Improvement in Inflation Forecasting: Ensembling Text Mining with Macro Data in Machine Learning Models

Pijush Kanti Das¹ & Prabir Kumar Das¹

¹ Indian Institute of Foreign Trade, Kolkata, India

Correspondence: Pijush Kanti Das, Indian Institute of Foreign Trade, 1583, Chowbaga Canal Side Rd, Madurgha, Chowbaga, Kolkata, West Bengal, 700107, India. Tel: 91-891-042-5811. E-mail: pijush_phdmp19@iift.edu

Received: April 26, 2024  Accepted: May 27, 2024  Online Published: May 30, 2024

doi:10.5539/ijef.v16n6p92  URL: https://doi.org/10.5539/ijef.v16n6p92

Abstract

We forecast inflation using a large news corpus and machine learning methods. Over 3.9 million daily newspaper headlines from January 2001 to June, 2023 are decomposed into monthly time series and integrated with machine learning models to predict inflation. The addition of Text mining in models outperformed the numerical predictions based on the machine learning models without text mining as published by the authors earlier in Das and Das (2024). In addition, the variable importance while analyzing the predictors provides further insights into new variables came out from text mining for which structured data was not available earlier. A dictionary of words sentimental to inflation forecasting has been prepared possibly for the first time. The forecasting model that used text words sentimental to inflation as additional inputs in artificial neural network performed better than all the other models in terms of forecast accuracy. Overall, we provide a novel representation of improvements in adding text mining in machine learning models in inflation forecasting.

Keywords: Inflation forecasting, text mining, India, machine learning

1. Introduction

Inflation forecasting stands as a pivotal element in economic decision-making, as highlighted by Medeiros et al. (2021) and Groen et al. (2013). Central to this process is the reliance of central banks on inflation projections to guide monetary policy and maintain stable inflation levels (Stock & Watson, 1999; D’Agostino & Surico, 2012). The attainment of dependable and precise forecasts for future inflation is paramount for the welfare of all economic stakeholders.

Traditionally, model-based approaches and consumer surveys serve as crucial tools for policymakers and practitioners in monitoring and forecasting inflation within the economy. In India, the Reserve Bank of India (RBI) conducts the Inflation Expectations Survey of Households (IESH) quarterly, aiming to gauge inflation expectations for the next three months and one year ahead. However, findings by Sharma and Bicchal (2018) suggest that survey respondents may not form expectations rationally, raising concerns about the credibility of the IESH conducted by the RBI.

A new frontier has emerged with the integration of machine learning algorithms into the traditional model-based approach for inflation forecasting. The primary drive behind this advancement likely stems from the capability of machine learning models to capture intricate, non-linear relationships inherent in economic data—something that traditional econometric models may struggle with. Machine learning models excel in their adaptability and self-adjustment to changing environments, outperforming econometric models that may falter in accommodating abrupt shifts in economic conditions, such as geopolitical events, technological advancements, or global economic crises. This shift towards machine learning heralds a promising direction for enhancing the accuracy and robustness of inflation forecasts, leveraging the power of advanced algorithms to navigate complex economic landscapes. Based on this assumption, in our earlier paper Das and Das (2024), we forecasted inflation in Indian context using machine learning models with high dimensional data and found the superiority of neural network and random forest model over traditional models.

Nevertheless, despite the advancements in machine learning and traditional forecasting methods, there remain notable shortcomings. Specifically, these approaches often overlook the impact of mass media, which can
substantially influence investor behavior and market dynamics. Recent studies have highlighted the potential of text mining technology in leveraging media sources to inform investment decisions, ranging from stock price and oil price forecasting to foreign exchange predictions (Yu et al., 2005, Chen et al., 2016, Gupta et al., 2020, Barbaglia et al., 2023). Though the application of text mining in the domain of finance was observed largely, but use of text mining for predicting macroeconomic indicator like inflation rate is scarce.

By leveraging large-scale text data from news articles, text mining techniques can uncover hidden patterns, sentiments, and contextual information relevant to economic conditions and inflation dynamics. In this study, the sentimental words derived from large corpus of newspaper headlines have been combined with structured macroeconomic data of independent variables to observe improvement in inflation forecasting, which is a pioneer work in this domain.

2. Literature Review

Machine learning methods are gaining significant traction in both academic and applied research, although their adoption in economics remains relatively new. One innovative aspect of machine learning is text mining, a computational technique used to process and distil vast amounts of text data—information that would be overwhelmingly complex or even impossible for an individual to handle manually. Text mining enables the extraction of valuable insights from novel data sources such as social media platforms (e.g., Twitter, Google) and public media sources (e.g., online news, communication reports). This approach facilitates a deeper analysis and understanding of economic relationships, including consumer behavior, thereby contributing to more informed policy-making and forecasting efforts. Notable studies by researchers like Askitas and Zimmerman (2009) and D’Amuri and Marcucci (2017) exemplify the growing interest and potential impact of text mining applications in economics, highlighting its role in unlocking new dimensions of economic data analysis and interpretation.

