Inflation Forecasting using Unstructured Data: The Benefits of News-based Indexes

Gilberto Boaretto
Dept. of Economics, PUC-Rio, Brazil
gilbertoboaretto@hotmail.com

Marcelo Fernandes
Sao Paulo School of Economics, FGV, Brazil
marcelo.fernandes@fgv.br

Marcelo C. Medeiros
Dept. of Economics, UIUC, US
marcelom@illinois.edu

Thiago Milagres
Vox Radar, Brazil
thiagogpsmilagres@gmail.com

This draft: March 16, 2023

Abstract

We propose forward-looking indexes for inflation based on tweets and newspaper articles employing a supervised machine-learning approach. Using Brazilian data, we verify that the news-based indexes were able to anticipate long-term trends as well as capture short-term movements of the accumulated inflation over 3, 6, and 12 months ahead at various periods. Furthermore, the proposed indexes could improve inflation forecast performance. More specifically, for short horizons (3 and 6 months ahead), a bias correction model for the median of available survey-based expectations benefits from including news-based indexes. On the other hand, for longer-term inflation (12 months ahead), models incorporating a large number of predictors can also be improved by incorporating the indexes. Thus, considering indexes from social media and news sources can improve inflation forecasting. The intuition for the result is that it pays to consider a broader set of information than solely that resulting from survey-based expectations that account only for experts' opinions.

Keywords: inflation forecasting; unstructured data; Twitter; newspapers; elastic net; adaLASSO.

JEL Codes: C22, C52, C53, C55, E37.

Acknowledgements: We are grateful to comments from Gustavo Gonzaga, Walter Novaes, Anna Catarina Tavella, and all seminar participants at PUC-Rio Workshop and Itaú Webinar.

XXIII Brazilian Finance Meeting – São Paulo, July 27-29, 2023

Track: Econometrics and Numerical Methods

1 Introduction

Unstructured data are becoming very popular in economic modeling and forecasting. Newspapers and social networks such as Twitter produce a considerable volume of unstructured data that should reflect well the information flow. Based on this thought, this work investigates whether indexes constructed from tweets and newspaper articles can help us anticipate future movements in inflation. In particular, inflation forecasting is an old and relevant research topic that presents new perspectives when considering unstructured data. The literature on this or directly related topic has been expanding by employing both new econometric techniques (Inoue and Kilian, 2008; Garcia, Medeiros, and Vasconcelos, 2017; Medeiros, Vasconcelos, Veiga, and Zilberman, 2021) and new databases such as Google Trends (Guzman, 2011; Li, Shang, Wang, and Ma, 2015; Niesert, Oorschot, Veldhuisen, Brons, and Lange, 2020), newspaper articles (Rambaccussing and Kwiatkowski, 2020; Larsen, Thorsrud, and Zhulanova, 2021), and Twitter (Angelico, Marcucci, Miccoli, and Quarta, 2022).

Experts write articles and opinion pieces in newspapers about economics, politics, social questions, and the international scene. Several economists, politicians, consumers, and entrepreneurs share their thoughts on social media about inflation, prices, and related topics. Could this information be used to obtain more accurate inflation forecasts than available expectations? This study aims to address this question and explore whether non-traditional data sources remain relevant and informative even in the presence of several macroeconomic and financial variables commonly used as predictors for inflation. Our application will address the Brazilian case. The Central Bank of Brazil managed the Focus Survey, a daily collection of inflation expectations provided by market specialists in the country. Traditionally, it is challenging to outperform these expectations, especially for shorter forecast horizons.

In this paper, we utilize a supervised machine learning procedure via the elastic net to construct forward-looking indexes for inflation using information gathered from Twitter and newspapers. This procedure is also called time-varying dictionary approach in the literature (Lima, Godeiro, and Mohsin, 2021). We base the methodology on the occurrence count of terms appearing in tweets or articles, with a broad set of predefined terms collected from Twitter and pre-selected n-grams from three relevant Brazilian newspapers used for articles. After selecting relevant terms for different inflation horizons employing elastic net estimator, we construct two versions of indexes from three distinct information sets. The non-standardized version predicts inflation based on the latest available counts. In a standardized version, we divide the previous predicted value by the sum of the absolute values of each term in the linear model. The information sets consist of only Twitter, only newspapers, or both. Throughout the paper, we detail the advantages and challenges of each version. Finally, in addition to visually verifying the adherence of the indexes to future inflation, we also conduct pseudo-out-of-sample forecasting exercises in which we compare models that include or disregard the indexes. We evaluate a simple historical bias correction model for available survey-based expectation, as well as a data-rich model that incorporates several predictors typically used in inflation forecasting.

Results overview. The news-based indexes were able to anticipate long-term trends and captured short-term movements in 3-, 6-, and 12-month-ahead cumulative inflation at various periods. Considering the benefits in forecasting inflation accumulated over 3 and 6 months ahead, the indexes contributed to a reduction in the root mean squared forecast error (RMSE) of a bias correction model for available Focus' inflation expectations. The model including an index based solely on articles achieved the best predictive performance for 3-month cumulative inflation among all the models considered, delivering a reduction of 26% of RMSE compared to the median of available inflation expectations. For 6-month-ahead inflation, the reductions were more modest, ranging from 7% to almost 13%, while the model that does not include any index registers a reduction of only 4%. In its turn, for inflation accumulated over 12 months, the inclusion of an index based solely on tweets improved the already good result of a high-dimensional model. More specifically, there was an extra reduction of 11 percentage points in terms of RMSE, totaling almost 50% reduction in this metric compared to the median of survey-based expectations. Our findings indicate that news-based indexes were particularly helpful from the beginning of Brazil's Coronavirus pandemic, i.e., from 2020 onwards, a period of great economic and social instability.

Literature and contributions. The predictive power of Central Bank statements has been extensively investigated for forecasting interest rates (Hubert and Labondance, 2021), output growth (Lima, Godeiro, and Mohsin, 2021), inflation (Dräger, Lamla, and Pfajfar, 2016), and multiple variables (Lin, Fan, Zhang, and Chen, 2022). Newspapers articles have been used to analyze economic fluctuations and growth (Larsen and Thorsrud, 2019; Thorsrud, 2020), inflation and inflation expectations (Larsen, Thorsrud, and Zhulanova, 2021), output growth (Martins and Medeiros, 2022), and several macroeconomic variables (Rambaccussing and Kwiatkowski, 2020; Kalamara, Turrell, Redl, Kapetanios, and Kapadia, 2022; Barbaglia, Consoli, and Manzan, 2022). Specifically using Twitter data, Angelico, Marcucci, Miccoli, and Quarta (2022) build a daily indicator of expected inflation for Italy, a country that only possesses a monthly survey-based expectation. The developed index proved to be a good proxy for daily inflation expectations, and the authors point out that Twitter can opportunely reflect the beliefs of economic agents.

Similar to Lima, Godeiro, and Mohsin (2021), we computed indexes to then use them in forecast models. Conversely, Kalamara, Turrell, Redl, Kapetanios, and Kapadia (2022), for instance, directly utilized time series of the counts of terms, alongside other predictors, for forecasting. Literature has been highlighting the benefits of both approaches in enhancing predictive accuracy. An advantage of obtaining an index is its ability to increase the possibilities of various analyses beyond forecasting. A practitioner may only be interested in observing some pattern, anticipating a trend or a turning point. For that, considering a news-based index in their repertoire may be useful. Concerning the construction of the news-based index, the time-varying dictionary approach via supervised machine learning presents the advantage of its simplicity of implementation and interpretation since it is a procedure with a target variable. These features distinguish the approach from more complex topic modeling approaches, such as those based on Latent Dirichlet Allocation employed by Larsen and Thorsrud (2019), Thorsrud (2020), Larsen, Thorsrud, and Zhulanova (2021), and Martins and Medeiros (2022).

