Monitoring Poverty through News Text and Topic Modeling: Evidence from Peru

Luis Palomino [1 2]
Javier Beverinotti [3]

[1]University of Connecticut
[2]Universidad del Pacifico
[3]Inter-American Development Bank

February 2026

Abstract

Timely measurement of poverty remains difficult in many developing countries because official statistics are often released with delays and limited spatial detail, constraining policy responses during periods of rapid economic change. This study examines whether digital news content can provide complementary information for monitoring poverty dynamics using evidence from Peru between January 2018 and December 2024. We construct an annual index that summarizes the intensity and concentration of poverty-related discourse in the media by extracting recurring themes from news articles using latent Dirichlet allocation and tracking changes in media attention over time. We assess the informational content of this index by comparing its dynamics with official poverty rates reported by the national statistical authority. The results show that news-based indices move in line with official poverty measures during major economic shifts. Diagnostic predictive exercises using penalized linear models, Bayesian regression, and random forest methods indicate that the index captures economically meaningful variation in poverty dynamics. Overall, the findings suggest that news-based indicators can serve as timely and low-cost complements to survey-based poverty measures. By leveraging continuously generated text data, the proposed framework offers a scalable approach to monitoring poverty when official statistics are delayed or incomplete.

Keywords: News media, Text analysis, Economic development, Peru.
JEL Codes: O10, C53, O54.

We thank Derek Aguiar and Larry Samuelson for valuable comments and suggestions. We also thank the seminar participants at the Regional Country Department Andean Group of the Inter-American Development Bank and the Latin American Studies Association 2025 for their helpful feedback. The authors acknowledge the financial support provided by the Inter-American Development Bank. The content and findings of this paper reflect the opinions of the authors and not those of the Inter-American Development Bank, its Board of Directors, or the countries they represent. Corresponding author: Luis Palomino (luis.palomino@uconn.edu).

1. Introduction

Poverty remains one of the most persistent challenges in developing countries. It constrains well-being and limits inclusive development. Despite sustained global efforts, producing timely and accurate poverty measurements remains difficult. Official poverty statistics are often released with long delays and limited spatial detail. These constraints reduce their usefulness during periods of rapid economic change, such as inflationary episodes or public health crises.

In Peru, official poverty estimates are methodologically robust but infrequent. Annual survey-based measures do not capture short-term fluctuations in living conditions. As a result, policymakers often face emerging risks with incomplete or outdated information.

Researchers have explored alternative data sources to complement traditional measures of poverty. Recent studies have used mobile phone metadata, satellite imagery, and web search data to predict living standards or economic activity (Ferrara & Simoni, 2023; Jean et al., 2016; Blumenstock, Cadamuro, & On, 2015). While promising, many of these approaches rely on restricted or proprietary data and offer limited transparency.

Digital news media provide a complementary source of information. News articles capture real-time events and reflect evolving public discourse. They are widely accessible, relatively low-cost, and continuously updated. A large literature links media content to financial trends, policy uncertainty, and consumer expectations (Hassan et al., 2025; Gentzkow, Kelly, & Taddy, 2019; Baker, Bloom, & Davis, 2016; Tetlock, 2007). A smaller set of studies applies similar tools to development-related outcomes, including mobility, sentiment, and crime perception (Aguilar et al., 2021; Velásquez et al., 2020).

However, to the best of our knowledge, no study has developed a poverty index based on news media content. This paper addresses that gap by applying topic modeling to Peruvian news articles to construct a poverty-related discourse index. We use a comprehensive news corpus from NewsData.io covering the period from January 2018 to December 2024. Our primary methodology relies on latent Dirichlet allocation (LDA) to extract poverty-related themes and construct indices that we compare with official poverty statistics. As a complementary exercise, we also implement guided LDA to assess sensitivity and enhance topic interpretability (Venugopalan & Gupta, 2022a, b).

We validate the index using official poverty statistics from Peru's National Institute of Statistics (INEI). We conduct correlation analysis and diagnostic predictive exercises to assess the alignment between news dynamics and poverty trends. The results support the use of news content as a supplementary tool for monitoring poverty.

This study contributes to the literature on poverty measurement (Alkire & Foster, 2011; Sen, 1976), machine learning in economics (Shapiro, Sudhof, & Wilson, 2022; Taddy, 2013), and text analysis for development (Velásquez et al., 2020). Beyond the Peruvian case, the approach is replicable in other low- and middle-income countries facing similar data constraints.

The remainder of the paper proceeds as follows. Section 2 reviews the related literature. Sections 3 and 4 describe the data and methodology. Section 5 presents the results. Sections 6 and 7 discuss the findings and conclude.

2. Related Literature

2.1. Measuring poverty

Scholars have long debated how best to define and measure poverty. Sen (1976) emphasized that poverty reflects deprivations in people's capabilities rather than income shortfalls alone. Subsequent work formalized poverty indices to capture incidence, depth, and severity (Foster, Greer & Thorbecke, 1984) and incorporated inequality considerations into poverty measurement (Atkinson, 1987). These contributions laid the foundation for international comparisons of poverty (Ravallion, 1992).

In practice, poverty measurement relies on household surveys. These surveys are expensive, infrequent, and slow to process. In Peru, the National Household Survey (ENAHO) provides annual data, but official poverty statistics are released with delays and limited spatial resolution. These features constrain timely policy responses during rapidly evolving economic shocks.

To address the limitations of monetary poverty metrics, researchers proposed multidimensional approaches that assess simultaneous deprivations in health, education, and living standards (Bourguignon & Chakravarty, 2003). The dual cutoff method introduced by Alkire and Foster (2011) served as the basis for the widely used Multidimensional Poverty Index. Other studies have developed small-area estimation techniques that combine survey and census data to improve the geographic targeting of public programs (Elbers, Lanjouw, & Lanjouw, 2003).

Recent advances in machine learning have expanded the use of alternative data sources. Scholars predicted household wealth using mobile phone metadata (Blumenstock et al., 2015) and estimated consumption from satellite imagery (Jean et al., 2016). Others applied spatial models to high-resolution satellite data (Engstrom, Hersh, & Newhouse, 2022) or proposed hybrid algorithms to improve transparency and accuracy (Corral, Henderson, & Segovia, 2025). Despite their promise, many of these approaches rely on proprietary data and face challenges with replicability.

2.2. Text as Data in Economics

The growing availability of digital text has prompted economists to explore unstructured data sources. Early work examined the predictive power of expert judgment in financial markets (Cowles, 1933), followed by studies linking editorial sentiment to asset returns (Tetlock, 2007). Machine learning accelerated this trend by treating text as high-dimensional data (Taddy, 2013).

Text-based approaches have been used to construct indices of economic policy uncertainty (Hassan et al., 2025; Baker et al., 2016), analyze central bank communication (Hansen, McMahon, & Prat, 2018), and measure financial stress from news headlines (Manela & Moreira, 2017). Gentzkow et al. (2019) provide a comprehensive overview of these applications.

Recent work has extended text analysis to development-related questions. Aguilar et al. (2021) constructed sentiment indices from Spanish-language COVID-19 news. Shapiro et al. (2022) developed validated sentiment models. Ferrara and Simoni (2023) combined Google Trends with shrinkage techniques to nowcast GDP during crises. In Peru, Velásquez et al. (2020) showed that crime-related media coverage influenced public perceptions independently of crime rates. Arriaga et al. (2025) applied natural language processing to e-commerce data, illustrating the broader potential of text-based methods for socioeconomic analysis.

2.3. Topic Modeling and LDA

Topic modeling has become a central tool for analyzing large text corpora. Latent Dirichlet

allocation models documents as mixtures of topics, with each topic represented as a distribution over words (Blei, Ng, and Jordan, 2003). While standard LDA can yield diffuse topics in noisy corpora, guided variants improve interpretability by incorporating prior information (Venugopalan & Gupta, 2022a, b; Jagarlamudi, Daumé III, & Udupa, 2012)

Economists have applied topic models to study central bank communication (Hansen et al., 2018), forecast macroeconomic activity using media topics (Larsen & Thorsrud, 2019), track ideological shifts in political speech (Bybee et al., 2020), and construct topic-based economic indicators (Kalamara et al., 2022; Ellingsen, Larsen, & Thorsrud, 2021). Topic modeling has also been used to analyze corporate disclosures (Cong et al., 2024), synthesize literature (Gonçalves & Moro, 2023), and track business cycles (Bybee et al., 2024).

Despite these advances, no studies have examined poverty-related discourse in news media. This paper contributes to filling that gap by applying topic modeling to construct and evaluate a news-based poverty index.