In a similar vein of research, Nyman et al. (2021) demonstrated a strong correlation between shifts in emotional narratives across various data sources, indicating their potential to predict economic variables within financial markets. Building upon this notion, Kalamara et al. (2022) leveraged timely signals extracted from three prominent UK newspapers to enhance forecasts of macroeconomic indicators such as GDP, CPI, and unemployment rates. Similarly, Basak et al. (2019) identified co-integration and causality between sentiment extracted from media coverage of Brexit and fluctuations in the British currency. Furthermore, Rambacussing and Kwiatkowski (2020) delved into the efficacy of newspapers in predicting inflation, output, and unemployment in the United Kingdom. Their findings suggest that while sentiment analysis of print media may aid in forecasting unemployment and output, its utility in improving inflation forecasts remains inconclusive.

An additional contribution of this research is to highlight the real-time forecasting of inflation using advanced machine learning methods. The significance of accurate inflation forecasting for informed decision-making is well recognized in the literature, yet improving upon traditional models has proven challenging. As noted by Medeiros et al. (2021), much of the existing literature has overlooked recent advancements in machine learning. Their work demonstrates the potential of machine learning and data-rich models to enhance inflation forecasts significantly. For instance, their LASSO and Random Forest models outperform standard benchmark models like autoregressive models, who also emphasize the effectiveness of high-dimensional models in inflation forecasting within data-rich environments. The literature (Das & Das, 2024), support this argument. In our research, the inclusion of neural network models alongside Random Forest further demonstrates superior performance, particularly in handling nonlinear and dynamic scenarios. This evidence underscores the transformative impact of leveraging machine learning techniques to advance inflation forecasting capabilities, highlighting the potential for more accurate and actionable predictions in real-time economic analysis.

The application of text mining for economic forecasting research remains relatively scarce in India, reflecting a broader trend where traditional methods continue to dominate. While text mining has gained traction in other regions for analyzing textual data from diverse sources such as news articles, social media, and reports, its adoption in Indian economic research is limited. Pandey et al. (2021) used a text analysis approach from the sentiments from the policy statements to make an observation on the monetary policy on inflation forecasts. In fact, as far our knowledge, this study may be pioneering research for inflation forecasting using ensemble model of text data and macroeconomic data in the context of developing countries like India. One key challenge may stem from the complexity and diversity of languages and dialects across India, which presents unique obstacles for natural language processing and sentiment analysis. Additionally, the availability and accessibility of comprehensive text data in structured formats suitable for analysis could be limited. However, as technological capabilities expand, there is potential for increased exploration and utilization of text mining techniques to enrich economic forecasting efforts in India, offering valuable insights into market sentiment, consumer behavior, and
economic trends.

In our study, we aimed to bridge this gap by concentrating on enhancing inflation rate forecasting in India through the application of machine learning models. Our principal aim was to evaluate the efficacy of integrating machine learning techniques with a comprehensive array of data encompassing macroeconomic variables and textual data sourced from news outlets. Notably, while prior studies, including the seminal work by Shapiro et al. (2022), relied on sentiment analysis from print and newspaper articles, our research represents a seminal contribution. We introduced a novel methodology by employing machine learning algorithms to extract sentiment-related words pertaining to inflation from newspaper headlines, augmenting traditional macroeconomic data. Subsequently, this amalgamated dataset was leveraged through machine learning algorithms as a tool for inflation forecasting. This approach not only introduces a fresh methodological dimension but also significantly reduces time and costs associated with the process.

3. Method

Constructing a quantitative measurement of newspaper headlines presents a significant challenge in our study. Our primary methodology for estimating news topics and forecasting inflation involves utilizing a substantial dataset of over 3.9 million daily newspaper headlines specific to India, spanning from January 2001 to June 2023. Leveraging the tm model proposed by Feinerer et al. (2008), we process this extensive corpus of headlines to extract relevant information for estimating monthly inflation, integrating this textual data with structured model data. Given the large dimensionality of this dataset, conventional time-series econometric models become impractical for analysis. To address this limitation, we capitalize on advances in machine learning methods to harness the predictive potential inherent in this vast dataset for inflation forecasting. By employing machine learning techniques, we aim to fully exploit the richness and complexity of the newspaper headlines corpus to enhance the accuracy and timeliness of inflation predictions, thereby advancing the state-of-the-art in economic forecasting methodologies.

Numerous prior studies have relied on traditional macroeconomic variables to forecast inflation (Atkeson et al., 2001; Stock & Watson, 1999; Forni et al., 2003; Medeiros et al., 2021). Given the inclusion of a high-dimensional set of economic variables, it is natural to question whether text data from newspaper headlines provides any additional predictive power in inflation forecasting. In this paper, we aim to investigate the potential incremental value of incorporating text data from newspaper headlines. To address this inquiry, we begin by evaluating the forecasting performance using macroeconomic variables alone, drawing from the extensive panel of macroeconomic variables outlined in our earlier work by Das and Das (2024). Subsequently, in this paper, we extend our analysis by integrating text data into the same machine learning models and training procedures to predict inflation alongside macroeconomic variables.