We can summarize the main contributions of this paper in the following four points. First, we propose alterations to the time-varying dictionary approach explored by Lima, Godeiro, and Mohsin (2021) for constructing our indexes for inflation using Twitter and newspapers. These alterations involve an alternative way of computing the indexes that employs the parameter estimates of the linear model used in selecting terms, smoothing through more recent fits as well as normalizing for stability over time. These changes can naturally be considered to obtain indexes for other economic variables. Second, our paper innovates by considering news-based indexes to forecast inflation *directly*, taking the Brazilian case as an application, thus extending the use of such an index compared to Angelico, Marcucci, Miccoli, and Quarta (2022). Brazilian inflation expectations from the Focus survey consist of expert opinions, most linked to the financial market. By showing that indexes based on tweets and articles help forecast inflation, a third contribution of our paper is to point out that information from a broader audience can be relevant to inflation prediction, an argument already defended in Angelico, Marcucci, Miccoli, and Quarta (2022) for Italy, for example. Fourth, we suggest a procedure for dealing with the secular increase in tweets over time, which, if not done, would contribute to artificially inflating the count of terms independent of the economic context.

Outline. This paper has four more sections in addition to this introduction. Section 2 details the Brazilian case and describes the news data as well as the construction of the news-based indexes for inflation. Section 3 describes the forecasting methodology employed to evaluate the contribution of the indexes to inflation forecasting. Section 4 analyses the adherence of the indexes to future inflation as well as presents and discusses the results of the forecasting exercises. Finally, section 5 concludes.

2 News-based indexes for the Brazilian inflation

2.1 The Brazilian context and the database for indexes

The Brazilian context. The *Instituto Brasileiro de Geografia e Estatística* (IBGE) computed the official Brazilian price index (IPCA) from which we compute the monthly inflation. The Central Bank of Brazil (BCB) manages the Focus survey, a daily-frequency expectation system based on expert opinion, mainly members of the financial market. The Focus survey collects expectations for several variables, including inflation, for multiple horizons. Although this survey has a daily periodicity, the current week's expectations are released to the public by the BCB only at the beginning of next week. Consequently, it is important to differentiate between the *available* Focus, which the econometrician observes when they compute their forecast, and the *ex-post* Focus, which is from the current day but will only be available days later. Thus, it is pertinent to know whether additional information generates more accurate forecasts for inflation at several horizons than Focus-based expectations. Furthermore, a survey-based expectation may reveal inside information unavailable to the econometrician and include signals not contained in other variables. In this context, it may be useful to use the available expectation as a predictor in a forecast model as well as to control for it to select which variables contribute at the margin to forecasting inflation.

Multi-horizon forecasts. We consider three forecast horizons: inflation accumulated over 3, 6, and 12 months ahead, which we indicate by `inf3m`, `inf6m`, and `inf12m`, respectively. These horizons can be relevant for managing monetary policy as well as pricing and investment make-decision.

Overview of indexes and data from Twitter and newspapers. The news-based indexes considered in this paper were developed in partnership with Vox Radar, a Brazilian technology company focused on monitoring social networks (social listening). We have daily data for both tweets and articles. For Twitter data, we counted mentions of various terms related to inflation in all tweets in Portuguese from 2010 onwards, disregarding tweets with terms about other economies such as "europa", "eua" (US), "fed", "alemanha" (Germany), "argentina", among other. The list of terms includes expressions about commodities, employment, exchange rate, expectations concerning prices and inflation, inflationary pressure, interest rates, investment, loans, costs, demand, supply shocks, taxes, and other macroeconomic-related terms. Some terms are similar to those found in Angelico, Marcucci, Miccoli, and Quarta (2022). We treat the data to control for Twitter usage over time. As is known, the utilization of Twitter experienced substantial growth in recent years. Consequently, there is a secular increase in tweets over time, which, if not accounted for, could artificially inflate the count of terms irrespective of the prevailing economic context. To mitigate this, we constructed a series of counts for generic terms and normalized each count of inflation-related terms by dividing it by the sum of counts of these generic terms. Table A1 in Appendix A presents the list of generic terms.

For newspapers, we counted 1-, 2-, and 3-grams (generically, n-grams) related to inflation after proceeding with tokenization, cleaning, and lemmatization of the articles obtained from three of the most relevant newspapers in Brazil (*Folha de São Paulo*, *Valor Econômico*, and *Estadão*), as done by Martins and Medeiros (2022). Tokenization divides the text into smaller units called tokens, usually comprising words and punctuation. Cleaning involves removing irrelevant elements such as stopwords, rare words, digits, and punctuation. Lemmatization reduces words to their base form. These procedures are widely used in the pre-processing of textual data. After this pre-processing, there are more than 36,000 n-grams. To reduce the universe of terms, we select those n-grams that contain specific words (or parts of words)[1]. Although the construction of the index employs a supervised machine learning method, which at first allows us to deal with the problem of dimensionality, including all this information would be counterproductive, besides the fact that many terms do not provide relevant information about future inflation.

We smoothed the series of counts by applying 132-day moving averages. This moving average aims to mitigate the effects that a great repercussion or unexpected increase of mentions of a certain term could have on obtaining the index. Other sizes of moving averages were investigated, but overall, 132 days produced good results. Note that 132 days corresponds to 6 months of business days. We also apply the transformation $\log(\text{count}_{i,t} + 1)$, where $\text{count}_{i,t}$ is the resulting moving average, with i indexing the n-grams, and t indicating the period. This transformation aims to mitigate possible asymmetries in the distribution of counts. Finally, the complete news database is ready to proceed with constructing the indexes for inflation, described below.

2.2 Construction of the indexes

Let be π_t the inflation rate at period t. Let be $\mathbf{news}_{t-h}$ a p-dimensional vector of the counts of terms of tweets and n-grams of newspaper articles observed at period $t-h$. These counts were obtained following the steps described in the previous subsection. At each period t and for each forecast horizon h, we estimate the linear model

$$\pi_t = \mu + \eta \,\text{Focus}_{t-h\,|\,t}^{\text{available}} + \boldsymbol{\phi}\,\mathbf{news}_{t-h} + \varepsilon_t, \tag{1}$$

where $\text{Focus}_{t-h\,|\,t}^{\text{available}}$ is the median of inflation expectations for period t from Focus survey *observed* at period $t-h$, ε_t is a projection error, and $(\mu, \eta, \boldsymbol{\phi}) \in \mathbb{R}^{p+2}$ is a vector of parameters. We estimate the model (1) employing the elastic net estimator. The estimator $\left(\widehat{\mu}, \widehat{\eta}, \widehat{\boldsymbol{\phi}}\right)$ for $(\mu, \eta, \boldsymbol{\phi})$ is the result of the problem

$$\min_{\mu, \eta, \boldsymbol{\phi}} \left\{ \sum_t \left(\pi_t - \mu - \eta \,\text{Focus}_{t-h\,|\,t}^{\text{available}} - \boldsymbol{\phi}\,\mathbf{news}_{t-h} \right)^2 + \lambda \left(\frac{1-\gamma}{2} \|\boldsymbol{\xi}\|_2^2 + \gamma \|\boldsymbol{\xi}\|_1 \right) \right\} \tag{2}$$

where $\boldsymbol{\xi} = (\eta, \boldsymbol{\phi})$, and λ and γ are hyperparameters.

[1] List of words (or parts of words), accompanied by the respective translations: "*preço*" (price), "*inflaç*" and "*inflac*" (root for inflation), "*ipca*" (Brazilian consumer price index), "*juro*" (interest), "*selic*" (Brazilian interest rate), "*demanda*" (demand), "*petróleo*" (oil), "*gasolina*" (gasoline), "*bacen*" and "*BC*" (Central Bank), "*commodit*" (root for commodities), "*camb*" and "*câmb*" (root for exchange rate), "*pib*" (GDP), and "*empreg*" (root for employment). We also include the 1-grams "*caged*" (a recording of hiring and firing employees in Brazil) and "*caro*" (expensive).

