UNIVERSIDADE FEDERAL DO PARANÁ

VITOR PESTANA OSTRENSKY

INTERNET DATA IN ECONOMIC FORECASTING

CURITIBA

2024

VITOR PESTANA OSTRENSKY

INTERNET DATA IN ECONOMIC FORECASTING

Tese apresentada como requisito parcial para obtenção do título de Doutor através do Programa de Pós-Graduação em Desenvolvimento Econômico, Setor de Ciências Sociais Aplicadas, da Universidade Federal do Paraná.

Orientador: Prof Marcos Minoru Hasegawa
Coorientador: Prof Maurício Vaz Lobo Bittencourt

CURITIBA

2024

DADOS INTERNACIONAIS DE CATALOGAÇÃO NA PUBLICAÇÃO (CIP)
UNIVERSIDADE FEDERAL DO PARANÁ
SISTEMA DE BIBLIOTECAS – BIBLIOTECA DE CIÊNCIAS SOCIAIS APLICADAS

Ostrensky, Vitor Pestana
Internet data in economic forecasting / Vitor Pestana Ostrensky. – 2024.
1 recurso on-line: PDF.

Tese (doutorado) - Universidade Federal do Paraná, Setor de Ciências Sociais Aplicadas, Programa de Pós-Graduação em Desenvolvimento Econômico.
Orientador: Marcos Minoru Hasegawa.
Coorientador: Maurício Vaz Lobo Bittencourt.

1. Desenvolvimento Econômico. 2. Previsão econômica. 3. Dados de Internet. 4. Aprendizado do computador. I. Hasegawa, Marcos Minoru. II. Bittencourt, Maurício Vaz Lobo. III. Universidade Federal do Paraná. Setor de Ciências Sociais Aplicadas. Programa de Pós-Graduação em Desenvolvimento Econômico. IV. Título.

Bibliotecário Eduardo Silveira – CRB – 9/1921

MINISTÉRIO DA EDUCAÇÃO
SETOR DE CIÊNCIAS SOCIAIS E APLICADAS
UNIVERSIDADE FEDERAL DO PARANÁ
PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO DESENVOLVIMENTO
ECONÔMICO - 40001016024P0

TERMO DE APROVAÇÃO

Os membros da Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação DESENVOLVIMENTO ECONÔMICO da Universidade Federal do Paraná foram convocados para realizar a arguição da tese de Doutorado de **VITOR PESTANA OSTRENSKY** intitulada: **Internet Data in Economic Forecasting**, sob orientação do Prof. Dr. MARCOS MINORU HASEGAWA, que após terem inquirido o aluno e realizada a avaliação do trabalho, são de parecer pela sua APROVAÇÃO no rito de defesa.

A outorga do título de doutor está sujeita à homologação pelo colegiado, ao atendimento de todas as indicações e correções solicitadas pela banca e ao pleno atendimento das demandas regimentais do Programa de Pós-Graduação.

CURITIBA, 29 de Fevereiro de 2024.

Assinatura Eletrônica
01/03/2024 16:48:35.0
MARCOS MINORU HASEGAWA
Presidente da Banca Examinadora

Assinatura Eletrônica
07/03/2024 18:26:21.0
MARCELO CUNHA MEDEIROS
Avaliador Externo (UNIVERSITY OF ILLINOIS)

Assinatura Eletrônica
01/03/2024 14:30:27.0
EDINALDO TEBALDI
Avaliador Externo (BRYANT UNIVERSITY)

Assinatura Eletrônica
01/03/2024 14:51:05.0
VICTOR RODRIGUES DE OLIVEIRA
Avaliador Interno (UNIVERSIDADE FEDERAL DO PARANÁ)

Assinatura Eletrônica
01/03/2024 14:46:59.0
ARMANDO VAZ SAMPAIO
Avaliador Interno (UNIVERSIDADE FEDERAL DO PARANÁ)

Assinatura Eletrônica
06/03/2024 14:30:47.0
MAURICIO VAZ LOBO BITTENCOURT
Coorientador(a) (UNIVERSIDADE FEDERAL DO PARANÁ)