The machine learning models utilized in this study encompass a range of candidates, including linear models with penalization methods such as Ridge, LASSO, and Elastic Net. Additionally, the study incorporates regression tree techniques like Random Forest, known for their ability to capture complex relationships in data, as well as an Artificial Neural Network (ANN) model, which excels at learning intricate patterns from large datasets.

3.1 Benchmark Model

We employed a non-seasonal Autoregressive Integrated Moving Average (ARIMA) model as our benchmark for comparison. The model is represented by the equation:

\[ y_t' = c + \sum_{i=1}^{p} \phi_i y_{t-i} + \sum_{j=1}^{q} \theta_j \epsilon_{t-j} + \epsilon_t \]  

(1)

Here, \( y_t' \) represents the dependent variable at time \( t \) and several lagged observations (previous time steps), \( \phi_i, \phi_2, ..., \phi_p \) are the coefficients of the lagged values of \( y_t' \), \( \theta_1, \theta_2, ..., \theta_q \) are the coefficients of the lagged values of the error term \( \epsilon_t \).

The parameter \( p \) denotes the number of lag observations included in the model, indicating the extent of memory of past values. The selection of the best ARIMA\( (p,d,q) \) model was determined through an automatic selection process optimizing the Bayesian Information Criterion (BIC) within our analysis.

3.2 Regularized Regression

Regularized regression is a variant of linear regression that incorporates a regularization term into the objective function.

3.2.1 Ridge Regression (L2 Regularization)

The objective of Ridge Regression (James et al., 2013) is to minimize the following loss function:
\[ \text{Objective} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \] (2)

Here, \( n \) represents the number of observations, \( p \) is the number of features or explanatory variables, \( y_i \) is the observed response for the \( i \)-th observation, \( \hat{y}_i \) is the predicted response for the \( i \)-th observation, \( \beta_j \) is the coefficient for the \( j \)-th feature, and \( \lambda \) is the regularization parameter.

The term \( \lambda \sum_{j=1}^{p} \beta_j^2 \) acts as a penalty on the model’s coefficients, discouraging large coefficient values. As the regularization parameter \( \lambda \) increases, this penalty’s effect intensifies, resulting in smaller coefficients and ultimately, a more straightforward model.

3.2.2 Lasso Regression (L1 Regularization)

The objective of Lasso Regression (Tibshirani, 1996) is to minimize the following:

\[ \text{Objective} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \] (3)

In this equation, all terms are analogous to those in Ridge Regression. However, the term \( \lambda \sum_{j=1}^{p} |\beta_j| \) introduces sparsity into the model. It acts as an L1 penalty on the coefficients \( \beta_j \), encouraging many coefficients to shrink all the way to zero. This characteristic of Lasso Regression makes it effective for feature selection, as it identifies and retains only the most relevant predictors in the model.

3.2.3 Elastic Net (ElNet) Regression

Elastic Net Regression combines both L1 and L2 regularization terms in its objective function. By integrating both types of penalties, Elastic Net addresses the limitations of individual regularization methods like Lasso (which can be unstable when the number of predictors is large) and Ridge (which does not perform feature selection).

\[ \text{Objective} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda (\rho \sum_{j=1}^{p} |\beta_j| + \frac{1-\rho}{2} \sum_{j=1}^{p} \beta_j^2) \] (4)

where, \( \rho \) is the mixing parameter with \( 0 \leq \rho \leq 1 \), which determines the balance between L1 (Lasso) and L2 (Ridge) penalties.

3.3 Random Forest (RF) Model

Breiman (2001) introduced the Random Forest (RF) model as an ensemble learning method designed for making predictions by aggregating the outputs of multiple decision trees.

Mathematically, the prediction for a new data point, \( x \), in a regression tree can be represented as:

\[ \hat{y}(x) = \frac{1}{n} \sum_{i=1}^{n} y_i \] (5)

where, \( n \) is the number of samples in the leaf node to which \( x \) belongs, and \( y_i \) is the target variable value for each sample in that leaf node. In this study randomforest package in R has been used.

3.4 Artificial Neural Network (ANN) Model

An Artificial Neural Network (ANN) is a machine learning model inspired by the structure and functioning of the human brain. It consists of interconnected nodes, or artificial neurons, organized into layers. The three main types of layers in a typical ANN are the input layer, hidden layer, and output layer.

