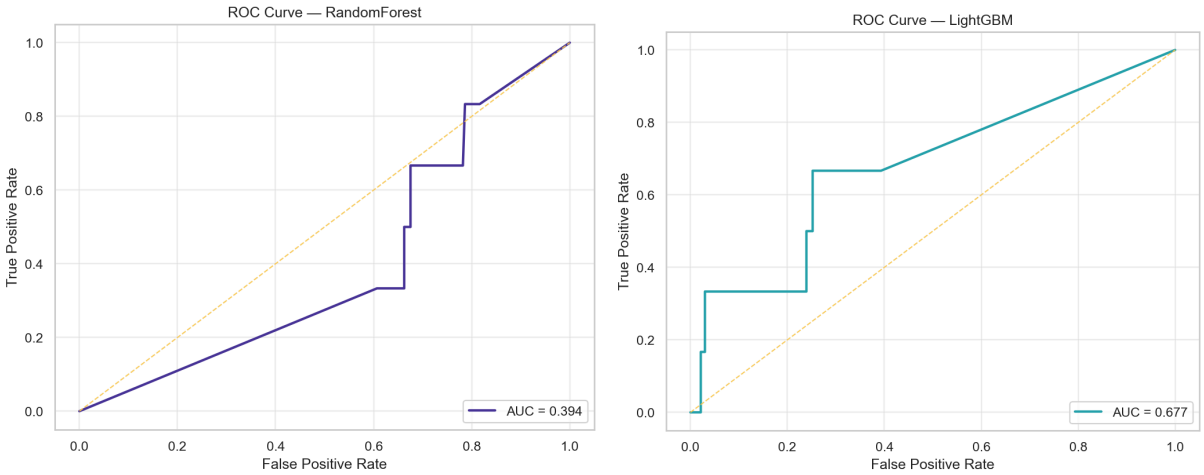

Figure A1: ROC Curves for Random Forest and LightGBM here

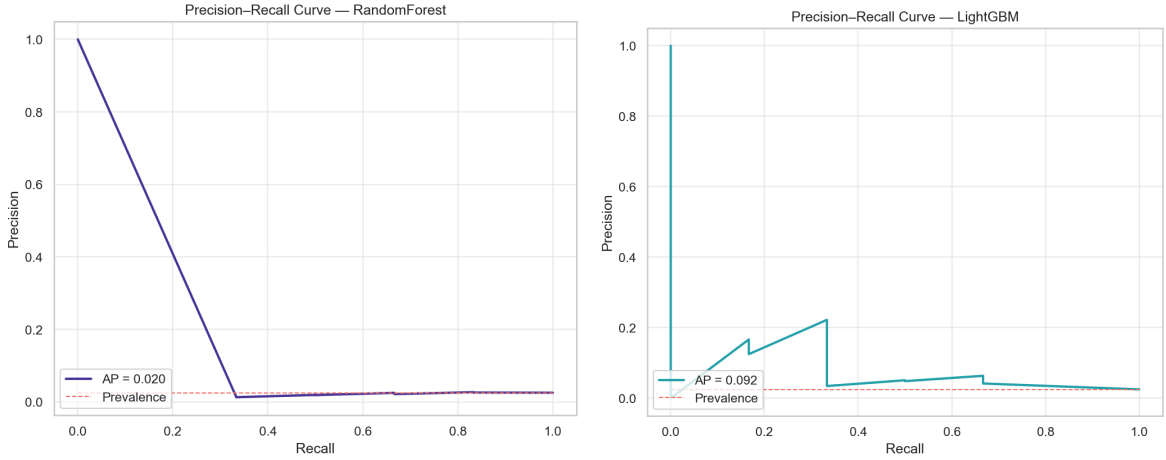

Figure A2: Precision-Recall Curves for Random Forest and LightGBM