The presence of the available Focus improves the "stability" of the indexes. It is a guarantee that the selected terms may contribute in some way to predicting inflation beyond what is summarized by the Focus-based expectation. Moreover, employing the elastic net increases the probability that two relevant and highly correlated terms will be selected – compared to the LASSO, for example. For more advantages of using the elastic net, see Lima, Godeiro, and Mohsin (2021).

Finally, we compute two *updated* indexes from the most recent vector of news ($\mathbf{news}_t$) in "standardized" and "non-standardized" versions:

$$\texttt{index}_t^1 = \sum_{i=1}^{p} \widehat{\phi}_i \, \text{news}_{it} \; \in \mathbb{R}, \tag{3}$$

$$\texttt{index}_t^2 = \frac{\sum_{i=1}^{p} \widehat{\phi}_i \, \text{news}_{it}}{\sum_{i=1}^{p} |\widehat{\phi}_i \, \text{news}_{it}|} \; \in \; [-1, 1]. \tag{4}$$

Pros. Following, we list five benefits of the proposed methodology:

(i) Flexibility and adaptability for any variable of interest (with due care);

(ii) The past values of the index do not change, i.e., a new update in time does not modify the previous values of the index;

(iii) There is the automation of the selection of relevant terms, despite the need for pre-selection of n-grams of articles;

(iv) Possibility of relevant terms changing over time; the sign of the coefficient associated with a term can include change over time;

(v) The standardized version of the index is limited to the range of -1 to 1, which avoids significant instabilities over time.

Potential disadvantages or difficulties. Following, we highlight four potential complications of the methodology:

(i) Since the index is based on an estimation, the model may delay capturing new relevant terms or exclude terms that are no longer relevant;

(ii) Arbitrariness of the rolling window length used in the estimation. A smaller window can make it possible to enter new terms more quickly. However, there is an increase in estimation uncertainty (instability). On the other hand, using an extended window may not be attractive – in this case, it was not attractive in our paper for the Brazilian inflation;

(iii) It requires care so that the index is not unstable over time, especially in the non-standard version, which may show strange behavior at times. The inclusion of an intercept that varies in time can generate significant instability, for example;

(iv) Need to condition on the available survey-based expectation to ensure “stability”. In the absence of something like the Focus expectation, one could consider an autoregressive (AR) term, for example. Along the same lines, the inclusion of monthly dummies could also contribute to obtaining the index, for example. Although it seems like a con, conditioning on the available Focus has economic interpretation – as will be argued further on.

2.3 Setup and important considerations

Selection of hyperparameters. We pick the λ from a grid of one hundred values with exponential decay whose definition follows the default of the package `glmnet` for R. For γ, we choose it from a grid of ten values that grows logarithmically according to the sequence

$$\Big\{ \big(\log(1.01 + j \cdot 0.2) \big)^{0.25} : j = 0, 1, \ldots, 8 \Big\} \cup \{1\}.$$

Then, both hyperparameters are selected via Bayesian Information Criterion (BIC).

Sensitivity to the pre-selection of terms and number of terms. There is a certain instability of the index concerning the pre-selection of terms. To avoid increasing the possibilities, we considered the same pre-selection (broad) of article terms for all horizons addressed. By taking 16 (pieces of) terms and adding two more specific terms (see previous Footnote 1), we count 762 n-grams of newspaper articles. Regarding tweets, we considered the count of 397 terms. Therefore, we consider 1,159 terms in the estimation that originates the indexes. We consider only tweets, articles or both in the information set for constructing news-based indexes for inflation.

Intercept zero and instability. We investigated several possibilities to improve the adjustment of the indexes to future inflation. The indexes presented and analyzed here were obtained by setting $\mu = 0$ (intercept zero) in the model (1) for the three horizons considered. Since the intercept varies (considerably) over time, it causes an increase in the “instability” of the indexes, which deteriorates the indexes visually and in terms of contribution to forecast performance. In a way, conditioning the model to some variable that generates stability in the estimation (such as controlling for the available Focus expectation or AR terms, for example) makes the requirement of the intercept dispensable.

Controlling for the available Focus. As previously mentioned, the presence of the available Focus expectation is necessary to guarantee the “adherence” of the indexes to future inflation rates. Furthermore, controlling for the available Focus survey generates an interesting economic interpretation: we managed to make the method include terms that generate “marginal gain” for the inflation adjustment after considering relevant available information from a survey. In other words, controlling for the Focus allows the estimator to select terms that capture the *inflationary surprise*, which may contribute to the relevance of the indexes in forecasting inflation.

Smoothing via averaging of fits of several models. A potential source of instability for the index is abrupt changes in the selection of terms by elastic net across the rolling windows. To alleviate this difficulty, we consider predictions from fits of models estimated in previous windows (with all models being evaluated in the most recent news vector). Formally, with $\widehat{\mathcal{M}}^{t-j}$ being the estimated model considering the period ending in $t-j$, we compute a "smoothed version" of the index via a simple average of the adjustments generated by evaluating each estimated model in the most recent vector of terms:

$$\texttt{index}_t^{i,s} = \frac{1}{J}\sum_{j=0}^{J-1} \widehat{\mathcal{M}}^{t-j}(\mathbf{news}_t), \qquad i \in \{1,2\},$$

where J is the number of fits we consider. Note that if we consider only the most recent fits, we will be left with $\widehat{\mathcal{M}}^{t}(\mathbf{news}_t)$, that is, one of the original versions presented in (3) and (4). Smoothing was necessary mainly for the 12-month cumulative inflation index. We consider the mean of the six most recent adjustments for all horizons. In general, this was the choice that generated the best results.

3 Evaluation of the relevance of news-based indexes

In addition to visually inspecting news-based indexes and comparing them to actual inflation, we conduct pseudo-out-of-sample forecasting exercises with models that include or exclude them. Besides the natural benchmarks delivered by the Focus survey, we consider the four models described following to verify the usefulness of news-based indexes in forecasting. For the presentation of the models, consider the following variable definitions:

- π_t is the cumulative inflation over h periods (months) at the period t;
- $\text{Focus}_{t+h\,|\,t}^{\text{available}}$ is the median of the Focus survey inflation expectations accumulated for h periods ahead and available at the period t;
- u_t is a forecast error;
- $\widehat{\pi}_{T+h\,|\,T}$ is a forecast for h-period-ahead cumulative inflation based on information *available* at T.

Bias correction via OLS. Following Mincer and Zarnowitz (1969), we take a linear model that considers both intercept (α) and slope (β) historical bias for a forecast. In particular, we are interested in the *available* Focus-based inflation expectation. Thus, we have the following model:

$$\pi_t = \alpha + \beta\,\text{Focus}_{t\,|\,t-h}^{\text{available}} + u_t, \qquad t = 1, \ldots, T-h.$$

After the estimation of the parameters employing least squares, we were able to obtain a forecast that corrects for historical bias for the period T by computing

$$\widehat{\pi}_{T+h\,|\,T} = \widehat{\alpha} + \widehat{\beta}\,\mathrm{Focus}^{\text{available}}_{T+h\,|\,T},$$

in which $\widehat{\alpha}$ and $\widehat{\beta}$ are OLS estimates.