The equations for an artificial neural network (ANN) involve the computation within each node, which includes the weighted sum of inputs, the application of an activation function, and the propagation of the output to subsequent layers (Das & Das, 2024). Here are the key equations for a simple feedforward neural network used in this study with one hidden layer:

Weighted Sum for a Node (including bias):

\[ Z_j = \sum_{i=1}^{n} (w_{ij} \times x_i + b_j) \] (6)

where, \( Z_j \) is the weighted sum for node \( j \) in the current layer, \( w_{ij} \) is the weight connecting the \( i \)-th node in the previous layer to the \( j \)-th node in the current layer, \( x_i \) is the output of the \( i \)-th node in the previous layer and \( b_j \) is the bias for node \( j \).

3.5 Model Evaluation

In order to evaluate the forecasting performance of each model, we relied on Root Mean Squared Error (RMSE) and relative RMSE (Das & Das, 2020). Model has been evaluated for both trained and test data. As the model was already built based on trained data, it is more important to observe the prediction accuracy in respect of test data.
Root Mean Squared Error (RMSE)\[ \text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \] (7)

Relative RMSE\[ \text{Relative RMSE} = \frac{\text{RMSE of Model}_i}{\text{RMSE of Benchmark model}} \] (8)

4. Data Sources and Preprocessing

4.1 Data Sources

Our primary text dataset comprises daily newspaper headlines from the Times of India (ToI), originally compiled by Kulkarni (2020) and subsequently updated on Kaggle by Kulkarni himself in 2023. The dataset spans from January 2001 to June 2023. The Times of India is among India’s largest newspapers and is considered a reliable indicator of public sentiment. It is reasonable to infer that the ToI provides timely coverage of real-time events in India, facilitating a comprehensive assessment of current economic conditions.

4.2 Data Preprocessing

We follow standard text processing steps to clean the news corpus, as outlined below:

Extraction of data: The data was obtained from Kaggle in CSV format and subsequently imported into R for analysis.

Preparation of time series: To prepare the daily newspaper data for time series analysis, a new variable representing the year and month was created. This variable allows for the aggregation of data at a monthly level, facilitating the construction of a monthly time series for modeling purposes.

Article Screening: Initially, the entire dataset of 3.9 million newspaper headlines was categorized into 1024 distinct categories. Among these, 887 categories deemed irrelevant (such as “Sports”, “Entertainment”, “Lifestyle”, “Festivals”, “Astrology”, etc.) were excluded from further analysis. This screening process resulted in retaining 137 categories that are relevant to topics like economics, inflation, business, etc.

After this screening, a subset of 828,626 newspaper headlines was identified, consisting of a total of 1,243,474 words spread across each month from January 2001 to June 2023 (spanning 270 months). These selected headlines and their associated words form the basis of the dataset used for subsequent analysis.

Term Frequency calculation: Frequency of each word in a particular month has been calculated.

Normalization and lemmatization: Normalization and lemmatization involved removing any non-alphabetical characters, punctuation marks, and numbers from the text. Each word was converted to lowercase and transformed to its base form (non-inflected form) using lemmatization. Finally, common stop words were eliminated from the processed text.

Preparation of Dictionary sentimental to inflation: Creation of an inflation-sensitive dictionary involved selecting words that appeared in at least 75% of the months (200 out of 270 months). This criterion yielded a set of 1,415 words identified as relevant to inflation sentiment. From this selection, a refined dictionary containing 166 words specifically associated with inflation sentiment was compiled (refer to Table A1 in Appendix A).

Document-Term Matrix (DTM) construction and finalization of data: We computed the relative frequencies of the selected words as per dictionary based on their percentage of count of a particular word among all words in that month and used this information to create a Document-Term Matrix (DTM). Each column in the DTM represents a distinct word from the dictionary, and each row corresponds to the relative frequency of that word in each month.

This matrix serves as a structured dataset that captures word usage patterns over time, facilitating subsequent analysis and interpretation.

Preparation of Dictionary sentimental to inflation: The words which are presents at least in 75% of months (200 months) are selected. 1415 words have been found. The words sentimental to inflation have been prepared and a dictionary of 166 words has been prepared. The DTM of these words with a dimension of 270X166 has been prepared.

Finalization of Data for model: Finally, the data has been truncated to January 2012 to December, 2022 (132 months) as Consumer Price Index (CPI) data in India is available from January, 2012 and also to keep parity with the structured macroeconomic data as described in Das and Das, 2024 for comparison and observe the improvements in adding text data in the model. A DTM of 132*166 with relative frequency has been used and added with earlier model data. In Das and Das (2024), structured monthly data for 56 features were used and, in this paper, finally the dimension of data has been used as 132*222. As this high-dimensional data is not suitable for
econometric approaches where n>p, we have used machine learning models for further investigations.

For using ML techniques, we divided our data set into two parts, namely, analysis sample, a data set on which the models are built, and test sample as test data, where the models are tested. The partitioning of the data set is based on 70% of the training set and 30% of test splitting scheme (Siami-Namini et al., 2019).

5. Results

Table 1 presents the RMSE (Root Mean Squared Error) values for various models used for inflation forecasting, evaluated on both training and test datasets. It also shows the RMSE value in both cases without text data and including text data and also improvements in forecasting error.