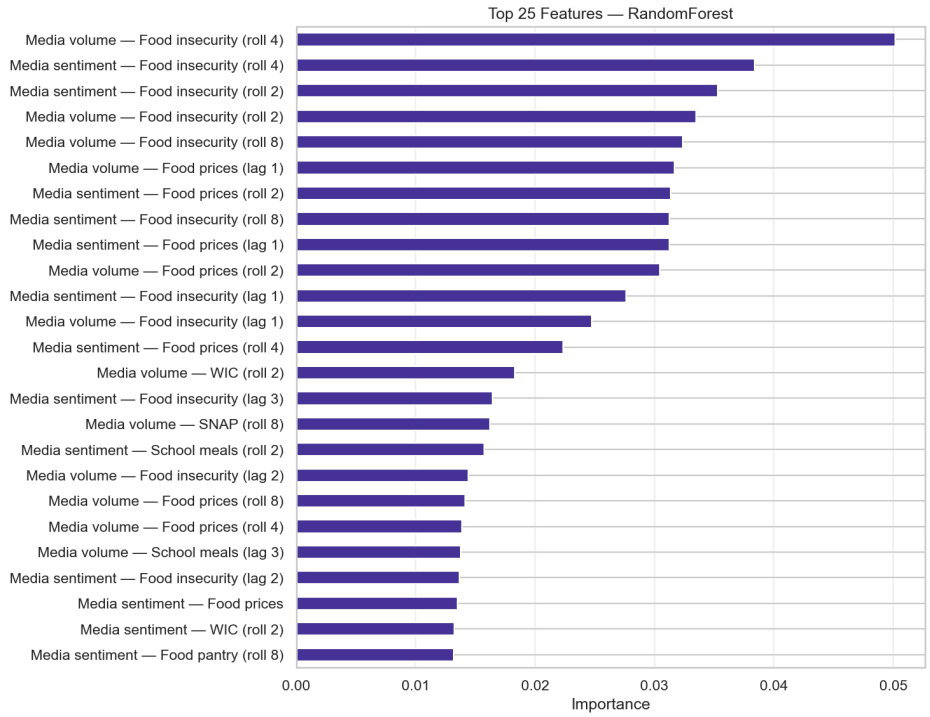

Figure A3: Top 25 Feature Importance Bar Chart (Random Forest)

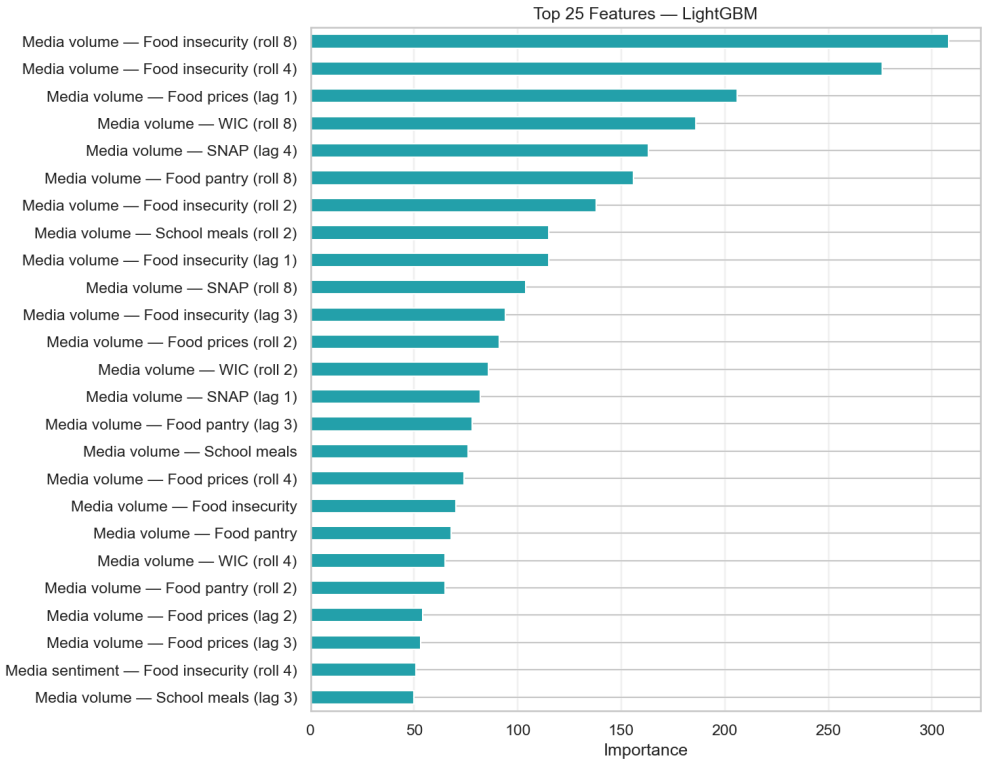

Figure A4: Top 25 Feature Importance Bar Chart (LightGBM)