Bias correction including news-based indexes. We can augment the previous simple bias correction model by adding indexes based on tweets and newspaper articles to test the forecasting performance. Thus, for each index in $\{\mathtt{index}_t^{i,s} : i \in \{1,2\},\, s \in \{\text{smooth}, \text{not smooth}\}\}$, we define the model

$$\pi_t = \alpha + \beta\,\mathrm{Focus}^{\text{available}}_{t\,|\,t-h} + \theta\,\mathtt{index}^{i,s}_{t-h} + u_t, \qquad t = 1, \ldots, T-h.$$

As before, we compute a forecast via

$$\widehat{\pi}_{T+h\,|\,T} = \widehat{\alpha} + \widehat{\beta}\,\mathrm{Focus}^{\text{available}}_{T+h\,|\,T} + \widehat{\theta}\,\mathtt{index}^{i,s}_{t-h},$$

in which the coefficients with hat are least squares estimates.

Data-rich environment and estimation via adaptive LASSO (adaLASSO). Previous models have the limitation of not including other potential predictors for inflation (macroeconomic variables, for example). Thus, we can consider including a large number of predictors, including their lags (about this, see Inoue and Kilian, 2008; Garcia, Medeiros, and Vasconcelos, 2017; Medeiros, Vasconcelos, Veiga, and Zilberman, 2021). Defining $\boldsymbol{x}_{t-h}$ to be a p-dimensional vector with such variables *available* at period $t-h$, we can write a general linear model as follows:

$$\pi_t = \alpha + \beta\,\mathrm{Focus}^{\text{available}}_{t\,|\,t-h} + \boldsymbol{\gamma}\,\boldsymbol{x}_{t-h} + u_t, \qquad t = 1, \ldots, T-h,$$

in which $\boldsymbol{\gamma}$ is a p-dimensional vector of parameters.

However, as the number of predictors exceeds the number of temporal observations, we must resort to machine learning methods. To display results, we opted for the adaptive LASSO (adaLASSO), which deals with the curse of dimensionality by selecting predictors.[2] After estimating the model, we calculate the forecast based on the latest available information, as previously done. Appendix B provides a description of the adaLASSO.

[2] We also consider other models such as LASSO, complete subset regression (CSR), and Random Forest (that admits nonlinearities). However, the adaLASSO performed better than the LASSO for inflation accumulated in 12 months, and both obtained a similar performance in the other horizons. Concerning the others, the adaLASSO was superior.

adaLASSO including news-based indexes. Finally, we also added a news-based index in linear model estimated via adaLASSO to verify the potential gains in forecast performance. The variable selection properties of the adaLASSO play an important role since it empirically determines whether or not indexes should be selected. Combined with evaluating forecasts based on a metric – e.g., root mean squared error (RMSE) – this will attest to the relevance (or not) of considering the indexes in a data-rich environment. In this case, for each $\{\texttt{index}_t^{i,s} : i \in \{1,2\},\, s \in \{\text{smooth}, \text{not smooth}\}\}$, the model to be estimated is given by

$$\pi_t = \alpha + \beta\, \text{Focus}_{t\,|\,t-h}^{\text{available}} + \boldsymbol{\gamma}\, \boldsymbol{x}_{t-h} + \theta\, \texttt{index}_{t-h}^{i,s} + u_t, \qquad t = 1, \ldots, T-h.$$

Again, following the same logic reported so far, we compute our forecasts.

Pseudo-out-of-sample exercise (setup). We set expanding windows to compute multi-horizon inflation forecasts starting in Jan/2019 and ending in Jul/2022. Thus, we compute 43 forecasts for each horizon. In the case of linear models estimated via the adaLASSO, we consider three lags for each time-varying predictor and include month dummies. Finally, we use root mean square error (RMSE) as a metric and the Diebold-Mariano test to assess the forecast performance of the models that include or do not include a news-based index. We consider a one-tailed Diebold-Mariano test with null and alternative hypothesis given by $\mathbb{H}_0 : \text{MSE}\big(\widehat{\pi}^1_{t+h\,|\,t}\big) = \text{MSE}\big(\widehat{\pi}^2_{t+h\,|\,t}\big)$ and $\mathbb{H}_1 : \text{MSE}\big(\widehat{\pi}^1_{t+h\,|\,t}\big) < \text{MSE}\big(\widehat{\pi}^2_{t+h\,|\,t}\big)$, where $\widehat{\pi}^1_{t+h\,|\,t}$ indicates the model does not include an index and $\widehat{\pi}^2_{t+h\,|\,t}$ indicates the model including an index.