Table 1. RMSE for both train and test data (January 2012-December, 2022)

<table>
<thead>
<tr>
<th>Model Name</th>
<th>RMSE_Trained</th>
<th>% improvement in forecasting</th>
<th>RMSE Test</th>
<th>% improvement in forecasting</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Excluding text mining</td>
<td>Including text mining</td>
<td>Excluding text mining</td>
<td>Including text mining</td>
</tr>
<tr>
<td>Benchmark: ARIMA</td>
<td>1.841</td>
<td>1.841</td>
<td>2.885</td>
<td>2.885</td>
</tr>
<tr>
<td>Ridge Regression</td>
<td>0.622</td>
<td>0.218</td>
<td>55.208</td>
<td>27.670</td>
</tr>
<tr>
<td>LASSO Regression</td>
<td>0.671</td>
<td>0.277</td>
<td>11.540</td>
<td>2.526</td>
</tr>
<tr>
<td>ElasticNet Regression</td>
<td>1.371</td>
<td>0.526</td>
<td>2.148</td>
<td>2.558</td>
</tr>
<tr>
<td>Random Forest</td>
<td>0.360</td>
<td>0.361</td>
<td>1.654</td>
<td>1.900</td>
</tr>
<tr>
<td>Artificial Neural Network</td>
<td>0.644</td>
<td>0.025</td>
<td>2.961</td>
<td>1.560</td>
</tr>
</tbody>
</table>

Note. Authors’ calculation.

Table 1 shows, incorporating text mining data has led to a notable decrease in RMSE (Root Mean Squared Error) for the LASSO, Ridge, and ANN (Artificial Neural Network) models, both in training and test datasets. Models such as Ridge Regression, LASSO Regression and Artificial Neural Network (ANN) demonstrate substantial improvements in forecasting accuracy when text mining is incorporated. Ridge Regression shows a significant improvement in forecasting precision when text mining is applied, with a 64.9% reduction in RMSE for training data and a 49.9% reduction for test data. LASSO Regression exhibits the highest improvement in forecasting accuracy among all models, with a remarkable 78.1% reduction in RMSE for training data and a 78.3% reduction for test data when text mining is incorporated. ElasticNet Regression and Random Forest, show mixed results, with ElasticNet Regression benefiting from text mining for training data but not for test data, while Random Forest experiences marginal changes in RMSE values. Artificial Neural Network (ANN) demonstrates a substantial 96.1% improvement in forecasting error for training data and a 47.3% improvement for test data when text mining is integrated.

These findings underscore the importance of integrating newspaper information into inflation forecasting models, highlighting how textual data can enhance predictive performance and provide valuable insights for economic analysis.

Table 2 below showcasing Relative RMSE values and rankings for different models compared to the benchmark Auto Regressive Moving Average (ARIMA) model, evaluated on both training and test datasets. The table includes Relative RMSE values for models trained with and without text data, along with improvements in forecasting accuracy and model rankings based on precision.

Table 2. Relative RMSE for both train and test data (January 2012-December, 2022)

<table>
<thead>
<tr>
<th>Model Name</th>
<th>Trained Data Relative RMSE</th>
<th>Test Data Relative RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Without text mining</td>
<td>Including text mining</td>
</tr>
<tr>
<td>Benchmark: ARIMA</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>Ridge Regression</td>
<td>0.338</td>
<td>0.119</td>
</tr>
<tr>
<td>LASSO Regression</td>
<td>0.364</td>
<td>0.150</td>
</tr>
<tr>
<td>ElasticNet Regression</td>
<td>0.745</td>
<td>0.286</td>
</tr>
<tr>
<td>Random Forest</td>
<td>0.196</td>
<td>0.196</td>
</tr>
<tr>
<td>Artificial Neural Network</td>
<td>0.350</td>
<td>0.014</td>
</tr>
</tbody>
</table>

Note. Authors’ calculation.
Table 2 presents Relative RMSE values, normalized to the benchmark ARIMA model (Relative RMSE of 1.000), for both training and test datasets. Lower Relative RMSE values indicate better performance compared to the ARIMA benchmark, with values closer to 0 representing higher precision in forecasting.

Models incorporating text mining consistently demonstrate lower Relative RMSE values compared to models trained without text mining, highlighting the importance of textual data in improving forecasting accuracy.

Models are ranked based on their Relative RMSE values for the test dataset (including text mining). The lower the Relative RMSE value, the higher the model's ranking in terms of forecasting precision. Among the models with text mining, the Artificial Neural Network (ANN) achieves the best performance with the lowest Relative RMSE of 0.541, followed by Random Forest (0.658), LASSO Regression (0.875), ElNet Regression (0.887), Ridge Regression (9.591), and ARIMA (1.000).

6. Discussion

Accurate and reliable inflation forecasts are essential for policymakers, investors, and laborers, particularly given the current context of unprecedented events like the Covid-19 pandemic and the dynamic nature of markets. In this paper, we propose a novel approach by integrating economic narratives into inflation forecasting.