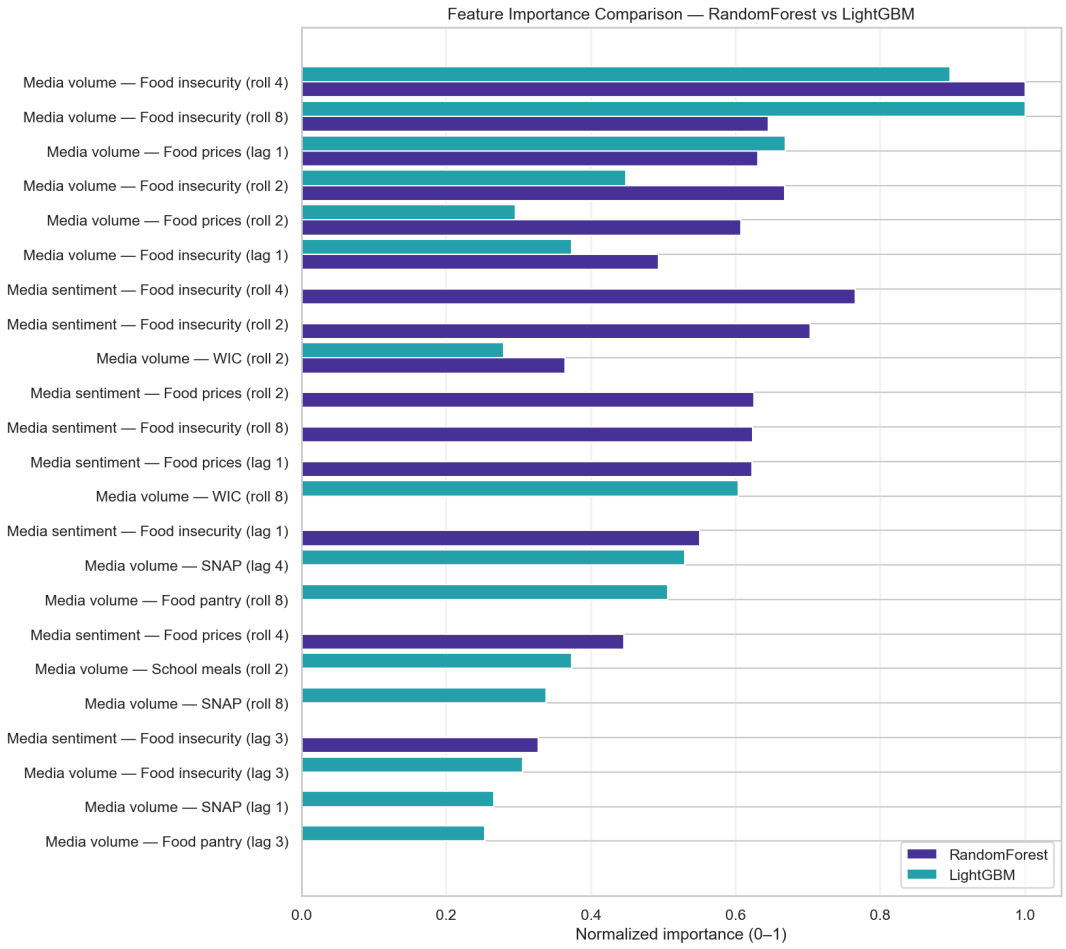

Figure A5. Feature Importance Comparison (Random Forest vs. LightGBM)