4 Results

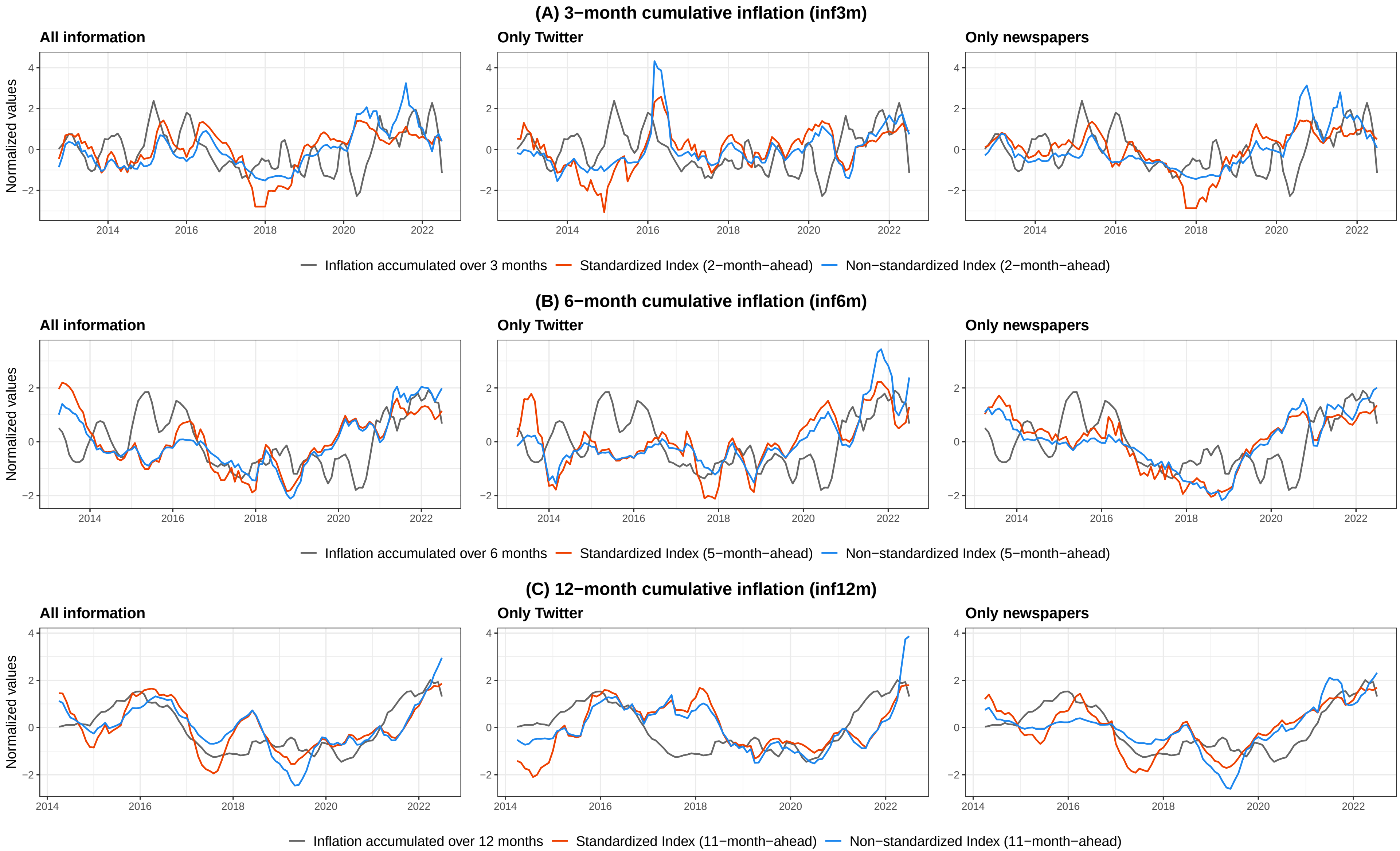
4.1 Visual inspection of the indexes

Figure 1 presents the actual inflation accumulated over 3, 6, and 12 months and news-based indexes in their different versions: standardized and non-standardized, and considering different sets of information – all information, only Twitter, and only newspapers. Each horizon is displayed in a row, and each information set is in a column. To facilitate the comparison, actual inflations (gray lines), as well as both standardized (red lines) and non-standardized (blue lines) indexes, were normalized over each period. We advance the indexes in time according to the horizon to compare them with the respective inflations. The start date differs for different horizons because we need more initial information in index construction (estimation) for longer horizons. Notice that standardized and non-standardized versions of the indexes often exhibit dissonant movements, reaffirming the relevance of considering both ways of constructing the index and verifying which better suits each situation. Despite indexes sometimes presenting discrepant magnitudes when compared to the respective actual inflations, we should verify whether the indexes can capture trends and track inflation's movements up and down over time.

For the inflation accumulated over 3 months, the indexes based on all information better capture inflation movements. The non-standardized index that uses all information tends to better track the ups and downs of inflation, especially from 2020 onwards. Even considering the smoothing via the average of different fits, we note that it was not possible to completely mitigate the noisy behavior that persisted over time in virtually all indexes for this horizon. In addition, the isolated peak of the non-standardized index based only on tweets in 2016 is a negative highlight. For 6-month cumulative inflation, all indexes adhere reasonably to the long-term inflation trend. However, they do not capture shorter-term cycles well. For this horizon, all indexes show trends that differ from the inflation realized in 2013. Finally, regarding inflation accumulated during 12 months, indexes show more abrupt fluctuations than inflation, but most of them capture the smooth ups and downs of the serie. A considerable divergence occurred in the magnitude and trend of the standardized index based solely on Twitter over 2014. Additionally, the non-standardized index considering only Twitter shows a considerable increase throughout 2022, which may indicate the relevance of the standardized version of the index to attenuate such situations. Visually, the best fit belongs to indexes that consider all information, i.e., join tweets and articles.

Despite difficulties anticipating some movements of inflation, news-based indexes have potential, and their consideration can contribute to decision-making regarding the prognostic of future inflation dynamics. Some movements not captured by the indexes, such as the sharp decline in 3- and 6-month cumulative inflation at the start of 2020 (at the beginning of the Corona Virus pandemic), are difficult to anticipate. On the other hand, it is worth highlighting that most of the indexes captured well the trend of increasing accumulated inflation over 6 and 12 months from 2021 onwards. From this period, the median of inflation expectations collected by the Central Bank of Brazil began to underestimate future inflation significantly (see Boaretto and Medeiros, 2023). Additionally, note that indexes based on all information or only on articles for accumulated inflation in the next 12 months efficiently anticipated the rapid disinflation during the second half of 2016 and the first half of 2017. The median of the Focus does not reasonably anticipate this rapid decline in inflation. Econometric models also do not easily anticipate it, even in an information-rich environment, as Boaretto and Medeiros (2023) pointed out. Following, we investigate the benefits of the employ of news-based indexes in pseudo-out-of-sample forecasting exercises.

Figure 1: News-based indexes and inflation, by horizon and information set



4.2 Evaluation of the predictive contribution of indexes

We generated 45 out-of-sample predictions for the 3-, 6-, and 12-month cumulative inflation, covering January 2019 to July 2022. Table 1 displays the forecast performance in terms of RMSE for *available* Focus (last available median expectation when we compute our forecasts), *ex-post* Focus (median expectation of the reference day, but released only days later), and addressed models that include or not a news-based index for inflation. We report the RMSE ratio using the available Focus RMSE as a reference point. Thus, for the three forecast horizons, the RMSE ratio of the available Focus is 1. If the RMSE ratio is less than 1, then the model performed better than the available Focus, and if greater than 1, then the model performed worse than the available Focus. From Panel A of Table 1, we notice that the *ex-post* Focus, which is more updated, improves the predictive performance slightly compared to the available Focus for all horizons considered. This result is expected since the experts have more updated information on the *ex-post* Focus. We also expected that the performance improvement would drop with the horizon increase since there is little relevant informational gain between a few days when looking at a longer horizon.

Table 2 reports the relative frequency in which the adaLASSO (high-dimensional model) automatically selects a news-based index. Figure 2 exhibits actual inflation and forecasts by the horizon (figure on the left) as well as the squared forecast errors (figure on the right) of main models/expectations. Each horizon appears in a different panel (from A to C). For the inflation accumulated over 3 months (`inf3m`), bias correction models for available Focus estimated by OLS, adding or not a news-based index as an extra predictor, registered the best performances. The RMSE reductions in comparison to the available median of the Focus range from 9% to 36%. For these low-dimensional models, including a news-based index contributes to a further reduction of up to 17 percentage points in relative RMSE, considering the non-standardized index based only on articles. The good performance of the low-dimensional model that includes indexes is mainly due to the reduction of (squared) forecast errors from the second half of 2021 (see Figure 2, Panel A). In its turn, in high-dimensional models estimated by the adaLASSO, from Table 2, we notice that the adaLASSO hardly selects news-based indexes. Non-standardized based only on articles was selected only approximately 14% of the time, which led to a reduction of 2 p.p. on relative RMSE, but statically not significant according to a one-tailed Diebold-Mariano test.

For the inflation accumulated during 6 months (`inf6m`), the bias correction model not including any news-based index led to a small reduction in RMSE (4%), which increased when an index was included (maximum reduction of almost 13%). This result highlights the contribution of a news-based index in a low-dimensional case. Intuitively, our news-based indexes still have predictive power conditional on the experts' available expectations. In contrast, high-dimensional models exhibited superior forecast performance, resulting in a significant decrease in RMSE of at least 24% relative to the available Focus. However, news-based indexes did not exhibit robust predictive power due to their infrequent selec-

Table 1: Out-of-sample RMSE with respect to available Focus

		inf3m	inf6m	inf12m
	A. Survey			
Focus	Available	1.000	1.000	1.000
	Ex-post	0.960	0.984	0.996
	B. Bias correction			
OLS (no index)		0.910	0.960	1.136
Including a non-std index	All information	*0.805* ***	0.859 ***	1.148
	Only tweets	0.920	0.906 **	1.174
	Only articles	**0.740** ***	0.863 ***	1.254
Including a std index	All information	0.881 ***	0.887 ***	0.901 ***
	Only tweets	0.917	0.928 ***	1.150
	Only articles	*0.843* ***	0.874 ***	0.813 ***
	C. High-dimensional model			
adaLASSO (no index)		0.939	*0.761*	*0.614*
Including a non-std index	All information	0.939	*0.761*	0.721
	Only tweets	0.939	**0.758**	0.615
	Only articles	0.917	*0.761*	*0.614*
Including a std index	All information	0.939	*0.761*	0.623
	Only tweets	0.939	*0.760*	**0.504** **
	Only articles	0.939	*0.761*	0.659

Notes: Forecasts covering the period from January/2019 to July/2022. The highlighted value in blue bold indicates the best result for each forecast horizon in terms of out-of-sample RMSE, while blue italics indicate the second- and third-best results. ***, **, and * indicate that a specific model that includes a news-based index performed statistically better than the corresponding model that did not include the index in a one-tailed Diebold-Mariano test at significance level of 10, 5, and 1%, respectively.

tion, except for indexes based solely on tweets. Among these, the non-standardized version was chosen 9.3% of the time by the adaLASSO, while the standardized version was chosen only once out of 43 time periods. Despite this, including these indexes did not result in a significant reduction in RMSE compared to the high-dimensional model that excluded them. A possible intuition for this result is that when we control for a more extensive information set, news-based indexes lose their relevance for the analyzed horizon, suggesting that the other variables already capture the same information.