AGRADECIMENTOS

Gostaria de expressar minha profunda gratidão a todos que contribuíram para a realização desta tese. Primeiramente, agradeço ao meu orientador, Prof. Marcos Minoru Hasegawa, cuja dedicação foi fundamental para a conclusão deste trabalho. De igual importância foi o apoio e a coorientação do Prof. Maurício Vaz Lobo Bittencourt, que generosamente compartilhou seu conhecimento e insights valiosos ao longo deste percurso. Estendo meus agradecimentos aos avaliadores que enriqueceram significativamente este trabalho com suas críticas e sugestões. Agradeço aos Professores Marcelo Cunha Medeiros, Edinaldo Tebaldi, Victor Rodrigues de Oliveira e Armando Vaz Sampaio por suas avaliações criteriosas e contribuições indispensáveis.

Agradeço imensamente ao meu pai e à minha mãe, cujo amor e apoio incondicionais me deram a liberdade de escolher meu próprio caminho, estando sempre ao meu lado em cada decisão.

Estendo meus agradecimentos aos meus amigos, pelo companheirismo constante ao longo desta jornada.

Por fim, e mais importante, sou grato à Ketlyn, minha esposa, que desempenhou um papel fundamental durante todo o período do meu doutorado. Sem seu apoio, amor e, principalmente, paciência, este trabalho não seria possível.

He who lives by the crystal ball
soon learns to eat ground glass.
(Edgar R. Fiedler)

RESUMO

Esta tese apresenta uma investigação sobre o uso de dados da internet para a previsão de indicadores econômicos, combinando três estudos distintos. A primeira parte da pesquisa examina dados do Google Trends para prever taxas de desemprego em 32 países da OCDE, utilizando modelos de aprendizado de máquina para analisar consultas de pesquisa relacionadas ao mercado de trabalho. Este estudo mostra que os dados de motores de busca podem superar fontes econômicas tradicionais para esta tarefa, e esse desempenho está associado ao uso da internet pela população. O segundo segmento explora a relação entre fontes de notícias brasileiras do Twitter e indicadores econômicos chave. Ele utiliza modelagem de tópicos e análise de sentimentos para demonstrar a associação entre o sentimento em certos tópicos de notícias e variáveis econômicas relacionadas. A parte final da tese foca no uso de dados do Twitter para prever taxas de inflação no Brasil. Tem como objetivo desenvolver indicadores de percepção de inflação baseados em postagens de mídia social sobre mudanças de preços, comparando-os com previsões de inflação estabelecidas. Os resultados indicam que este tipo de dado contém informações úteis para as expectativas de inflação. Juntos, estes estudos contribuem para o campo de previsão econômica ao demonstrar a utilidade de fontes de dados online no fornecimento de insights sobre condições econômicas.

Palavras-chaves: Previsão Econômica. Dados de Internet. Aprendizado de Máquina.

ABSTRACT

This thesis presents an investigation into the use of internet data for forecasting economic indicators, combining three distinct studies. The first part of the research examines Google Trends data to predict unemployment rates in 32 OECD countries, utilizing machine learning models to analyze job market-related search queries. This study shows that search engine data may outperform traditional economic sources for this task and this performance is associated with internet usage in the population. The second segment explores the relationship between Brazilian news sources from Twitter and key economic indicators. It utilizes topic modeling and sentiment analysis to demonstrate the association between the sentiment in certain news topics and related economic variables. The final part of the thesis focuses on using Twitter data to forecast inflation rates in Brazil. It aims to develop indicators of inflation perception based on social media posts about price changes, comparing these with established inflation forecasts. The results indicate that this kind of data contains useful information for inflation expectations. Together, these studies contribute to the field of economic forecasting by demonstrating the usefulness of online data sources in providing insights into economic conditions.

Key-words: Economic Forecasting. Internet data. Machine Learning.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

INTRODUCTION

The last decade has witnessed a remarkable surge in data generated by internet activities. Social media platforms, search engines, and e-commerce websites collectively produce an immense volume of data daily. This internet-generated data stands in stark contrast to conventional economic indicators, characterized by its vast volume, wide variety, high velocity and often a unstructured format. Unlike traditional data, which is often lagged and aggregated, internet data may offer a more granular, real-time view of economic activities. Recent literature has demonstrated the potential of this type of data in enhance economic forecasting. However, it remains open what would be the best ways of extracting information from large volumes of data, after removing the noise that comes with them.