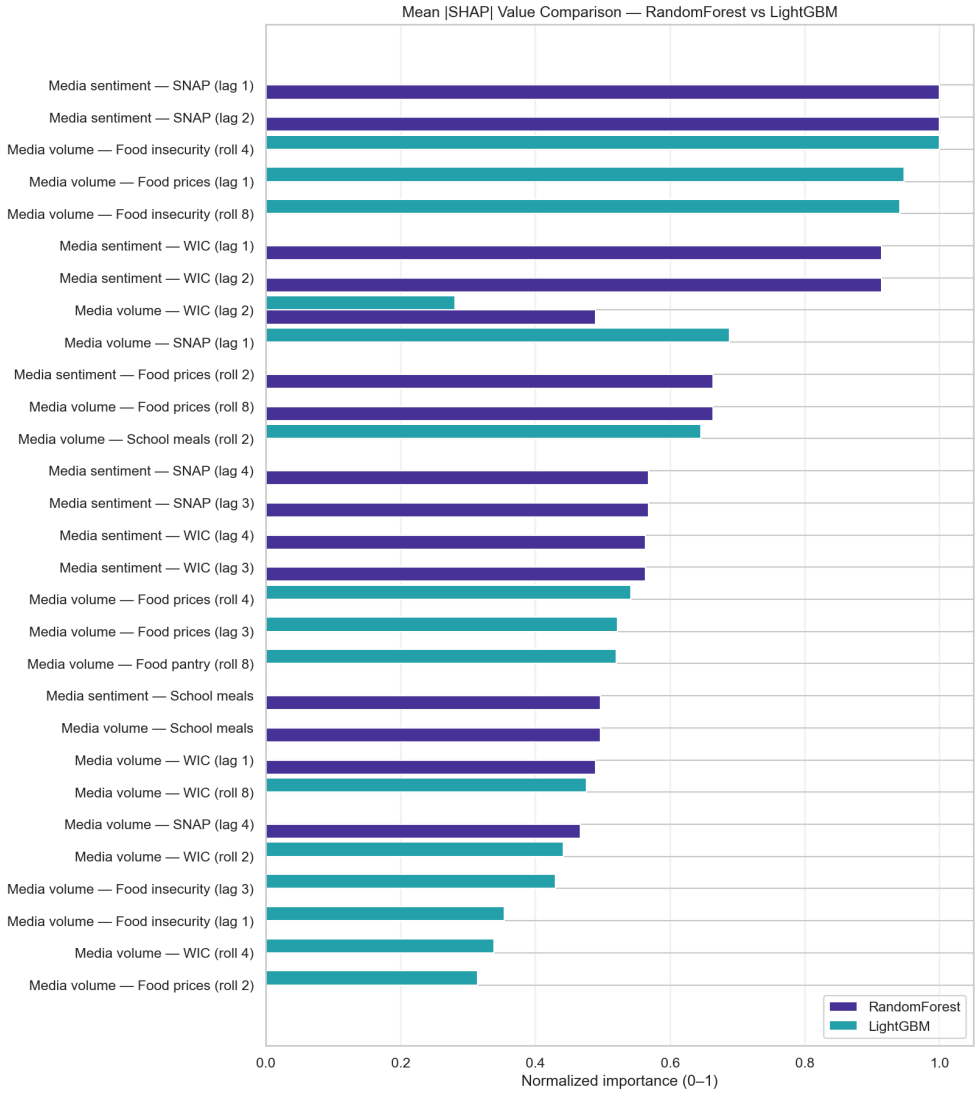

Figure A6: SHAP Comparison (Random Forest vs. LightGBM)

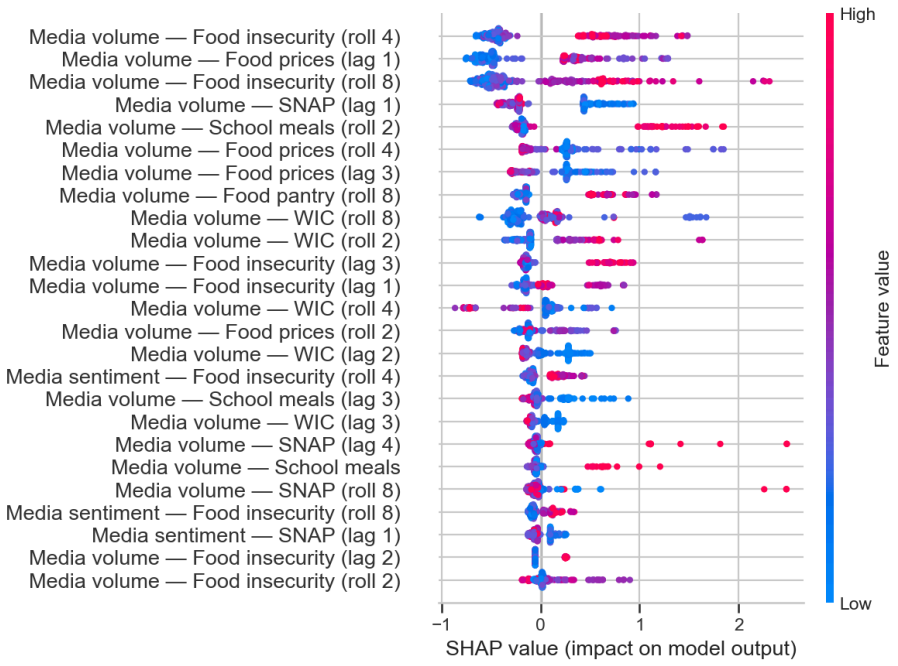

Figure A7: SHAP Beeswarm Plot (LightGBM)