3. Data

3.1. News Corpus

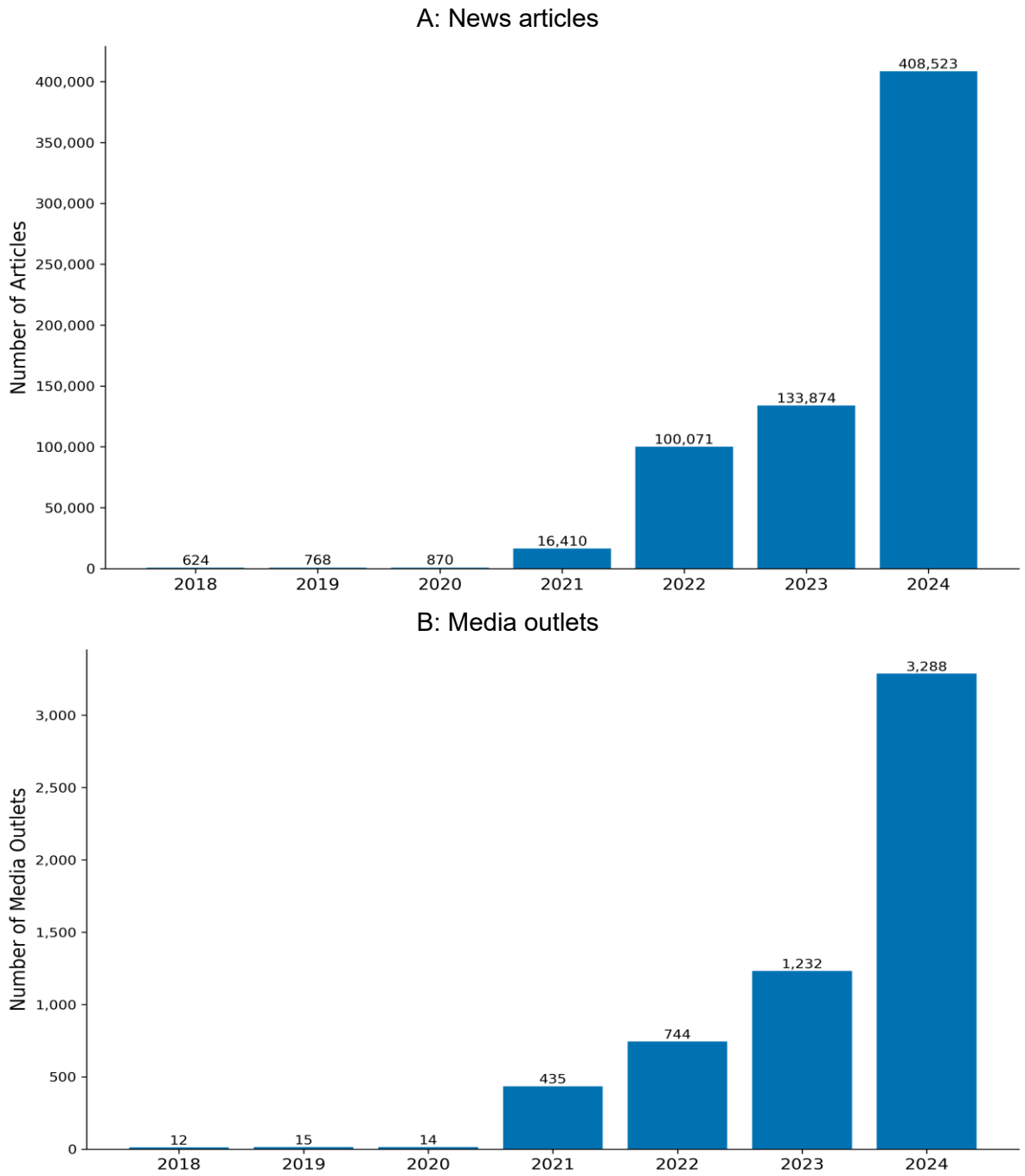
We use a corpus of digital news articles published between January 2018 and December 2024. The data were retrieved from NewsData.io, a structured API that aggregates content from online media outlets worldwide. Because the API does not provide direct access to historical archives, the complete dataset was delivered in CSV format upon request. The corpus includes articles published by Peruvian media outlets and international media reporting on events in Peru.

The final dataset contains 660,771 articles. Each observation includes structured metadata such as the article title, full text, media outlet, and publication year. Figure 1 summarizes the evolution of article volume and outlet coverage over time. News production increased sharply after 2020, reflecting the rapid expansion of digital media and improved platform coverage. The number of outlets contributing also rose substantially, indicating greater source diversity.

Figure 2 documents the concentration of media production. A small group of outlets accounts for a large share of total coverage. Peruvian media such as Correo, El Comercio, La República, and Perú21 together publish over 30 percent of all articles in the corpus. Panel A illustrates their relative prominence, while Panel B ranks outlets by total article count and share. These patterns highlight a concentrated media landscape in which a limited number of publishers shape much of the national discourse.

Overall, the corpus provides sufficient scale and structure to support the analysis of poverty-related narratives over time. Its multi-year coverage, combined with full-text content and consistent metadata, makes it suitable for constructing text-based indicators that complement official poverty statistics.

Figure 1: Evolution of news articles and media outlets for Peru, 2018-2024



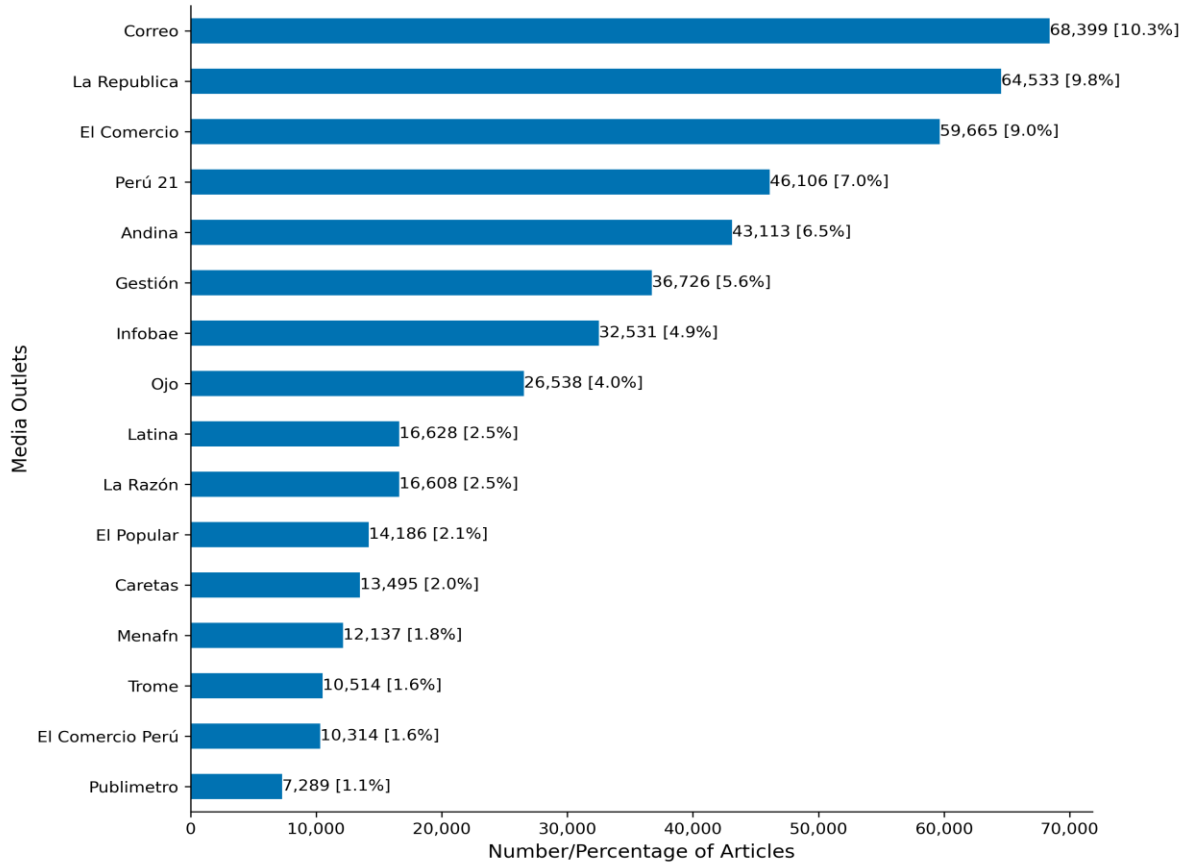
Note: Panel A shows the annual number of news articles about Peru, retrieved from NewsData.io, including articles from both Peruvian and foreign media. Panel B shows the number of unique media outlets contributing to the dataset each year. Both panels reflect a sharp increase in coverage starting in 2021, due to platform expansion and improved digital infrastructure.