This thesis contributes to recent advancements by focusing specifically on employing Google Trends and Twitter data for economic forecasting. In the first essay, our investigation centered on whether integrating Google Trends data enhances unemployment rate forecasts across 32 OECD countries. We found that this data source can match, and often outperform, traditional sources, especially in countries with high internet penetration among the population. The second essay delves into the connection between sentiments derived from Twitter news accounts and various economic indicators. The topics chosen for each variable generally have theoretical links to that variable. Furthermore, the sentiment indices related to economic topics have notably enhanced GDP nowcasting exercises using economic data. The final essay explores the potential of Tweets from the general public as a reliable source for inflation forecasting. Our findings indicate that this data offers improvements over survey expectations, particularly over longer forecast horizons.

Parte I

Ensaio 1 - Google Trends Data for Unemployment Forecasting in OECD Countries

Google Trends Data for Unemployment Forecasting in OECD Countries

Resumo

This essay investigates the potential of Google Trends data as a predictive tool for unemployment rates in 32 OECD countries using data from 2004 to 2021. Employing machine learning models—Lasso, Support Vector Machine, Complete Subset Regression, and Random Forest—the study analyzes 118 different categories of job market and economic conditions-related search queries. The research broadens the scope of previous studies by examining multiple countries and leveraging the comprehensive categories feature of Google Trends to address the limitations of single-term queries. The main contributions include demonstrating the enhanced forecast accuracy using internet-based data, especially in digitally engaged societies, and introducing a scalable, country independent approach to the use of Google Trends in economic forecasting.

1 INTRODUCTION

In recent years, there have been a number of new trends in macroeconomic forecasting that have showed potential to improve the accuracy and reliability of the predictions. One of these trends is the growing availability of high-quality data sources, such as real-time economic indicators and internet-based data, which can provide valuable information for forecasting models. One of these internet-based data is Google Trends, which is a tool provided by Google that allows users to see how often specific terms are searched for on the internet. As it is the most used tool on the internet, the data coming from it can be very informative. The use of internet search engines for forecasting has grown in recent years. Since the seminal paper by Choi e Varian (2012), which was one of the firsts to rely on Google Trends data for predicting macroeconomic variables, several studies have shown that this type of data can be helpful to predict economic activity (WOLOSZKO, 2020; BANTIS; CLEMENTS; URQUHART, 2023), inflation (BLEHER; DIMPFL, 2021), and exchange rate (BULUT, 2018), for example. The primary advantages over the commonly used exogenous variables are their high frequency and low publication lag.

One macroeconomic aspect that gained special attention from this kind of study was the job market (DILMAGHANI, 2018; NAGAO; TAKEDA; TANAKA, 2019; BORUP; SCHÜTTE, 2020; MIHAELA, 2020; AARONSON et al., 2021), due to its more direct causal connection. With the spread of access to the network, the internet has become one of the critical places for job offers and searches. Job search activity reflects the proactive measures individuals take in seeking employment, characterized by searches related to vacancies, resumes, and career advice, as well as active job listings and opportunities. Conversely, unemployment rates are indicative of the broader economic health and could be reflected in increased searches for unemployment benefits and job loss, signifying an actual or feared loss of employment. Essentially, the link between Google searches and unemployment rates hinges on the idea that people's search behavior changes when they are concerned about their job status. For example, during times of rising unemployment, we expect an increase in searches related to job loss, unemployment benefits, as well as an increase in searches for new job opportunities. These changes in search patterns can be captured almost instantly through Google Trends, providing a real-time indicator of shifts in the job market.