Recently, economists such as Shiller (2017) have hypothesized that narratives play a significant role in shaping economic expectations and influencing macroeconomic outcomes. Our research goes beyond traditional methods by enhancing inflation forecasting using novel narrative data extracted from Indian newspaper headlines. Additionally, our study contributes empirical evidence supporting the notion that narratives or textual data play a crucial role in macroeconomics.

By leveraging economic narratives from newspaper headlines, we aim to improve the accuracy and effectiveness of inflation forecasting models, ultimately providing valuable insights for decision-makers and stakeholders navigating complex economic environments. This research underscores the importance of incorporating textual data analytics into macroeconomic analysis to capture nuanced trends and behaviors that impact economic outcomes.

6.1 Salient Findings

The key empirical findings derived from our study are described below.

Initially, we leverage a substantial text corpus comprising newspaper headlines within the Indian context, pioneering a methodology that integrates machine learning models. This methodology amalgamates structured data, encompassing macroeconomic variables recorded on a monthly basis, with unstructured data sourced from daily newspaper headlines through the utilization of text mining techniques. Additionally, we construct a specialized dictionary comprising sentiment-laden terms tailored specifically for inflation forecasting, augmenting the efficacy of our predictive models.

Furthermore, our analysis delves into whether the incorporation of textual data yields any incremental value in mitigating model error within the realm of inflation forecasting. Our investigation reveals a notable enhancement in forecasting accuracy subsequent to the fusion of structured macroeconomic data with textual data, underscoring the substantial utility of integrating diverse data sources in predictive modeling endeavors.

Lastly, our inquiry scrutinizes the efficacy of various machine learning (ML) models, culminating in the determination that the Artificial Neural Network (ANN) model demonstrates superior forecasting performance, even subsequent to the integration of ensembled data. This underscores the pivotal role of employing advanced ML techniques in the realm of economic forecasting, highlighting their potential to yield unparalleled predictive accuracy and insight.

6.2 Implications

Having established the efficacy of integrating text mining into inflation forecasting, our focus shifts to elucidate the origins of the predictive potency inherent in economic narratives. One plausible avenue lies in the influence economic narratives exert on, or their reflection of, the expectations harbored by economic agents regarding inflation. Indeed, the trajectory of inflation is often shaped by the anticipations and decisions of these agents, as emphasized in studies by Coibion et al. (2018), Coibion and Gorodnichenko (2015), and Berge (2018). Thus, the nexus between economic narratives and inflation expectations emerges as a pivotal channel worthy of exploration.

This paper contributes to and intersects with two prominent strands of literature. Primarily, our work aligns directly with the extensive body of research dedicated to the pursuit of reliable and precise inflation forecasts. Noteworthy contributions in this domain include seminal works by Stock and Watson (1999, 2016); Faust and Wright (2013), Ang et al. (2007), and Medeiros et al. (2021), among others. A pivotal departure between these studies and ours lies...
in our novel incorporation of textual data extracted from newspapers. While existing literature predominantly relies on quantitative variables for inflation forecasting, our approach introduces a pioneering dimension by integrating the rich informational content embedded within economic narratives. By doing so, we transcend the confines of purely quantitative analysis and gain insights into the multifaceted qualitative factors that exert substantial influence on the economy, encompassing aspects such as political risks and societal sentiments. This broader perspective facilitated by economic narratives enhances our ability to comprehensively capture the intricate determinants underlying inflation dynamics.

Furthermore, our paper contributes to the burgeoning literature that integrates narratives extracted from newspapers into investigations of economic decision-making and macroeconomic outcomes. Shiller (2017) lays down a foundational framework for examining the interplay between narratives and economic fluctuations. An expanding body of scholarship explores the pivotal role of news media in propagating narratives (Bybee et al., 2021; Basak et al., 2019) and shaping economic expectations (Coibion et al., 2022). However, quantitatively measuring narratives poses a formidable challenge. To address this challenge, our paper devises a model that seamlessly integrates structured macroeconomic variables with textual data derived from newspaper headlines pertaining to inflation. In doing so, we contribute to advancing the understanding of the role economic narratives play in the realm of inflation forecasting, thereby enriching the analytical toolkit available for studying economic phenomena.

In conclusion, the insights gleaned from this study underscore the pivotal role of economic narratives sourced from newspapers in the prediction of inflation. These narratives adeptly encapsulate a spectrum of factors influencing inflation, irrespective of the availability of structured macroeconomic data. Furthermore, our findings suggest that the fusion of economic narratives with machine learning models enables the capture of valuable information aligned with economic intuition and theoretical frameworks, transcending mere data mining exercises. Thus, this study underscores the intrinsic value of economic narratives as potent predictors in the realm of inflation forecasting, complementing traditional quantitative approaches.