The accuracy of the 12-month cumulative inflation forecast (`inf12m`) deteriorates when a historical bias correction is applied to the available expectation. The results in Table 1 indicate an increase of more than 13% in RMSE compared to the available Focus. The situation worsens when each of the three non-standardized indexes is considered. However, standardized indexes based on all information or only on newspaper articles led to a substantial reduction in RMSE, ranging from 10% to 19%, compared to the available Focus. These im-

Table 2: Selection of news-based indexes by the adaLASSO (%)

	inf3m		inf6m		inf12m	
	Non-std	Std	Non-std	Std	Non-std	Std
All information	–	–	–	–	60.47	100.00
Only tweets	–	–	9.30	2.33	39.53	100.00
Only Articles	13.95	–	–	–	–	93.02

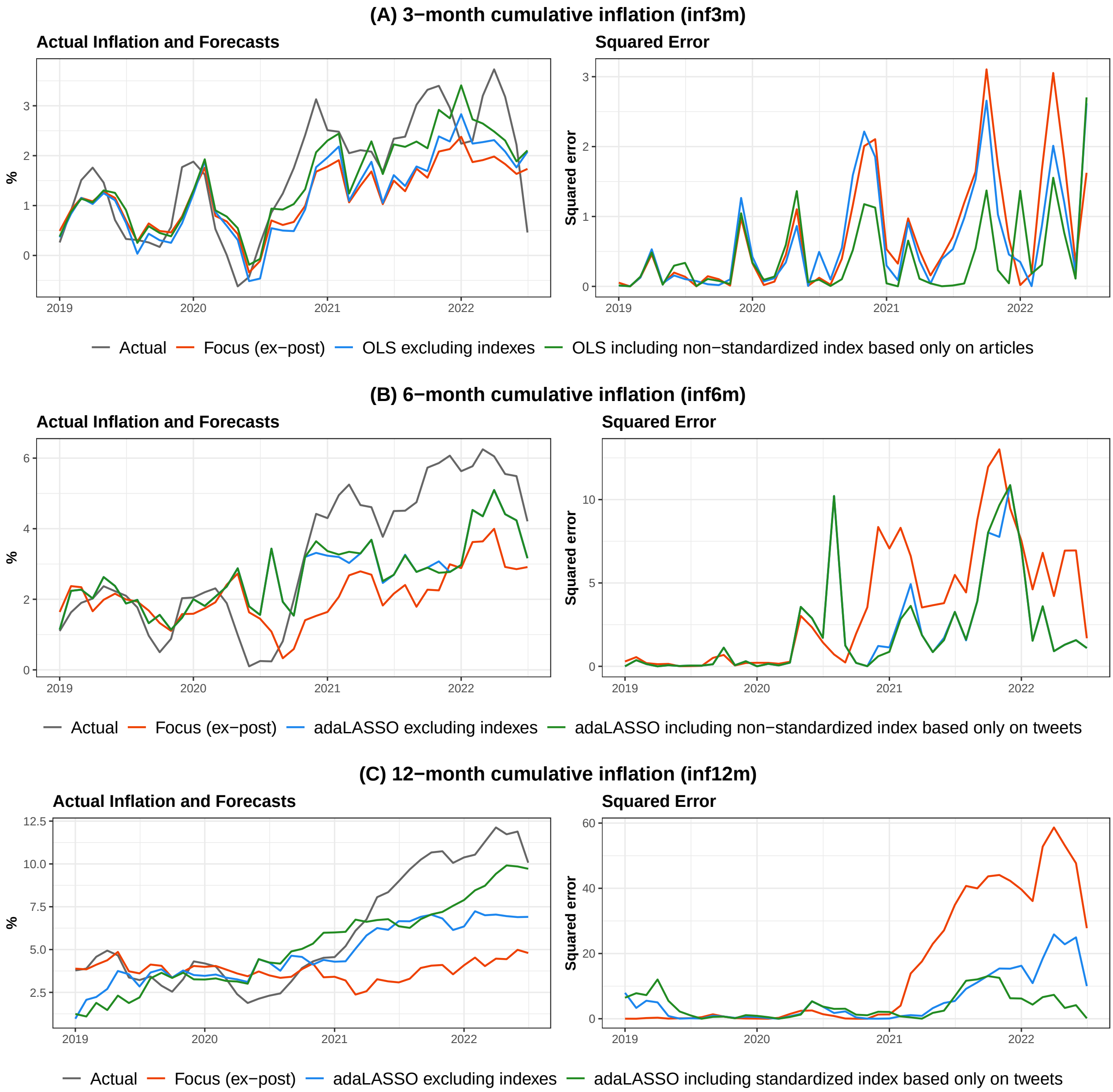
Notes: "–" indicates that the adaLASSO did not select a specific index for any period. $T = 43$.

provements are statistically significant, as attested by a one-tailed DM test. These findings underscore the benefits of implementing discipline through standardized versions of the indexes, given that the series of the count of terms can be volatile even with some smoothing applied. When we consider a large number of predictors in a linear model estimated employing adaLASSO, there is an expressive reduction of almost 39% in terms of RMSE. In this context, only the standardized index based solely on Twitter information can deliver an even better result: a reduction of almost 50% in RMSE, which is a decrease of 11 percentage points compared to the model that did not include any index. Notably, this index was automatically picked in 100% of the opportunities, as shown in Table 2. Moreover, according to Panel C of Figure 2, the predictive improvements come from better forecasts starting from 2021.

The results of the pseudo-out-of-sample forecasting exercises indicate that news-based indexes were particularly useful during periods of high instability[3], such as the onset of the COVID-19 pandemic in 2020 and onwards. As shown in Figure 2, except for the accumulated inflation in 6 months, the indexes significantly reduced the squared forecast error from the second half of 2020. Panel C of Figure 2 also suggests that disregarding the inaccurate forecasts generated for early 2019, the high-dimensional models, including or not news-based indexes, would deliver an even greater RMSE reduction. Another result indicates that smaller models performed better for the shorter horizon of 3 months, whereas models incorporating several predictors performed better for the longer horizon of 12 months. This result may have occurred because there is little room to improve the predictive performance of a survey-based expectation as we shorten the forecast horizon. Moreover, while a restricted model using the "right variables" still generates some improvement concerning the inflation expectation in shorter horizons, a more extensive model is more susceptible to specification errors and estimation uncertainty. However, in long horizons, there is room for the effective contribution of other predictors – including news-based indexes, even considering an information-rich environment.

[3]This result is similar to the finding by Kalamara, Turrell, Redl, Kapetanios, and Kapadia (2022), who used newspaper data to forecast GDP, inflation, and unemployment in the United Kingdom.

Figure 2: Inflation forecasts and squared forecast errors, by horizon



5 Conclusion

This study presents novel approaches to constructing forward-looking inflation indexes utilizing data from Twitter and newspapers through a supervised learning method shown as time-varying dictionary approach. Considering the Brazilian case and different horizons for cumulative inflation, our news-based indexes were able to anticipate long-term trends. Furthermore, they captured short-term movements in inflation at various periods. We also highlight the benefits of news-based indexes for inflation forecasting by conducting pseudo-out-of-sample exercises. News-based indexes can improve forecast performance for different horizons. For short ones (3 and 6 months ahead), a low-dimensional model that considers the median of expectations from a survey as the unique predictor benefits from including news-based indexes. On the other hand, for larger horizons (12 months ahead), high-dimensional models, which incorporate many predictors, can also be improved by incorporating these indexes, at least marginally. Thus, incorporating news-based indexes from social media and news sources can improve inflation forecasting.