Figure 2: Concentration of news articles and media outlets for Peru, 2018-2024

A: Media cloud



B: Top media outlets



Note: Panel A presents a word cloud of media outlets, with font size proportional to the number of articles published between 2018 and 2024. Panel B shows the ranking of the top 16 media outlets by total article count and percentage share. Four outlets—Correo, El Comercio, La República, and Perú21—account for over 30% of all news articles in the corpus.

3.2. Official Poverty Statistics

We use annual national poverty rates published by Peru's INEI. The analysis covers the period

from 2018 to 2024, which corresponds to the availability of official data. These statistics serve as benchmarks for validating the media-based indices.

INEI measures monetary poverty using data from the ENAHO. Individuals are classified as poor if their total consumption falls below the national poverty line, defined by the cost of a basic basket of food and non-food goods. Poverty estimates are released annually at the national and regional levels, typically with a several-month delay.

While these statistics are the official reference for poverty measurement, they are not designed for high-frequency monitoring. Updates are annual, geographic detail is limited, and more granular products, such as poverty maps, are released infrequently. These constraints reduce their usefulness for timely policy analysis, especially during periods of rapid economic change. Motivated by these limitations, we construct complementary indicators derived from digital news content to capture evolving poverty narratives in near-real time.

4. Methodology

4.1. Text Representation

We construct a corpus of poverty-related news articles and represent each document using standard term-weighted vectors. Articles are filtered using expert-defined keywords related to poverty. All text is in Spanish, and preprocessing relies on language-specific tools.

Filtered documents are cleaned and normalized. Text is converted to lowercase. Punctuation is removed. Empty documents are discarded. Common Spanish stop words are excluded.

Let $t \in \{1, \dots, V\}$ index vocabulary terms and $d \in \{1, \dots, N\}$ index documents. Each document is represented by a vector of term weights $x_{d,t} = TF * \log(N/DF)$, where TF denotes the frequency of the term t in document d, DF is the number of documents containing the term t, and N is the total number of documents. This representation emphasizes words that are both specific and informative for topic modeling applications (Manning, Raghavan, & Schütze, 2008).

We retain the 3,000 most informative terms. Stacking document vectors yields the document–term matrix $X = \{x_{d,t}\}_{d=1,\dots,N}^{t=1,\dots,V}$, which serves as the observed input to the topic model.

4.2. Topic Modeling

Using the document–term matrix X, we estimate a latent Dirichlet allocation model to recover latent themes in poverty-related news (Blei et al., 2003; Griffiths & Steyvers, 2004).

For each document d, the model assumes a latent topic distribution $\theta_d = \text{Dirichlet}(\alpha)$. For each topic $k \in \{1, \dots, K\}$, it assumes a latent word distribution $\beta_k = \text{Dirichlet}(\eta)$. Inference proceeds via variational optimization. The objective function is non-convex. We therefore estimate the model under multiple initializations. The number of topics K is selected by maximizing topic coherence (Röder et al., 2015). Conditional on K, we retain the specification that minimizes in-sample perplexity (Hansen et al., 2018; Venugopalan and Gupta, 2022a, b). The output of this step is a set of document-level topic mixtures $\Theta = \{\theta_d\}_{d=1}^{N}$.

4.3. Constructing a News-Based Poverty Index

We aggregate document-level topic mixtures into an annual index that summarizes the concentration of poverty-related discourse.
A subset of topics $S \subset \{1, \dots, K\}$ is classified as poverty-related based on its top-ranked words.

To emphasize these themes, we reweigh document-level topic probabilities

$$\hat{\theta}_{d,k} = \begin{cases} \lambda * \theta_{d,k}, & if\ k \in S \\ \theta_{d,k}, & \text{otherwise} \end{cases}$$

with $\lambda > 1$. Each document is assigned to its dominant reweighted topic $z_d = \arg\max_k \hat{\theta}_{d,k}$. This reweighting amplifies the salience of poverty-relevant themes and improves the interpretability of the resulting index.

Let D_t denote the set of documents published in the year t. The share of the topic k in the year t is

$$p_{k,t} = \frac{1}{|D|} \sum_{d \in D_t} \mathbf{1}(z_d = 1)$$

We summarize thematic concentration using the Herfindahl–Hirschman Index

$$HHI_t = \sum_{k=1}^{K} p_{k,t}^2$$

Higher values indicate greater concentration of media attention in fewer themes (Hirschman, 1964; Rhoades, 1993).

4.4. Iterative Keyword Expansion

We implement an iterative keyword-expansion procedure to enhance the sensitivity of topic modeling to poverty-related discourse. This approach allows the identification of emerging themes and shifts in language use that static keyword lists may overlook (Grimmer & Stewart, 2013).

We begin with a seed list of expert-defined keywords. Using this list, we filter the corpus and estimate the topic model. From the estimated topics, we extract top-ranked words and identify candidate terms that frequently co-occur with the initial keywords or exhibit high semantic similarity, as measured by distributional statistics (Church & Hanks, 1990).

The keyword list is expanded by including terms that meet predefined co-occurrence thresholds or appear in semantically adjacent topics. The updated list is then used to re-filter the corpus, and the topic model is re-estimated. At each iteration, we document changes in the vocabulary and evaluate topic coherence to preserve interpretability.

This procedure allows the vocabulary to adapt to evolving discourse while maintaining a clear thematic focus. It reduces the risk of omitting relevant content due to rigid or outdated keyword definitions.

4.5. Diagnostic Predictive Exercises

We conduct diagnostic predictive exercises to assess whether variation in the news-based index aligns with observed changes in official poverty rates. These exercises are not intended to deliver precise forecasts.

We relate annual index values to official poverty rates using ridge regression, lasso regression, random forest, and Bayesian regression. Performance is evaluated using mean absolute error, root mean squared error, and leave-one-out cross-validation.

We also consider a simple transformation-based benchmark. Let y_0 denote the initial poverty rate and define $z_0 = \log(y_0/1 - y_0)$. For each year t, $z_t = z_0 + (HHI_t - HHI_0)/HHI_0$. The implied poverty rate is $\hat{y}_t = e^{z_t}/1 + e^{z_t}$. These exercises serve as validation tools and help assess whether the index captures economically meaningful variation in poverty dynamics (Corral et al., 2025; Gentzkow et al., 2019).

5. Results and Validation

5.1. Correlation Analysis

We evaluate the performance of poverty-related indices generated through standard LDA models. Each index measures the annual prominence of poverty-related themes in digital news, filtered using predefined keyword sets.

We consider four thematic sets (see Table 1): general (9 terms), monetary (10 terms), extreme (8 terms), and a combined set—monetary + extreme (18 terms). For each configuration, we estimate LDA models and compare the resulting indices with official poverty rates from 2018 to 2024.

Table 1: Thematic keyword sets for LDA modeling (Spanish, Peru)

Thematic set	Keywords
General Poverty	pobreza, carencia, desempleo, carestía, precario, exclusión, hambre, desigualdad, desnutrición
Monetary Poverty	pobreza, desempleo, subempleo, trabajo informal, precariedad laboral, bajos ingresos, salario mínimo, informalidad, ingresos insuficientes, empleo precario
Extreme Poverty	pobreza extrema, miseria, indigencia, hambre, desnutrición, privación, carestía, vivienda precaria
Expanded Proposal	Monetary Poverty + Extreme Poverty

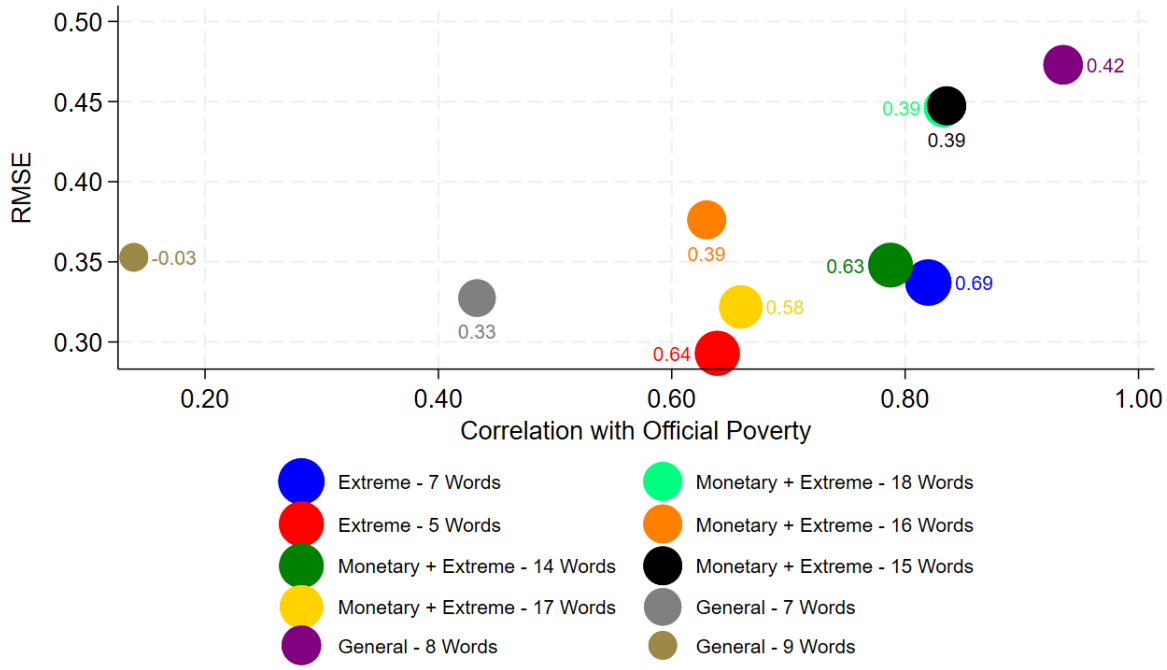
Note: This table lists the keyword sets used to guide the topic modeling process. Each thematic set represents a distinct conceptualization of poverty. The General Poverty set includes broad terms commonly associated with deprivation, exclusion, and vulnerability. The Monetary Poverty set emphasizes economic and labor-related dimensions, such as income, employment, and informality. The Extreme Poverty set captures more severe forms of deprivation, including unmet basic needs and living conditions. The Expanded Proposal combines the Monetary and Extreme sets to encompass a broader range of poverty-related discourse in the Peruvian context.