6.3 Way Forward

This study has demonstrated the efficacy of utilizing newspaper headlines as a text corpus for inflation forecasting, yielding remarkable results. Building upon these findings, future research endeavors could extend the scope by incorporating larger text datasets, such as comprehensive newspaper corpora or data from sources like Bloomberg. By leveraging expansive textual data, researchers can delve deeper into understanding the nuanced impact of textual information on macroeconomic forecasting. Expanding the breadth of text data under examination holds promise for unraveling additional insights and refining predictive models, thereby advancing our comprehension of the intricate dynamics governing economic phenomena.

References


Appendix A

Table A1. Dictionary of Words related to inflation forecasting derived from newspaper headlines

<table>
<thead>
<tr>
<th>Sl. No.</th>
<th>Words</th>
<th>Sl. No.</th>
<th>Words</th>
<th>Sl. No.</th>
<th>Words</th>
<th>Sl. No.</th>
<th>Words</th>
<th>Sl. No.</th>
<th>Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Aircraft</td>
<td>36</td>
<td>Emerg</td>
<td>71</td>
<td>Income</td>
<td>106</td>
<td>Petrol</td>
<td>141</td>
<td>Sale</td>
</tr>
<tr>
<td>2.</td>
<td>airlin</td>
<td>37</td>
<td>Employe</td>
<td>72</td>
<td>India</td>
<td>107</td>
<td>Pharma</td>
<td>142</td>
<td>Sanction</td>
</tr>
<tr>
<td>3.</td>
<td>Auto</td>
<td>38</td>
<td>Equity</td>
<td>73</td>
<td>Industry</td>
<td>108</td>
<td>Poor</td>
<td>143</td>
<td>Save</td>
</tr>
<tr>
<td>4.</td>
<td>Balance</td>
<td>39</td>
<td>Exchang</td>
<td>74</td>
<td>Inflat</td>
<td>109</td>
<td>Port</td>
<td>144</td>
<td>Sbi</td>
</tr>
<tr>
<td>5.</td>
<td>Bank</td>
<td>40</td>
<td>Expand</td>
<td>75</td>
<td>Info</td>
<td>110</td>
<td>Price</td>
<td>145</td>
<td>Scam</td>
</tr>
<tr>
<td>6.</td>
<td>Benefit</td>
<td>41</td>
<td>Expans</td>
<td>76</td>
<td>Infrastructure</td>
<td>111</td>
<td>Privat</td>
<td>146</td>
<td>Scheme</td>
</tr>
<tr>
<td>7.</td>
<td>Bill</td>
<td>42</td>
<td>Export</td>
<td>77</td>
<td>Insur</td>
<td>112</td>
<td>Produc</td>
<td>147</td>
<td>Season</td>
</tr>
<tr>
<td>8.</td>
<td>Billion</td>
<td>43</td>
<td>Factori</td>
<td>78</td>
<td>Interest</td>
<td>113</td>
<td>Product</td>
<td>148</td>
<td>SEBI</td>
</tr>
<tr>
<td>9.</td>
<td>Bond</td>
<td>44</td>
<td>Farm</td>
<td>79</td>
<td>Internet</td>
<td>114</td>
<td>Profit</td>
<td>149</td>
<td>Sector</td>
</tr>
<tr>
<td>10.</td>
<td>Boost</td>
<td>45</td>
<td>FDI</td>
<td>80</td>
<td>Invest</td>
<td>115</td>
<td>Progress</td>
<td>150</td>
<td>Sell</td>
</tr>
<tr>
<td>11.</td>
<td>Brand</td>
<td>46</td>
<td>Fee</td>
<td>81</td>
<td>Investor</td>
<td>116</td>
<td>Project</td>
<td>151</td>
<td>Sensex</td>
</tr>
<tr>
<td>12.</td>
<td>Budget</td>
<td>47</td>
<td>Finance</td>
<td>82</td>
<td>Ipo</td>
<td>117</td>
<td>Promot</td>
<td>152</td>
<td>Share</td>
</tr>
<tr>
<td>13.</td>
<td>Busi</td>
<td>48</td>
<td>Financi</td>
<td>83</td>
<td>Job</td>
<td>118</td>
<td>properti</td>
<td>153</td>
<td>Shock</td>
</tr>
<tr>
<td>14.</td>
<td>Buy</td>
<td>49</td>
<td>Fiscal</td>
<td>84</td>
<td>Labour</td>
<td>119</td>
<td>Pull</td>
<td>154</td>
<td>Shop</td>
</tr>
<tr>
<td>15.</td>
<td>Buyer</td>
<td>50</td>
<td>Flood</td>
<td>85</td>
<td>Land</td>
<td>120</td>
<td>Push</td>