In this domain, Google Trends has demonstrated its capability to enhance forecasts for unemployment insurance claims and employment growth in the United States, as evidenced by Aaronson et al. (2021) and Borup e Schütte (2020), respectively. These studies employ distinct methodologies utilizing Google Trends data. The former relies on data from either a single topic or an aggregation of queries, specifically "unemployment", while the latter adopts a broader approach by starting with a base set of queries and expanding to include related queries, ultimately encompassing over a hundred time series queries. Although the majority of research tends to focus on more straightforward approaches, typically employing a single search term, recent advancements highlight the effectiveness of Google Trends in forecasting unemployment rates across various countries. This is evidenced by studies conducted in Italy (NACCARATO et al., 2018), Romania (MIHAELA, 2020), Spain and Portugal (SIMIONESCU; CIFUENTES-FAURA, 2022), and Ghana (ADU; APPIAHENE; AFRIFA, 2023), which collectively attest to the ability of Google Trends to deliver forecast improvements.

In this essay, we aim to examine the potential of using Google Trends data as a predictor of unemployment rates in a set of 32 OECD countries. For this, extracted the data from 2004 to 2021 for 118 different categories of search queries that are related to the job market and economic conditions. Then, we use targeted predictors with Lasso model (TIBSHIRANI, 1996) to reduce the dimensionality of our data. Following the recent literature that showed the overall improvements of using Machine Learning models in economic forecasting (SERMPINIS et al., 2014; NG, 2014; DÖPKE; FRITSCHE; PIERDZIOCH, 2017; DIEBOLD; SHIN, 2019; MEDEIROS; VASCONCELOS et al., 2021), we estimate three different methods from this framework: Support Vector Machine, Complete Subset Regression and Random Forest.

This essay makes two main contributions to the growing body of research. First, it broadens the narrow focus of previous studies that only looked at single countries. By making forecasts for 32 OECD countries, this paper offers a more comprehensive and global perspective, uncovering trends and patterns that were previously overlooked. Moreover, the essay explores how Google trends may enhance forecasts in certain contexts. Specifically, our study includes an exercise demonstrating that the benefits of using Google trends are more pronounced in countries with higher levels of internet usage, indicating that the extent of improvement from using online data in forecasts correlates directly with the country's digital engagement.

Secondly, most previous studies have used single term queries to analyze Google Trends data. However, this approach has some limitations, as the terms used for search queries may vary depending on the language and culture of the country. Another aspect is that using a single query limit the potential information that Google Trends may provide, The use of the "categories" section of Google Trends, similar to the method from Woloszko (2020), eliminates the first issue by providing standardized categories that remain the same across all countries, regardless of language. This approach allows for better scalability of the model to multiple countries[1], while returning more information than can typically be achieved with individual terms. Importantly, by utilizing multiple categories, this method comprehensively extracts the maximum amount of relevant data, effectively addressing the second limitation and ensuring a more thorough and expansive analysis.

Beyond this introduction, this paper is organized as follows: Section 2 provides a description of the data used in the study, including sources and characteristics. In Section 3, the methods applied to analyze the data are outlined, detailing the techniques and algorithms used. The results of the study are presented and discussed in Section 4. Finally, Section 5 provides the main conclusions of the study, summarizing the findings, and addressing the limitations and future research agenda.

2 DATA

The sample has a monthly periodicity and runs from January 2004, which is the first period of data availability for Google Trends, to July 2021. We begin the evaluation using out-of-sample data starting from January 2012, ensuring a substantial period for both training and assessment. Our sample contains 32 OECD countries, which includes: Austria, Australia, Belgium, Canada, Chile, Czechia, Germany, Denmark, Estonia, Spain, Finland, France, Greece, Hungary, Ireland, Iceland, Italy, Japan, Lithuania, Latvia, Mexico, Netherlands, Norway, Poland, Portugal, Slovenia, Slovakia, South Korea, Sweden, Turkey, United Kingdom, and United States. Figure 1 shows that the unemployment rate in those countries has similar patterns over time. This is particularly evident in the 2008 crisis and in the outbreak of the Covid 19 pandemic.

[1] While individual search queries in Google Trends may vary due to language differences and local terminologies, the categories within the platform are standardized and consistently defined across all countries.