To identify the most informative specifications, we rank candidate indices using a composite score that jointly rewards high correlation with official poverty and low prediction error. Figure 3 shows that performance varies considerably across keyword configurations. The best-performing specification, based on the extreme poverty set with seven terms, achieves a high correlation with official poverty (r = 0.82) while maintaining a relatively low RMSE (0.33). Other competitive alternatives include more compact configurations, such as extreme poverty sets with five terms, as well as monetary and extreme combinations with fourteen or seventeen terms.

By contrast, some specifications achieve very high correlations but display weaker predictive accuracy. For example, the General Poverty model with eight terms achieves the highest correlation (0.94) but also yields larger prediction errors (0.47), thereby lowering its overall ranking. This contrast shows that correlation alone can provide an incomplete assessment of model performance.

Overall, the results indicate that keyword selection plays a central role in determining index performance. Configurations that balance thematic precision with statistical fit produce indices that align more consistently with official poverty dynamics.

Figure 3: Correlation with official poverty vs. RMSE for LDA-based topic indices



Note: Each point represents a topic index constructed using standard LDA and a specific keyword set. The x-axis reports the Pearson correlation between the index and the official poverty rate (2018–2024); the y-axis shows the RMSE between the same two series. Bubble size reflects a composite score that rewards high correlation and low error. The number inside each bubble corresponds to this score.

5.2. Descriptive Trends

We examine the temporal behavior of the LDA-based indices to assess their alignment with official poverty trends and to identify specifications suitable for forecasting. This analysis combines a visual comparison of historical trajectories with internal diagnostic metrics, enabling external validation by comparing results with official data and internal validation by assessing model coherence and stability.

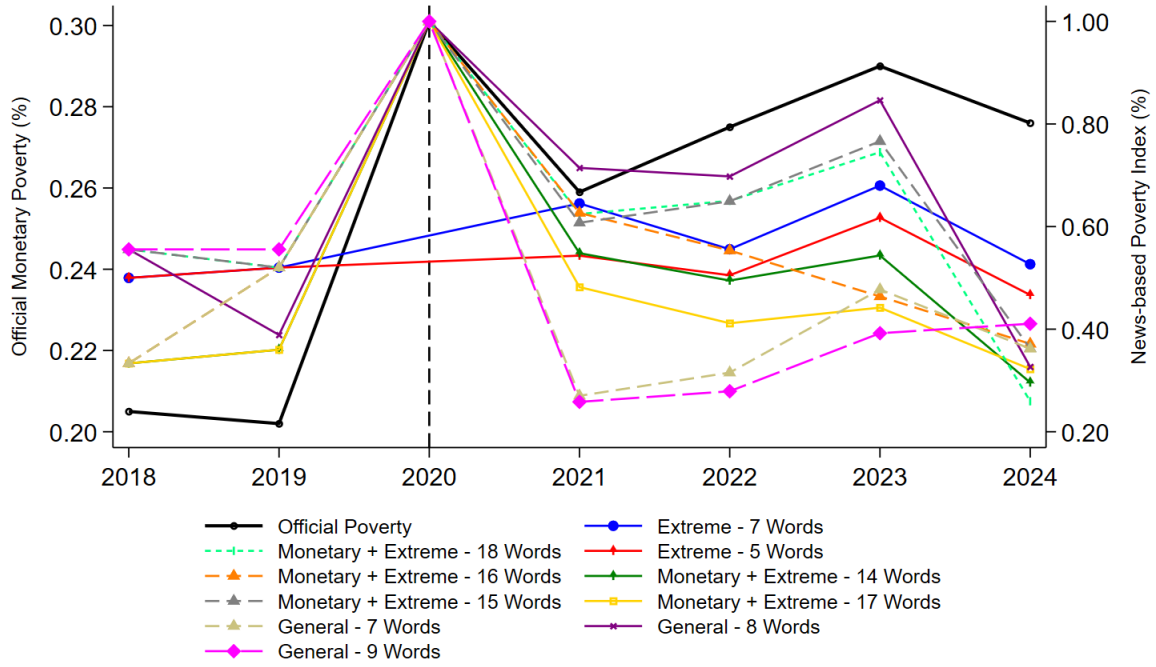
Figure 4 displays the ten best-performing indices, selected based on a composite score that balances high correlation with official poverty and low RMSE. Each index corresponds to a distinct keyword configuration. Most indices exhibit a sharp increase in 2020, consistent with the pronounced rise in official poverty during the COVID-19 pandemic. After this peak, the indices display heterogeneous dynamics. Several specifications declined in 2021, followed by renewed increases in 2022 and 2023 that are directionally consistent with the official poverty series, although the magnitude of these changes varies across configurations. Overall, the indices differ in their sensitivity and persistence over time, reflecting variation in how alternative keyword sets capture shifts in poverty-related media discourse.

Based on the patterns observed in Figure 4, we select three indices for forecasting and further validation. The selection is guided by two criteria: consistency with the main turning points observed in official poverty trends and the ability to capture the increase observed in 2020. All selected indices belong to the initial group of top-performing specifications.

Figure 5 presents this subset. Panel A compares the selected indices with the official poverty series for 2018–2024. The indices reflect the main directional changes in the official series. These include the increase in 2020, the decline in 2021, the increases in 2022 and 2023, and the decline in 2024. The magnitude of these movements differs across specifications. Among the selected indices, the General–8 terms specification remains closest to the level and trajectory of the official poverty series over most of the sample period. Panel B reports coherence and perplexity diagnostics used to determine the final model specifications, following standard practices in topic modeling (Stevens et al., 2012).