There are several possibilities for extending the results of this paper that can be investigated in future research. The most natural extension is to look at sub-components of a price index and predict their variations individually. Since different disaggregates have specific characteristics and some are more difficult to predict, indexes based on tweets and articles can be interesting in predicting future values of these disaggregations. Moreover, one potential avenue of exploration beyond inflation forecasting is to utilize news-based indexes to model and predict demand for various goods and services.

Appendix A Terms and Variables

Generic terms on Twitter. Table A1 contains the generic terms whose count was used to normalize the count of terms related to inflation over time in order to control for the secular trend in the number of tweets. The translations to English are also presented.

Table A1: List of generic terms and their translations

Generic term	Translation	Generic term	Translation
oi	hi	ok	okay
olá	hello	sim	yes
bom dia	good morning	não	no
boa noite	good night	galera	folks
boa tarde	good afternoon	bora	let's go (slang)
escrever	to write	fazer	to do, to make
ler	to read	valeu	thanks (slang)
vamos	let's go	obrigado	thanks, thank you

Other predictors. In addition to news-based indexes, we consider the *available* Focus-based inflation expectation, seasonal dummies, and eighty more time-varying variables and their respective lags as predictors for inflation. These variables can be divided into ten categories: Prices and Money (19), Commodities Prices (4), Economic Activity (9), Employment (5), Electricity (4), Confidence (3), Finance (12), Credit (4), Government (12), and Exchange and International Transactions (9). The choice of the variables was similar to the variables used in Garcia, Medeiros, and Vasconcelos (2017).

Table A2 presents a description of all variables as well as the transformations implemented to guarantee the stationarity of the series. To get as close as possible to a real-time database, we considered the average disclosure delay of each variable. The penultimate column of Table A2 contains this information. We consider the last day of each month as the reference day on which multi-period forecasts are computed.

Table A2: Description of predictor variables

#	Variable	Description	Unit	Source	Lag	Transformation
		A. Prices and Money				
1	inf	Consumer Price Index (IPCA)	index	IBGE	1	% change
2	expec	Focus-based inflation expectations (*available*)	% per month	BCB	0	–
3	ipca15	Consumer Price Index - 15 (IPCA-15)	index	IBGE	0	% change
4	inpc	Consumer Price Index (INPC)	index	BCB	1	% change
5	ipc	Consumer Price Index - Brazil (IPC-Br)	index	FGV	1	% change
6	igpm	General Price Index - M (IGP-M)	index	FGV	1	% change
7	igpdi	General Price Index - DI (IGP-DI)	index	FGV	1	% change
8	igp10	General Price Index - 10 (IGP-10)	index	FGV	1	% change
9	ipc_fipe	Fipe Consumer Price Index (IPC-Fipe)	index	Fipe	1	% change
10	ipa	Wholesale Price Index (IPA)	index	FGV	1	% change
11	ipa_ind	IPA – industrial Products	index	FGV	1	% change
12	ipa_agr	IPA – agricultural Products	index	FGV	1	% change
13	incc	National Index of Building Costs (INCC)	index	FGV	1	% change
14	bm_broad	Broad Monetary Base – end-of-period balance	index	BCB	2	% change
15	bm	Monetary Base – working day balance average	Index	BCB	2	% change
16	m1	Money supply M1 – working day balance average	Index	BCB	2	% change
17	m2	Money supply M2 – end-of-period balance	Index	BCB	2	% change
18	m3	Money supply M3 – end-of-period balance	Index	BCB	2	% change
19	m4	Money supply M4 – end-of-period balance	Index	BCB	2	% change
		B. Commodities prices				
20	icbr	Brazilian Commodity index – all	index	BCB	1	% change
21	icbr_agr	Brazilian Commodity index – agriculture	index	BCB	1	% change
22	icbr_metal	Brazilian Commodity index – metal	index	BCB	1	% change
23	icbr_energy	Brazilian Commodity index – energy	index	BCB	1	% change
		C. Economic Activity				
24	ibcbr	Brazilian IBC-Br Economic Activity index	index	BCB	3	% change
25	month_gdp	GDP monthly – current prices	R$ million	BCB	1	% change
26	tcu	Use of installed capacity – manufacturing industry	%	FGV	1	first difference
27	pimpf	Industrial Production – general	index	IBGE	2	% change
28	pmc	Retail sales volume – total	index	IBGE	2	% change
29	steel	Steel production	index	BCB	1	–
30	prod_vehicles	Vehicle production – total	units	Anfavea	1	% change
31	prod_agr_mach	Production of agricultural machinery – total	units	Anfavea	1	% change
32	vehicle_sales	Vehicle sales by dealerships – total	units	Fenabrave	1	% change
		D. Labor Market				
33	unem	Unemployment (combination of PME and PNADC)	%	IBGE	3	first difference
34	employment	Registered employess by economic activity - Total	units	IBGE	1	first difference
35	aggreg_wage	Overall Earnings (broad wage income)	R$ (million)	IBGE	2	% change
36	min_wage	Federal Minimum Wage	R$	MTb	0	% change
37	income	Households gross disposable national income	R$ (million)	BCB	2	% change
		E. Electricity				
38	elec	Electricity consumption - total	GWh	Eletrobrás	3	% change
39	elec_res	Electricity consumption - residential	GWh	Eletrobrás	3	% change
40	elec_com	Electricity consumption - commercial	GWh	Eletrobrás	3	% change
41	elec_ind	Electricity consumption - industry	GWh	Eletrobrás	3	% change
		F. Confidence				
42	cons_confidence	Consumer Confidence index	index	Fecomercio	1	% change
43	future_expec	Future expectations index	index	Fecomercio	1	% change
44	conditions	Current economic conditions index	index	Fecomercio	1	% change
		G. Finance				
45	ibovespa	Ibovespa index	% per month	BM&FBOVESPA	1	–
46	irf_m	Anbima Market Index of the prefixed federal bonds	index	Anbima	1	% change
47	ima_s	Anbima Market Index of the federal bonds tied to the SELIC rate	index	Anbima	1	% change
48	ima_b	Anbima Market Index of the federal bonds tied to the IPCA index	index	Anbima	1	% change
49	ima	General Anbima Market index	index	Anbima	1	% change
50	saving_deposits	Savings deposits - end-of-period balance	R$ (mil)	BCB	2	% change
51	selic	Selic Basic Interest rate	% per month	BCB	1	–
52	cdi	Cetip DI Interbank Deposits rate	% per month	Cetip	1	–
53	tjlp	TJLP Long Term Interest rate	% per year	BCB	1	–
54	ntnb	3-Year Treasury (real) Rate indexed to the IPCA (NTN-B)	% per year	Anbima	0	–
55	embi	Emerging Markets Bond Index Plus – Brazil	b.p. acc. month	JP Morgan	0	first difference
56	vix	CBOE Volatility Index (VIX)	index	CBOE	0	
		H. Credit				
57	cred_total	Credit outstanding - total	R$ (million)	BCB	2	% change
58	cred_dgp	Credit outstanding as a percentage of GDP	% of GDP	BCB	2	first difference
59	indebt_house1	Household debt to income ratio – all	% of 12m income	BCB	2	first difference
60	indebt_house2	Household debt to income ratio without mortgage loans	% of 12m income	BCB	2	first difference
		I. Government				
61	net_debt_gdp	Net public debt (% GDP) - Consolidated public sector	% of GDP	BCB	2	first difference
62	net_debt	Net public debt - Total - Consolidated public sector	R$ (million)	BCB	2	first difference
63	net_debt_fedgov_bcb	Net public debt - Federal Government and Central Bank	R$ (million)	BCB	2	first difference
64	net_debt_states	Net public debt - State governments	R$ (million)	BCB	2	first difference
65	net_debt_cities	Net public debt - Municipal governments	R$ (million)	BCB	2	first difference
66	primary_result	Primary result - Consolidated public sector	R$ (million)	BCB	2	first difference
67	debt_fedgov_old	Gross general government debt - Method used until 2007	R$ (million)	BCB	2	% change
68	debt_fedgov_new	Gross general government debt - Method used since 2008	R$ (million)	BCB	2	% change
69	treasury_emit	National Treasury domestic securities - Total issued	R$ (million)	BCB	2	% change
70	treasury_mkt	National Treasury domestic securities - Total on market	R$ (million)	BCB	2	% change
71	treasury_term	National Treasury securities debt - medium term	months	BCB	2	first difference
72	treasury_dur	National Treasury securities debt - medium duration	months	BCB	2	first difference
		J. Exchange and International Transactions				
73	reer	Real Effective Exchange Rate	R$/other	BIS	2	% change
74	usd_brl_end	USD-BRL rate – end of period	R$/US$	BCB	0	% change
75	usd_brl_av	USD-BRL rate – monthly average	R$/US$	BCB	0	% change
76	eur_brl_end	EUR-BRL rate – end of period	R$/€	Bloomberg	0	% change
77	eur_brl_av	EUR-BRL rate – monthly average	R$/€	Bloomberg	0	% change
78	current_account	Current account – net	US$ (million)	BCB	2	% change
79	trade_balance	Balance on goods and services - net (Brazilian trade balance)	US$ (million)	BCB	2	% change
80	exports	Imports	US$ (million)	BCB	2	% change
81	imports	Exports	US$ (million)	BCB	2	% change