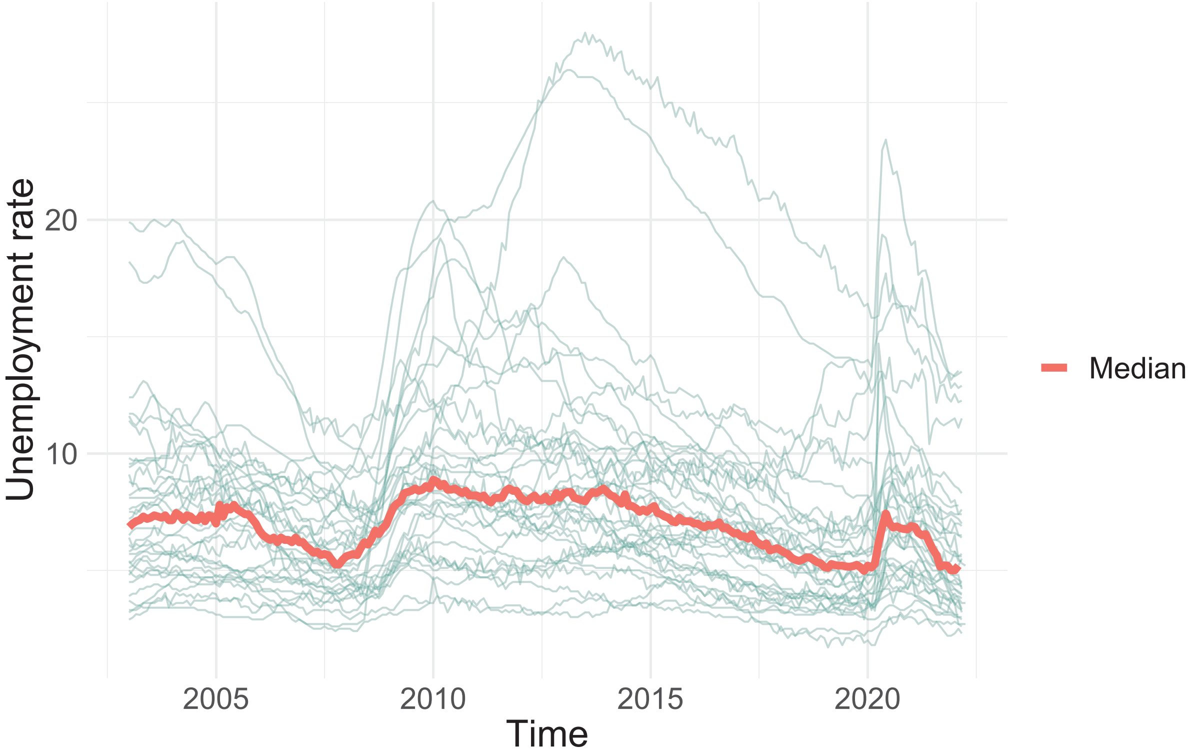


FIGURE 1 – Comparative Analysis of Unemployment Rates Among 32 Selected OECD Countries.

2.1 Google Trends

Google Trends is a public web service launched by Google in 2006. It allows users to see how often certain terms are searched for on the internet. With the internet increasingly present in people's reality, data from Google Trends provide an accurate survey of social behavior. There are a number of reasons why Google Trends data can be useful for predicting unemployment. First, the data is collected from many users, which gives a good representation of the general population. Second, the data is collected in real-time, which means it can be used to predict future trends. And finally, the data is publicly available, which means it can be used by anyone interested in using it for forecasting.

We used Google Trends to build our collection of data to predict search volume on the internet, as it provides a time series index of the proportion of queries for a search term in a given geographical location. The number of queries for a particular term is divided by the total number of Google searches in the selected geographic area and by time period. The result is then scaled in a range from 0 to 100.

Instead of relying on the use of a single or multiple queries - terms that one can search in Google - we use the category metric that is provided by Google Trends API. This metric aggregates the queries into topics, based on what the search is about. For example, a search about jobs can fall into the category "Job"or "Job Listings". There are 1133 categories in total for Google Trends. After a manual review process, we have filtered 118 of them that are related to jobs or the status of the economy itself[2]. Table 8 shows the full list in the appendix.

Medeiros e Pires (2021) highlights an issue of sample variation with Google Trends where identical queries for the same days yield inconsistent time series values. Although this issue is much less pronounced in category searches, we adopt his solution to mitigated it by averaging multiple time series obtained on the same day.