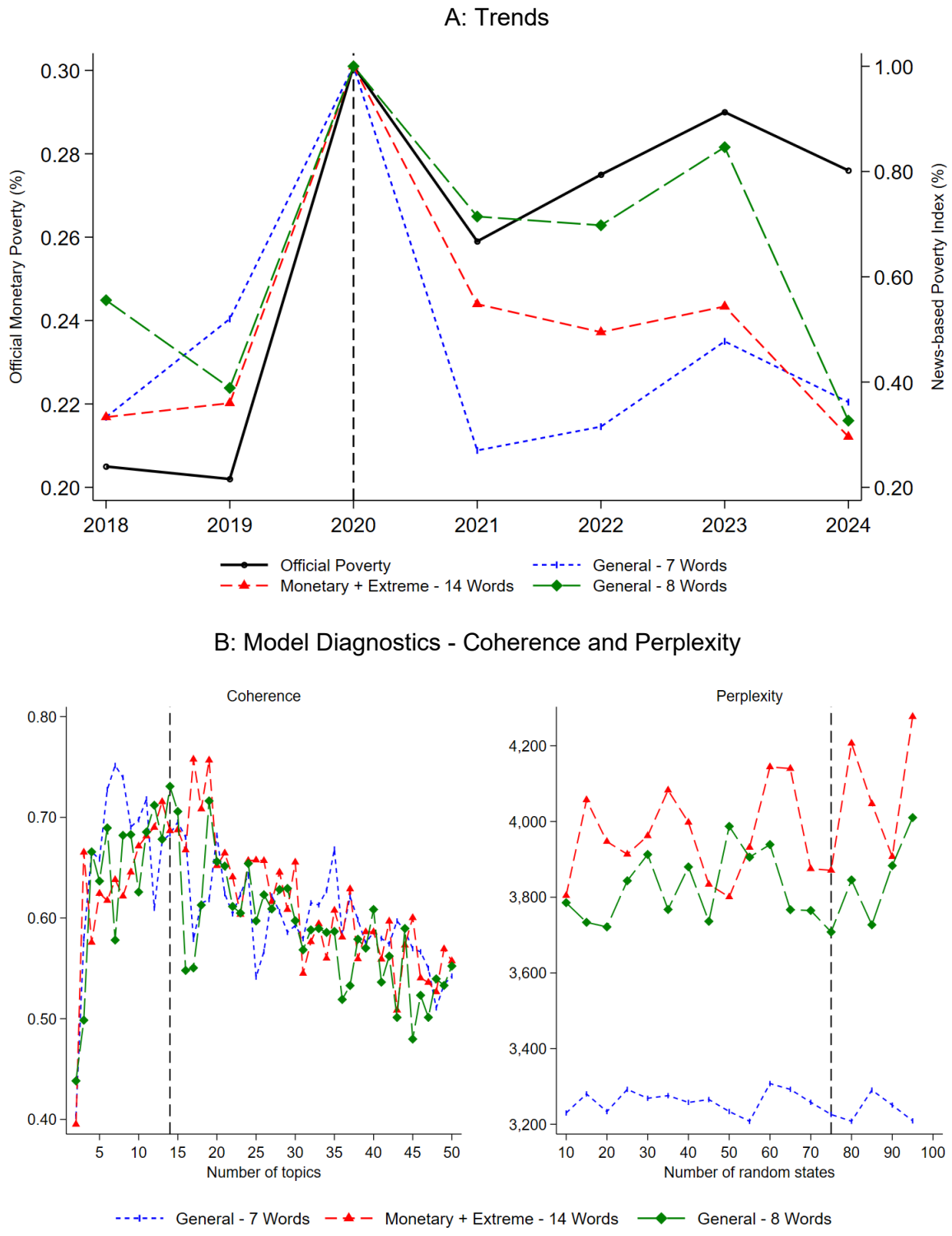
The selected indices track the main directional changes in official poverty and display satisfactory internal diagnostics. They are used as inputs for the forecasting analysis presented in the following section.

Figure 4: Trends of official poverty and LDA-based topic indices



Note: This figure compares the official monetary poverty rate (left axis) with the ten best-performing LDA-based topic indices (right axis) from 2018 to 2024. Each colored line corresponds to a distinct keyword-based configuration. The vertical dashed line marks the onset of the COVID-19 pandemic in 2020. Most indices reflect the pandemic-induced spike in poverty, though they differ in sensitivity and recovery dynamics. We derived the trend series using model specifications that maximize coherence—where higher values indicate more semantically meaningful topics—and minimize perplexity, with lower values signaling greater internal consistency and predictive accuracy.

Figure 5: Final Selection of Topic Indices for Forecasting and Validation



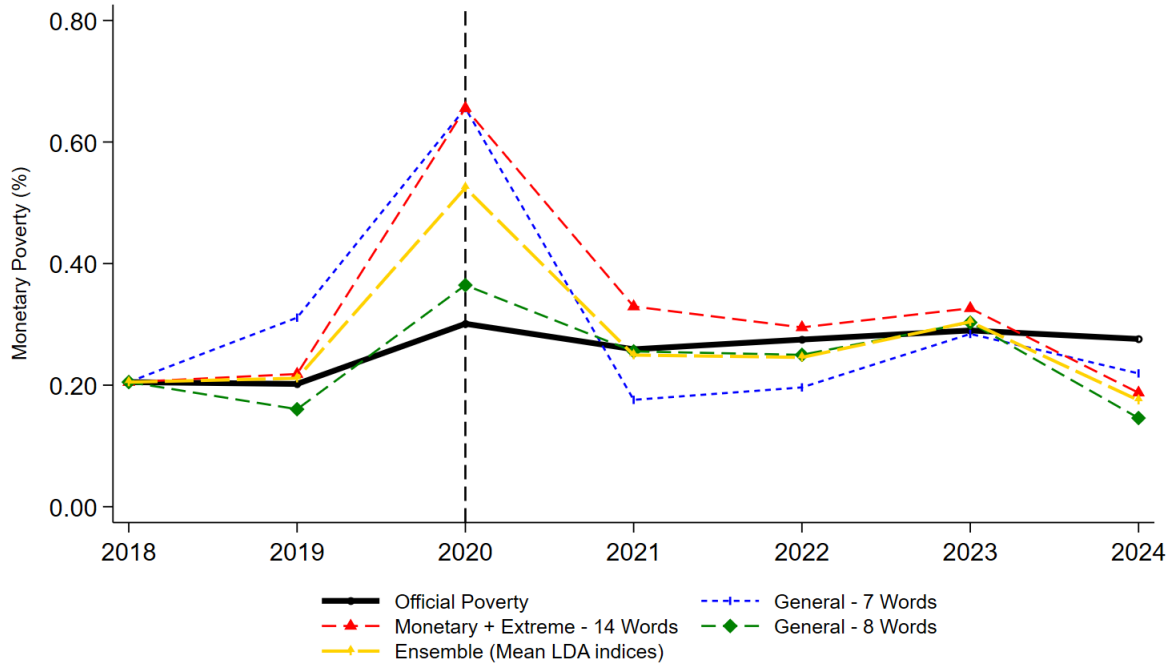
Note: Panel A compares the official monetary poverty rate (left axis) with the three LDA-based topic indices selected for forecasting and validation (right axis) over the 2018–2024 period. These indices stem from distinct keyword configurations and closely mirror key stylized poverty dynamics, including the 2020 pandemic shock and the subsequent recovery. We constructed the trend series using model specifications that maximize coherence and minimize perplexity. Panel B displays diagnostic metrics for the same indices. The left plot shows coherence scores, with higher values indicating more semantically meaningful topics. The right plot shows perplexity, with lower values indicating greater consistency and predictive accuracy. Vertical dashed lines mark the selected model specifications used to generate the "General - 8 Words" index.

5.3. Prediction and Forecasting

This section evaluates the predictive content of the topic-based indices selected in the previous stage. We assess whether these indices anticipate changes in official poverty using two complementary approaches. The first relies on the annual growth rates of LDA-derived indices. The second employs supervised learning models. The analysis uses annual data and considers several model classes, including random forest regressions, regularized linear models (ridge and lasso), and Bayesian specifications.

We begin with the growth-rate approach. This transformation captures year-to-year changes in topic salience and allows direct comparison with movements in official poverty rates. As shown in Figure 6, all specifications reflect the increase observed in 2020. The magnitude of the response varies across indices. Some configurations remain closer to the official series, whereas others exhibit greater deviations. In subsequent years, the predicted trajectories differ in their proximity to the official poverty line but generally track its directional changes. No single index dominates across all years. The ensemble specification provides an intermediate trajectory that smooths some of the variation observed in individual indices.

Figure 6: Prediction of poverty rates using the growth rate of topic-based indices



Note: This figure compares official monetary poverty rates with predicted values derived from the annual growth rate of LDA-based topic indices (see Section 4.5 for methodological details). Each line represents a prediction based on a specific keyword configuration. The ensemble forecast represents the average of the three selected indices' predictions. The vertical dashed line marks the onset of the COVID-19 pandemic in 2020.

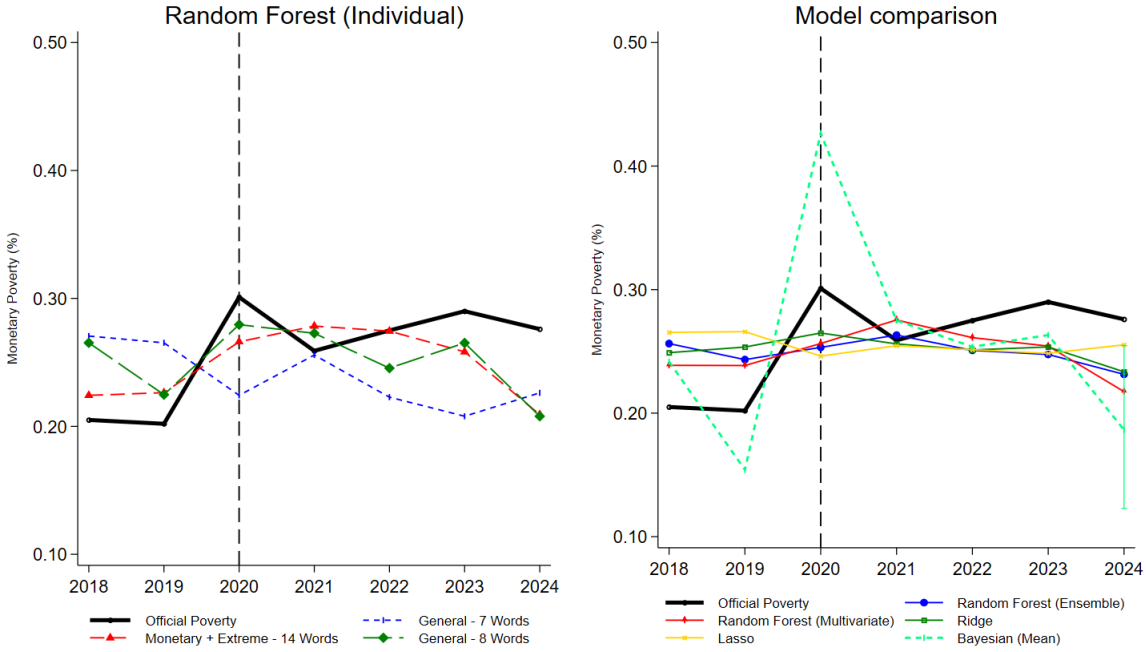
Table 2 summarizes the results from the supervised learning models. Overall, models based on random forests perform better than linear specifications. Among individual predictors, indices constructed from monetary and extreme poverty terms tend to perform better than those based on more general configurations. Multivariate and ensemble specifications reduce prediction errors relative to some individual models, although improvements are not uniform across metrics. Linear models perform worse, and some specifications exhibit limited explanatory power. The Bayesian approach reflects greater uncertainty, consistent with its probabilistic structure.