Appendix B Adaptive LASSO (adaLASSO)

Consider a predictive linear model given by $\pi_t = \boldsymbol{\beta}\, \boldsymbol{x}_{t-h} + \varepsilon_t$, in which π_t is inflation at period t, $\boldsymbol{x}_{t-h}$ is a J-dimensional vector of predictors (and their lags) *observed* at period $t-h$, and ε_t is a forecast error. We can estimate the parameter vector $\boldsymbol{\beta}$ via adaptive LASSO (adaLASSO). Introduced by Zou (2006), this method solves

$$\widehat{\beta}_{\text{adaLASSO}}(\lambda, \boldsymbol{\omega}) = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ \frac{1}{T-h} \sum_{t=1}^{T-h} (\pi_t - \alpha - \boldsymbol{\beta}\, \boldsymbol{x}_{t-h})^2 + \lambda \sum_{j=1}^{J} \omega_j \, |\beta_j| \right\}$$

in which λ is a regularization parameter, and $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_J)$ is a vector of weights obtained previously via LASSO, an estimator that assumes $\omega_j = 1$ for all j (see Tibshirani, 1996). More precisely, we compute the adaLASSO weights via

$$\omega_j = \left(\left| \widehat{\beta}_{\text{LASSO},j} \right| + \frac{1}{\sqrt{T}} \right)^{-1},$$

in which the presence of $T^{-1/2}$ makes possible a variable that the LASSO had not selected in the first stage, i.e., the case in which $\beta_{\text{LASSO},j} = 0$, can be selected by the adaLASSO.

Finally, we get an h-periods-ahead forecast by computing $\widehat{\pi}_{t+h\,|\,t} = \widehat{\alpha} + \widehat{\boldsymbol{\beta}}_{\text{adaLASSO}}\, \boldsymbol{x}_t$.

References

ANGELICO, C., J. MARCUCCI, M. MICCOLI, AND F. QUARTA (2022): "Can we measure inflation expectations using Twitter?," *Journal of Econometrics*, 228(2), 259–277.

BARBAGLIA, L., S. CONSOLI, AND S. MANZAN (2022): "Forecasting with economic news," *Journal of Business & Economic Statistics*, pp. 1–12.

BOARETTO, G., AND M. C. MEDEIROS (2023): "Forecasting Aggregate Inflation using Disaggregates and Machine Learning," Unpublished manuscript.

DRÄGER, L., M. J. LAMLA, AND D. PFAJFAR (2016): "Are survey expectations theory-consistent? The role of central bank communication and news," *European Economic Review*, 85, 84–111.

GARCIA, M. G., M. C. MEDEIROS, AND G. F. VASCONCELOS (2017): "Real-time inflation forecasting with high-dimensional models: The case of Brazil," *International Journal of Forecasting*, 33(3), 679–693.

GUZMAN, G. (2011): "Internet search behavior as an economic forecasting tool: The case of inflation expectations," *Journal of Economic and Social Measurement*, 36(3), 119–167.

HUBERT, P., AND F. LABONDANCE (2021): "The signaling effects of central bank tone," *European Economic Review*, 133, 1–27.

INOUE, A., AND L. KILIAN (2008): "How Useful Is Bagging in Forecasting Economic Time Series? A Case Study of U.S. Consumer Price Inflation," *Journal of the American Statistical Association*, 103(482), 511–522.

KALAMARA, E., A. TURRELL, C. REDL, G. KAPETANIOS, AND S. KAPADIA (2022): “Making text count: economic forecasting using newspaper text,” *Journal of Applied Econometrics*, 37, 896–919.

LARSEN, V. H., AND L. A. THORSRUD (2019): “The value of news for economic developments,” *Journal of Econometrics*, 210(1), 203–218.

LARSEN, V. H., L. A. THORSRUD, AND J. ZHULANOVA (2021): “News-driven inflation expectations and information rigidities,” *Journal of Monetary Economics*, 117, 507–520.

LI, X., W. SHANG, S. WANG, AND J. MA (2015): “A MIDAS modelling framework for Chinese inflation index forecast incorporating Google search data,” *Electronic Commerce Research and Applications*, 14(2), 112–125.

LIMA, L. R., L. L. GODEIRO, AND M. MOHSIN (2021): “Time-varying dictionary and the predictive power of FED minutes,” *Computational Economics*, 57(1), 149–181.

LIN, J., J. FAN, Y. ZHANG, AND L. CHEN (2022): “Real-time macroeconomic projection using narrative central bank communication,” *Journal of Applied Econometrics*, pp. 1–20.

MARTINS, L. C. L., AND M. C. MEDEIROS (2022): “Nowcasting GDP with News and Google Trends,” Unpublished manuscript.

MEDEIROS, M. C., G. F. VASCONCELOS, Á. VEIGA, AND E. ZILBERMAN (2021): “Forecasting inflation in a data-rich environment: the benefits of machine learning methods,” *Journal of Business & Economic Statistics*, 39(1), 98–119.

MINCER, J. A., AND V. ZARNOWITZ (1969): “The Evaluation of Economic Forecasts,” in *Economic forecasts and expectations: Analysis of forecasting behavior and performance*, pp. 3–46. NBER.

NIESERT, R. F., J. A. OORSCHOT, C. P. VELDHUISEN, K. BRONS, AND R.-J. LANGE (2020): “Can Google search data help predict macroeconomic series?,” *International Journal of Forecasting*, 36(3), 1163–1172.

RAMBACCUSSING, D., AND A. KWIATKOWSKI (2020): “Forecasting with news sentiment: Evidence with UK newspapers,” *International Journal of Forecasting*, 36(4), 1501–1516.

THORSRUD, L. A. (2020): “Words are the New Numbers: A Newsy Coincident Index of the Business Cycle,” *Journal of Business & Economic Statistics*, 38(2), 393–409.

TIBSHIRANI, R. (1996): “Regression Shrinkage and Selection via the Lasso,” *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1), 267–288.

ZOU, H. (2006): “The Adaptive Lasso and Its Oracle Properties,” *Journal of the American Statistical Association*, 101(476), 1418–1429.