Following the procedures of Borup e Schütte (2020), we account for seasonality by estimating a linear regression for each Google Trends series, using monthly dummies as dependent variables. The seasonally adjusted data is then obtained by extracting the residuals from this model. To avoid non-stationarity and/or a deterministic trend in the series, we use the sequential testing strategy of Ayat e Burridge (2000). In this strategy, the Augmented Dickey-Fuller (ADF) test is applied sequentially for stationarity and trend. Thus, if the series has a trend, they are reversed. If they have a unit root, we take the first difference. In order to choose the lag length, we use the Bayesian information criterion (BIC) and a significance level of 1%. To avoid forward-looking bias, this procedure is applied considering the time series frame with an extended window starting after 60 months from the beginning of the data and coinciding with the first estimation window.

2.2 Benchmark - Rolling Mean Model (RM) and Economic dataset

To compare the performance of our models using Google Trends, we use the moving average of the dependent variable for each of the forecast horizons. This is the same approach Borup e Schütte (2020) used. The window for this moving average is 60 months[3], which is the same as the moving window used to estimate each model in this paper.

We also utilize macroeconomic and financial dataset from sources like the OECD, World Bank, and IMF to create a more robust benchmark. While not all variables are available for every country each month, the missing data is minimal (<5%)[4]. Our approach prioritizes the inclusion of a wide array of relevant variables to ensure extensive coverage. We also adjust

[2] In our selection process for Google Trends categories pertinent to economic and employment analysis, we focused on those with a direct correlation to job markets and economic indicators while excluding those of lesser relevance. Less relevant categories for our task, such as "Celebrities & Entertainment News", "Hair Loss", and "Computer & Video Games", were omitted due to their negligible connection to economic health or employment trends. Conversely, we prioritized relevant categories like "Job Listings,"indicative of job market activity; "Welfare & Unemployment,"reflecting economic distress; Economic Indicators like "Fuel Economy & Gas Prices", offering insights into economic conditions; and "Real Estate", as housing and real estate often mirror broader economic trends. This deliberate and focused selection of categories ensures that our analysis remains relevant, targeted, and robust in understanding and predicting economic and employment landscapes.

[3] The 60-month window was chosen after testing alternative lengths, specifically 48-month and 72-month duration. Outcomes exhibited minimal differences.

[4] To ensure the integrity of our economic dataset benchmark, we conducted a correlation analysis between the performance metrics of our final model and the proportion of missing data for each country. The results indicated no correlation, suggesting that the performance of our benchmarks is not affected by the minor gaps in data.

for publication lags, making our dataset a pseudo-real-time resource. Detailed information on the economic indicators used and their respective transformations is provided in Table 1.

TABLE 1 – Economic variables used in the economic benchmark dataset

Description	Transformation	Source
Exports (goods)	Growth rate same period previous year	World Bank
Imports (goods)	Growth rate same period previous year	World Bank
Industrial production index	Growth rate same period previous year	OECD
Share Prices	Growth rate previous month	OECD
10-year government bond yields	Difference previous month	IMF
3-month interbank rates	Difference previous month	IMF
Consumer price index - general	Growth rate same period previous year	IMF
Consumer price index - Electricity, water and fuel	Growth rate same period previous year	IMF
Production tendency (survey)	Balance s.a	OECD
Employment tendency - manufacturing (survey)	Balance s.a	OECD
Employment tendency - general (survey)	Balance s.a	OECD
Employment tendency - retail (survey)	Balance s.a	OECD
Employment tendency - services (survey)	Balance s.a	OECD
Consumer opinion on economic situation (survey)	Balance s.a	OECD
Consumer opinion on prices (survey)	Balance s.a	OECD
Consumer opinion - composite indicators (survey)	Balance s.a	OECD
Business situation - construction (survey)	Balance s.a	OECD
Business situation - services (survey)	Balance s.a	OECD
Business situation - retail (survey)	Balance s.a	OECD
Leading indicators - OECD	s.a	OECD
Passenger cars registration	Growth rate previous month	OECD
Residential buildings permits	Growth rate previous month	World Bank
Production - intermediate goods	Growth rate previous month	OECD
Production - investment goods	Growth rate previous month	OECD
Total retail trade	Growth rate previous month	OECD
Construction production	Growth rate previous month	OECD

Note: 's.a' denotes seasonally adjusted.