Table 2: Prediction performance of LDA-based indices with supervised learning models

Model	Index	RMSE	MAE	R^2	Predicted 2024	CI 95
Random Forest (Individual)	Gen. 7w	0.063	0.057	-1.611	0.226	
Random Forest (Individual)	Mon. Ext. 14w	0.024	0.022	0.607	0.209	
Random Forest (Individual)	Gen. 8w	0.032	0.029	0.306	0.208	
Random Forest (Ensemble)	Mean	0.039	0.035	0.004	0.231	
Random Forest (Multivariate)		0.032	0.030	0.316	0.217	
Ridge		0.036	0.032	0.140	0.233	
Lasso		0.047	0.041	-0.440	0.255	
Bayesian (Multivariable)		0.059	0.046		0.186	[0.123, 0.255]

Note: This table summarizes the predictive performance of supervised learning models that use LDA-based topic indices as predictors. We evaluate individual and ensemble forecasts constructed from three selected indices, using both univariate and multivariable specifications. The models include random forest, ridge, and lasso regression, as well as Bayesian regression. Performance metrics include root mean squared error (RMSE), mean absolute error (MAE), and the coefficient of determination (R^2). The final two columns report the 2024 predicted poverty rate and the corresponding 95% confidence interval (when available). We trained all models using annual data from 2018 to 2023.

Figure 7: Forecast accuracy by model type and individual LDA indices



Note: This figure presents supervised poverty forecasts generated from LDA-based topic indices. The left panel displays predictions from Random Forest models trained individually on each selected index. The right panel compares out-of-sample predictions across model types, including ensemble and multivariate random forests, ridge regression, lasso regression, and Bayesian regression. The vertical dashed line marks the onset of the COVID-19 pandemic in 2020.

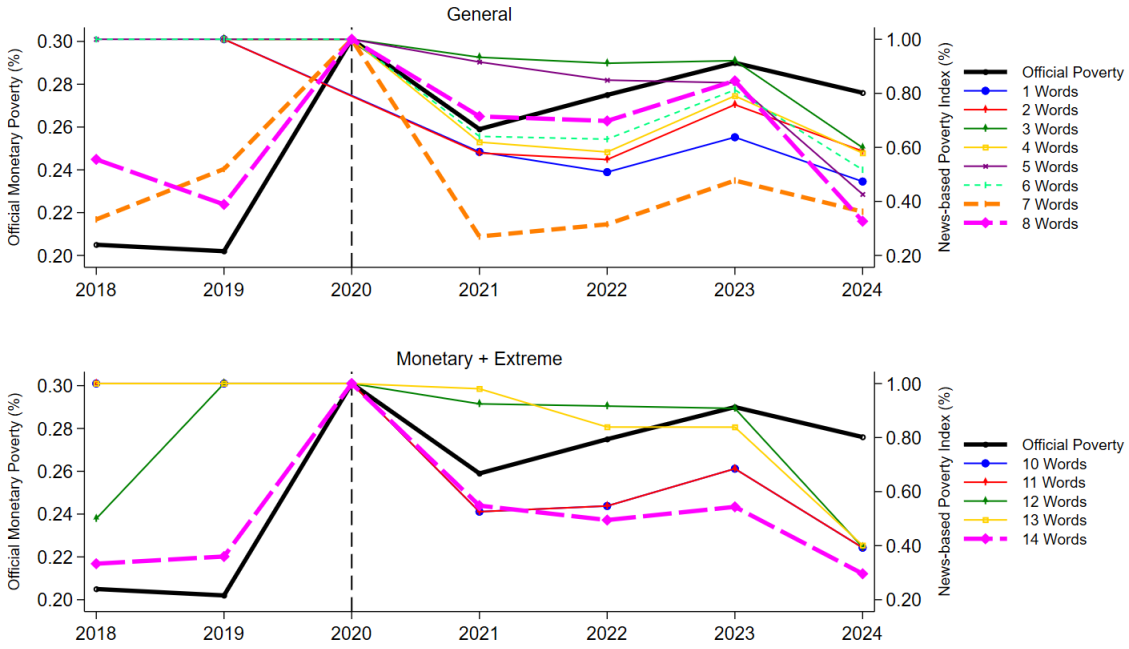
Figure 7 compares forecast trajectories across models. The left panel reports predictions from random forest models trained on individual indices. Most specifications reflect the increase in poverty observed in 2020. In subsequent periods, the individual models differ in their proximity to official poverty levels. The right panel contrasts predictions across model classes. Nonlinear models tend to follow the overall trajectory of official poverty more closely than linear specifications, while Bayesian predictions exhibit greater temporal variability.

Overall, the results indicate that news-based topic indices contain a predictive signal for monitoring poverty dynamics. The forecasting exercise emphasizes relative performance across indices and modeling approaches rather than precise point prediction, highlighting the potential of topic-based indicators to complement traditional data sources when timely information is limited.

5.4. Sensitivity analysis

This section examines the sensitivity of the news-based poverty indices to keyword selection and model choice. We assess how the indices respond to iterative keyword expansion and compare the results with those obtained using standard LDA and Guided LDA. The analysis begins with a small set of seed terms and progressively adds terms until it meets the specifications specified in the previous sections. Guided LDA is implemented as a semi-supervised extension of LDA that uses seed words to guide topic–word associations (Jagarlamudi et al., 2012; Nguyen et al., 2015; Venugopalan and Gupta, 2022a, b). This approach provides a complementary benchmark to the unsupervised LDA results by explicitly emphasizing poverty-related language in the news.

Figure 7: Sensitivity of LDA Poverty Indices to Iterative Keyword Expansion

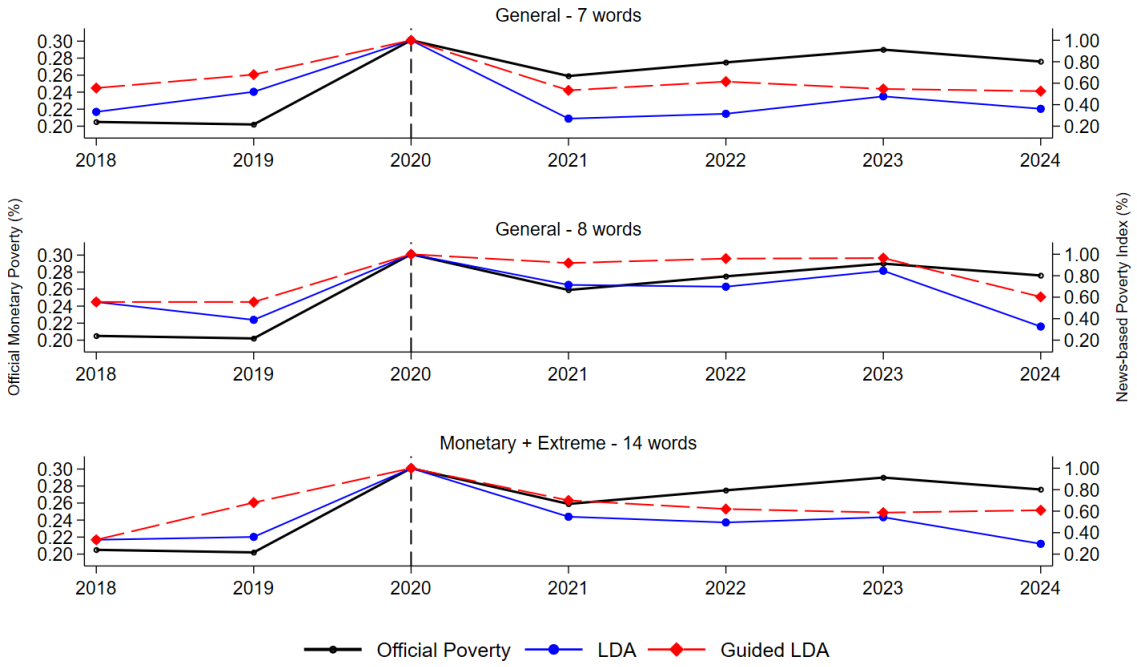


Note: This figure compares official monetary poverty rates (black line, left axis) with news-based poverty indices derived from LDA models using the iterative keyword expansion approach (colored lines, right axis). The top panel reports results for general poverty, while the bottom panel combines monetary and extreme poverty. The dashed vertical line marks the year of the COVID-19 shock (2020).