<td>155</td>
<td>Stock</td>
</tr>
<tr>
<td>16.</td>
<td>Car</td>
<td>51</td>
<td>Food</td>
<td>86</td>
<td>Loan</td>
<td>121</td>
<td>Pvt</td>
<td>156</td>
<td>Store</td>
</tr>
<tr>
<td>17.</td>
<td>Cash</td>
<td>52</td>
<td>Foreign</td>
<td>87</td>
<td>Manufactur</td>
<td>122</td>
<td>Rate</td>
<td>157</td>
<td>Strateg</td>
</tr>
<tr>
<td>18.</td>
<td>Chang</td>
<td>53</td>
<td>Fuel</td>
<td>88</td>
<td>Margin</td>
<td>123</td>
<td>Rais</td>
<td>158</td>
<td>Strategi</td>
</tr>
<tr>
<td>19.</td>
<td>Cheaper</td>
<td>54</td>
<td>Fund</td>
<td>89</td>
<td>Market</td>
<td>124</td>
<td>Rbi</td>
<td>159</td>
<td>Stress</td>
</tr>
<tr>
<td>20.</td>
<td>Claim</td>
<td>55</td>
<td>Gain</td>
<td>90</td>
<td>Merger</td>
<td>125</td>
<td>Realti</td>
<td>160</td>
<td>Suppli</td>
</tr>
<tr>
<td>21.</td>
<td>Corpor</td>
<td>56</td>
<td>Gdp</td>
<td>91</td>
<td>Mf</td>
<td>126</td>
<td>Recruit</td>
<td>161</td>
<td>Tourism</td>
</tr>
<tr>
<td>22.</td>
<td>Cost</td>
<td>57</td>
<td>Grow</td>
<td>92</td>
<td>Million</td>
<td>127</td>
<td>Reform</td>
<td>162</td>
<td>Touris</td>
</tr>
<tr>
<td>23.</td>
<td>Cr</td>
<td>58</td>
<td>Gold</td>
<td>93</td>
<td>Missil</td>
<td>128</td>
<td>Regul</td>
<td>163</td>
<td>Trade</td>
</tr>
<tr>
<td>24.</td>
<td>Crisi</td>
<td>59</td>
<td>Govern</td>
<td>94</td>
<td>Mobil</td>
<td>129</td>
<td>Research</td>
<td>164</td>
<td>Valu</td>
</tr>
<tr>
<td>25.</td>
<td>Creore</td>
<td>60</td>
<td>Govt</td>
<td>95</td>
<td>Money</td>
<td>130</td>
<td>Reserve</td>
<td>165</td>
<td>Work</td>
</tr>
<tr>
<td>26.</td>
<td>Credit</td>
<td>61</td>
<td>Grant</td>
<td>96</td>
<td>Mutual</td>
<td>131</td>
<td>Restor</td>
<td>166</td>
<td>Worker</td>
</tr>
<tr>
<td>27.</td>
<td>Cut</td>
<td>62</td>
<td>Green</td>
<td>97</td>
<td>Need</td>
<td>132</td>
<td>Retain</td>
<td></td>
<td></td>
</tr>
<tr>
<td>28.</td>
<td>Debt</td>
<td>63</td>
<td>Grow</td>
<td>98</td>
<td>Oil</td>
<td>133</td>
<td>Return</td>
<td></td>
<td></td>
</tr>
<tr>
<td>29.</td>
<td>Deposit</td>
<td>64</td>
<td>Growth</td>
<td>99</td>
<td>Onlin</td>
<td>134</td>
<td>Retail</td>
<td></td>
<td></td>
</tr>
<tr>
<td>30.</td>
<td>Demand</td>
<td>65</td>
<td>Hail</td>
<td>100</td>
<td>Open</td>
<td>135</td>
<td>Revamp</td>
<td></td>
<td></td>
</tr>
<tr>
<td>31.</td>
<td>Diesel</td>
<td>66</td>
<td>Highway</td>
<td>101</td>
<td>Oversea</td>
<td>136</td>
<td>Revenue</td>
<td></td>
<td></td>
</tr>
<tr>
<td>32.</td>
<td>Dollar</td>
<td>67</td>
<td>Hire</td>
<td>102</td>
<td>Owner</td>
<td>137</td>
<td>Rich</td>
<td></td>
<td></td>
</tr>
<tr>
<td>33.</td>
<td>Earn</td>
<td>68</td>
<td>Hospit</td>
<td>103</td>
<td>Paid</td>
<td>138</td>
<td>Rise</td>
<td></td>
<td></td>
</tr>
<tr>
<td>34.</td>
<td>Econom</td>
<td>69</td>
<td>Impact</td>
<td>104</td>
<td>Payment</td>
<td>139</td>
<td>Rs</td>
<td></td>
<td></td>
</tr>
<tr>
<td>35.</td>
<td>Econom</td>
<td>70</td>
<td>Import</td>
<td>105</td>
<td>Pension</td>
<td>140</td>
<td>Salari</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. Authors’ calculation.

Copyrights

Copyright for this article is retained by the author(s), with first publication rights granted to the journal.

This is an open-access article distributed under the terms and conditions of the Creative Commons Attribution license (http://creativecommons.org/licenses/by/4.0/).