Figure 7 illustrates the indices' sensitivity to iterative keyword expansion. In the General panel, several specifications exhibit limited variation prior to 2020, with some indices remaining at or near the upper bound of the scale. As a result, not all configurations display a visible increase in 2020. Indices based on a single or a pair of words do not reproduce the main changes observed in official poverty measures. From 2021 onward, the indices diverge in both level and variation, and no monotonic relationship emerges between the number of words included and alignment with the official series. In 2023, several configurations moved in the same direction as official poverty, although with different magnitudes. In 2024, most indices declined, while dispersion across specifications increased.

In the Monetary + Extreme panel, the indices show more similar trajectories across keyword sets. Several specifications exhibit values close to the upper bound of the index around 2020, limiting the visibility of a distinct spike for those configurations. After 2020, the indices separate into different patterns, with some remaining at relatively high levels and others declining more sharply. Differences across keyword sets persist through 2024, indicating that expansion of the vocabulary does not yield uniform improvements in tracking official poverty dynamics.

Figure 8: Sensitivity of News-based Poverty Indices: LDA vs. Guided LDA



Note: This figure compares official monetary poverty rates (black line, left axis) with news-based poverty indices estimated using standard LDA (blue line, right axis) and Guided LDA (red dashed line, right axis). The top and middle panels report results for general poverty, using 7 and 8 keywords, respectively, while the bottom panel shows results for monetary and extreme poverty combined, using 14 keywords. The dashed vertical line marks the year of the COVID-19 shock (2020). Guided LDA follows the semi-supervised algorithm implemented in the original guidedlda Python library, which we adapted to our research setting.

Figure 8 compares indices estimated using standard LDA and Guided LDA for selected specifications. In all cases, both approaches capture the main temporal changes observed in official poverty, including the increase in 2020 and the subsequent decline. However, differences in level emerge after 2020. For the General specifications with seven and eight words, Guided LDA produces higher index values than standard LDA in most years, with larger differences in the seven-word case. For the Monetary + Extreme specification with fourteen words, both methods follow similar directions, although Guided LDA remains systematically above the LDA index after 2020. In 2024, the gap between methods becomes more pronounced as the Guided LDA index declines more slowly.

Overall, the sensitivity analysis indicates that news-based poverty indices are influenced by both keyword selection and the modeling approach. Iterative expansion of the keyword set does not guarantee improved alignment with official poverty statistics, and some configurations yield indices that are weakly informative in certain periods. Guided LDA alters the level of the indices relative to standard LDA, while preserving their broad temporal patterns. These results underscore the importance of careful specification choices when constructing text-based poverty indicators.

6. Discussion, Limitations, and Policy Implications

This study shows that digital news content can be used to construct informative indicators of poverty dynamics. Topic-based indices derived from media text align with official monetary poverty measures and reflect major shifts over time, especially during large aggregate shocks. These patterns support an interpretation of news-based indices as complementary signals rather than substitutes for survey-based statistics.

The results can be framed in terms of latent representations under uncertainty. Policymakers and economic agents do not directly observe poverty. Instead, they rely on imperfect and noisy information. Media coverage aggregates dispersed signals about economic conditions and conveys them through recurring themes. Topic models extract these themes as latent representations that summarize public concern related to poverty. In this sense, the indices provide operationally useful representations for reasoning about poverty dynamics, even though they do not directly measure poverty levels. This interpretation is consistent with recent work on latent representations in economics, where agents infer coarse structures that preserve decision-relevant information while abstracting from underlying complexity (Samuelson & Steiner, 2025).

The forecasting exercises reinforce this interpretation. Differences across indices and modeling strategies show that the objective is not precise point prediction. Instead, the exercises assess the informational content of alternative representations of media discourse. Some indices capture turning points and directional changes more closely than others. No single specification dominates across all periods. This outcome is consistent with the short annual time series and with the difficulty of forecasting poverty using limited data. The forecasting analysis should therefore be interpreted as diagnostic rather than predictive.

Several limitations deserve emphasis. First, topic models are estimated using iterative algorithms that depend on initialization, approximate optimization, and numerical implementation. Model specifications are selected using standard fit and coherence criteria, including perplexity and topic coherence. While the code produces identical results when run repeatedly within the same computational environment, applying the same pipeline across different machines can lead to meaningful variation in inferred topics and, in some cases, in the resulting index dynamics. Such variation arises from differences in numerical backends, parallelization schemes, and hardware architectures. Fixing random seeds or library versions alone is not sufficient to ensure replicability. Improving the stability of latent representations across computational settings remains an important direction for future work.

Second, the framework estimates topic structures using the full news corpus. This choice maximizes information. However, it may distort temporal interpretation if topic definitions evolve with economic conditions. An alternative approach is to estimate topics using an initial training window and then fix the topic structure when constructing indices for later periods. This design would strengthen the ex-ante interpretation of the indices.

Third, the forecasting analysis relies on a short annual time series. The limited number of observations restricts the reliability of standard predictive metrics, especially for flexible machine learning models. Predictive performance should therefore be interpreted with caution. The results are best understood as comparative evidence on information content rather than precise measures of out-of-sample accuracy.

Fourth, the methodology relies on classical topic models that prioritize interpretability. This choice is deliberate. However, it excludes recent advances in natural language processing. Embedding-based representations and neural topic models offer richer semantic structure and improved stability. Extending the framework to these methods is a natural next step.

Fifth, media-based indicators are shaped by the incentives and constraints of news production. Editorial priorities, political pressures, and variation in press freedom influence

which topics receive attention and how they are framed. In media environments that are restricted or highly concentrated, news-based indices may underrepresent socioeconomic hardship or respond weakly to underlying conditions. Contextual validation, therefore, remains essential.

Despite these limitations, the approach has relevant policy implications. News-based indices can complement official poverty statistics when survey data are delayed or infrequent. They provide timely signals related to labor markets, prices, and household vulnerability. Their value lies in timeliness and flexibility rather than precision.

The usefulness of these indicators depends on institutional arrangements that translate information into action. When embedded in predefined decision frameworks such as policy dashboards, contingency planning tools, or social protection systems, the signals can support more responsive policy design. However, institutional capacity and political context remain critical. In settings with weak governance or frequent policy turnover, informational signals may not lead to timely intervention.

This limitation is particularly relevant for development finance. One promising avenue is for multilateral organizations to link disbursements or program triggers to short-term poverty dynamics, as measured by independent, transparent indices. Such arrangements, combined with automated alerts or pre-approved interventions, could improve responsiveness during periods of economic stress. They would not replace traditional measures of poverty.

In summary, topic-based indices provide a scalable and timely complement to household surveys. They are most valuable in contexts marked by data scarcity, economic volatility, or institutional fragility.

7. Conclusions and Future Work

This paper develops a framework for transforming digital news content into informative indicators of poverty dynamics. Using evidence from Peru, the study shows that topic-based indices constructed from media text contain systematic signals related to poverty that complement official statistics.

The paper's main contribution lies in measurement rather than prediction. The proposed indices summarize public concern related to poverty and capture changes in narratives over time. Interpreted as latent representations, they provide a practical way to reason about poverty under uncertainty, using continuously generated, widely available information.

The results show that news-based indices move in line with official poverty measures during major economic shifts. At the same time, the analysis highlights important limits. The indices do not measure poverty levels directly and should not be interpreted as precise predictors. Their value lies in monitoring, comparison, and interpretation rather than in point estimation.

Several directions for future research follow from this work. One priority concerns the stability and replicability of latent representations. As shown in this study, topic-based indices can be sensitive to computational settings. Developing estimation protocols that yield stable representations across environments is therefore essential. Promising avenues include deterministic algorithms, explicit stability-based model selection, and standardized computational frameworks.

A second direction is methodological. Future work can extend the framework to modern natural language processing methods (Dell, 2025). Embedding-based representations, neural topic models, and supervised topic models may capture richer semantic structure and improve robustness. Comparing these approaches with classical topic models would clarify trade-offs

between interpretability, stability, and informational content.

A third direction concerns spatial and temporal scope. Within-country extensions could construct subnational indices that track poverty-related narratives across regions or cities. Extending the analysis to historical news archives would also allow the study of poverty dynamics during earlier episodes of economic crisis or structural change, especially when household survey data are sparse or unavailable.

A fourth direction involves broader measurement applications. Because news text reflects multiple dimensions of economic and social conditions, the framework can be adapted to construct indicators of other development outcomes. These include labor market stress, food insecurity, access to public services, and social protection coverage. Text-based indicators may capture dimensions that are difficult to observe with traditional data sources and can complement image-based or administrative measures.

A fifth direction focuses on bias and triangulation. Future research can examine heterogeneity across media sources and adjust for changes in news intensity or editorial incentives. Linking news-based indices with other high-frequency indicators, such as labor market data, inflation measures, administrative records, or remote sensing data, would help assess when media-based signals reflect development dynamics.

Finally, future research can further develop the theoretical foundations of news-based measurement. Formal models linking underlying conditions, information signals, media coverage, and latent representations can clarify when topic-based indices are informative for policy-relevant decision-making and how they interact with institutional constraints.

Overall, this study highlights the potential of combining economic analysis with text-based methods to improve poverty monitoring and related development measurement. News-based indices do not replace household surveys. They offer a scalable and timely complement, particularly in contexts characterized by data scarcity, economic volatility, or institutional fragility.

References

Aguilar, P., Ghirelli, C., Pacce, M., & Urtasun, A. (2021). Can news help measure economic sentiment? An application in COVID-19 times. Economics Letters, 199, 109730.

Alkire, S., & Foster, J. (2011). Counting and Multidimensional Poverty Measurement. Journal of Public Economics, 95(7–8), 476–487.

Arriaga, B., Gómez, A., Palacios, A., & Aliaga, W. (2025). A Peruvian Machine learning model for E-commerce product matching in South America, Spain and Portugal. Journal of Open Innovation: Technology, Market, and Complexity, 100561.

Atkinson, A. B. (1987). On the measurement of poverty. Econometrica: Journal of the Econometric Society, 749-764.

Baker, S. R., Bloom, N., & Davis, S. J. (2016). Measuring economic policy uncertainty. Quarterly Journal of Economics, 131(4), 1593–1636.

Bellstam, G., Bhagat, S., & Cookson, J. A. (2021). A text-based analysis of corporate innovation. Management Science, 67(7), 4004-4031.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. Journal of Machine Learning Research, 3, 993–1022.

Blumenstock, J. E., Cadamuro, G., & On, R. (2015). Predicting Poverty and Wealth from Mobile Phone Metadata. Science, 350(6264), 1073–1076.

Bourguignon, F., & Chakravarty, S. R. (2003). The measurement of multidimensional poverty. The Journal of Economic Inequality, 1, 25-49.

Bybee, L., Kelly, B., Manela, A., & Xiu, D. (2024). Business news and business cycles. The Journal of Finance, 79(5), 3105-3147.

Church, K. W., & Hanks, P. (1990). Word association norms, mutual information, and lexicography. Computational Linguistics, 16(1), 22–29.

Cong, L. W., Liang, T., Zhang, X., & Zhu, W. (2024). Textual factors: A scalable, interpretable, and data-driven approach to analyzing unstructured information (No. w33168). National Bureau of Economic Research.

Corral, P., Henderson, H., & Segovia, S. (2025). Poverty mapping in the age of machine learning. Journal of Development Economics, 172, 103377.

Cowles, A. (1933). Can stock market forecasters forecast? Econometrica, 1(3), 309–324.

Dell, M. (2025). Deep learning for economists. Journal of Economic Literature, 63(1), 5-58.

Elbers, C., Lanjouw, J. O., & Lanjouw, P. (2003). Micro-level Estimation of Poverty and Inequality. Econometrica, 71(1), 355–364.

Ellingsen, J., Larsen, V. H., & Thorsrud, L. A. (2021). News media versus FRED-MD for macroeconomic forecasting. Journal of Applied Econometrics, 37(1), 63-81.

Engstrom, R., Hersh, J. S., & Newhouse, D. L. (2022). Poverty from Space: Using High-Resolution Satellite Imagery for Estimating Economic Well-Being. World Bank Economic Review, 36(1), 118–139.

Ferrara, L., & Simoni, A. (2023). When are Google data useful to nowcast GDP? An approach via preselection and shrinkage. Journal of Business & Economic Statistics, 41(4), 1188-1202.

Foster, J., Greer, J., & Thorbecke, E. (1984). A class of decomposable poverty measures. Econometrica: journal of the econometric society, 761-766.

Gentzkow, M., Kelly, B., & Taddy, M. (2019). Text as data. Journal of Economic Literature, 57(3), 535–574.

Gonçalves, H. S., & Moro, S. (2023). On the economic impacts of COVID-19: A text mining literature analysis. Review of Development Economics, 27(1), 375-394.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. Proceedings of the National Academy of Sciences, 101(suppl_1), 5228–5235. https://doi.org/10.1073/pnas.0307752101

Grimmer, J., & Stewart, B. M. (2013). Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. Political Analysis, 21(3), 267–297.

Hansen, S., McMahon, M., & Prat, A. (2018). Transparency and deliberation within the FOMC: A computational linguistics approach. Quarterly Journal of Economics, 133(2), 801–870.

Hassan, T. A., Hollander, S., Kalyani, A., van Lent, L., Schwedeler, M., & Tahoun, A. (2025). Text as Data in Economic Analysis. Journal of Economic Perspectives, 39(3), 193-220.

Hirschman, A. O. (1964). The paternity of an index. The American Economic Review, 54(5), 761–762.

Jagarlamudi, J., Daumé III, H., & Udupa, R. (2012). Incorporating lexical priors into topic models. In Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (pp. 204-213).

Jean, N., Burke, M., Xie, M., Davis, W. M., Lobell, D. B., & Ermon, S. (2016). Combining Satellite Imagery and Machine Learning to Predict Poverty. Science, 353(6301), 790–794.

Kalamara, E., Turrell, A., Redl, C., Kapetanios, G., & Kapadia, S. (2022). Making text count: economic forecasting using newspaper text. Journal of Applied Econometrics, 37(5), 896-919.

Larsen, V. H., & Thorsrud, L. A. (2019). The value of news for economic developments. Journal of econometrics, 210(1), 203-218.

Manela, A., & Moreira, A. (2017). News implied volatility and disaster concerns. Journal of Financial Economics, 123(1), 137-162.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to Information Retrieval. Cambridge University Press, Cambridge, UK.

Nguyen, V. A., Boyd-Graber, J., Resnik, P., Cai, D., & Midberry, J. (2015). Lexical and hierarchical topic regression. Proceedings of the 32nd International Conference on Machine Learning (ICML), 53–62.

Ravallion, M. (1992). Poverty Comparisons: A Guide to Concepts and Methods. The World Bank.

Rhoades, S. A. (1993). The Herfindahl–Hirschman Index. Federal Reserve Bulletin, 79, 188–189.

Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the space of topic coherence measures. In Proceedings of the Eighth ACM International Conference on Web Search and Data Mining (pp. 399–408). https://doi.org/10.1145/2684822.2685324

Samuelson, L., & Steiner, J. (2025). Robust Latent Representations.

Sen, A. (1976). Poverty: an ordinal approach to measurement. Econometrica: Journal of the Econometric Society, 219-231.

Shapiro, A. H., Sudhof, M., & Wilson, D. J. (2022). Measuring news sentiment. Journal of econometrics, 228(2), 221-243.

Stevens, K., Kegelmeyer, P., Andrzejewski, D., & Buttler, D. (2012). Exploring topic coherence over many models and many topics. In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (pp. 952–961). Association for Computational Linguistics.

Taddy, M. (2013). Multinomial inverse regression for text analysis. Journal of the American Statistical Association, 108(503), 755–770.

Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. Journal of Finance, 62(3), 1139–1168.

Velásquez, D., Medina, S., Yamada, G., Lavado, P., Nunez-del-Prado, M., Alatrista-Salas, H., & Morzán, J. (2020). I read the news today, oh boy: The effect of crime news coverage on crime perception. World Development, 136, 105111.

Venugopalan, D., & Gupta, R. (2022a). A machine learning method to monitor online discourse: Topic-based tracking of COVID-19 pandemic-related news. Knowledge-Based Systems, 239, 107942. https://doi.org/10.1016/j.knosys.2021.107942

Venugopalan, S., & Gupta, D. (2022b). Machine learning methods for predicting credit default: A comparative analysis. Knowledge-Based Systems, 238, 107924. https://doi.org/10.1016/j.knosys.2021.107924

Venugopalan, M., & Gupta, D. (2022). An enhanced guided LDA model augmented with BERT based semantic strength for aspect term extraction in sentiment analysis. Knowledge-based systems, 246, 108668.