3 METHODOLOGY

Our approach to make predictions for different time horizons will be to estimate a different model for each of them. Thus, we will have a model to estimate the h-month ahead unemployment rate growth for each country c:

$$y^h_{c,t+h} = \frac{1}{h}\sum_{j=1}^{h} y_{c,t+h} \tag{1}$$

where $y_{c,t}$ is the log-difference of the seasonally adjusted employment rate in a country c at time t.

The forecast horizons are set at 1, 3, 6, 9, and 12 months, similar to (BORUP; SCHÜTTE, 2020), as these intervals effectively capture capabilities across both short-term and long-term timeframes. As previously discussed, we execute the model estimates within a 60-month (5-year) moving average window. This approach involves averaging the values over the specified period to smooth out short-term fluctuations and highlight longer-term trends, providing a more stable and reliable basis for our model. The choice to employ a moving average window for estimation over an expanding window involves a complex trade-off. Essentially, while including more data in the estimation generally enhances the model's accuracy and robustness — arguing for an expanding window approach — it can also introduce noise, especially for parameters that are highly volatile over time. This concern is particularly pronounced with search engine data, which can reflect abrupt changes in trends or sudden bursts of interest. Consequently, by limiting our analysis to a 60-month window, we aim to maintain a balance between capturing sufficient data to ensure model reliability and avoiding the pitfalls of overextending the time frame, which can lead to decreased sensitivity to recent developments and increased potential for distortion due to outdated or irrelevant data points.

3.1 Targeted predictors

Given our data structure, we are presented with the challenging scenario of high dimensionality: for each country, we have in excess of a hundred Google Trends variables (k), yet our estimation windows are fixed at 60-month (n), which implies a $k > n$ problem. This disproportion not only poses statistical complications, but also raises concerns regarding the potential overfitting and limited generalization of models (FRIEDMAN; HASTIE; TIBSHIRANI et al., 2001). Furthermore, it is a reasonable assumption that not all variables equally contribute to the predictability of labor market dynamics.

To mitigate this high-dimensionality issue, we adopt a "Targeting predictors"approach, which essentially involves pre-selecting a subset of the variables. As proposed by Bai e Ng (2008), and later endorsed by studies such as Borup e Schütte (2020), this technique has demonstrated improved forecasting performance in contexts characterized by sparse data, similar to our situation (BULLIGAN; MARCELLINO; VENDITTI, 2015; APRIGLIANO; ARDIZZI; MONTEFORTE, 2019; BORUP; CHRISTENSEN et al., 2022). It could be contended that the models employed for the final forecast possess the capability to process high-dimensional datasets, which is certainly accurate. However, through our tests, comparing the performance of models both with and without the incorporation of targeting predictors, we have observed significant performance enhancements when these predictors are included before the models are run.

We use Lasso (Least Absolute Shrinkage and Selection Operator), the most common regularization method (BORUP; CHRISTENSEN et al., 2022), for targeting predictors. Conceived by Tibshirani (1996), Lasso seeks to mitigate the dimensionality problem by introducing a penalty based on the absolute magnitude of coefficients, making it especially useful in instances marred by multicollinearity. The objective function can be described as:

$$\sum_{i=1}^{n}\left(y_i - \sum_{j} x_{ij}\beta_j\right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \tag{2}$$

The regularization parameter, λ, assumes a critical role in shaping the estimator's bias-variance trade-off. An increasing λ amplifies the estimator's bias, while reducing it escalates the variance. If set to 0, it reverts to the conventional OLS estimator. In our modeling process, we calibrated the λ to yield a fixed set of variables, streamlining our predictors to a manageable and more informative subset. In our preliminary tests, a selection of 15 variables emerged as an optimal balance. Utilizing too many variables risks overfitting, while selecting too few can lead to underfitting. Within the literature, the number of variables chosen for soft thresholding varies considerably, ranging from as few as 10 (BORUP; CHRISTENSEN et al., 2022), to as many as 30 (KOPOIN; MORAN; PARÉ, 2013). Our choice of 15 variables, therefore, strikes a harmonious midpoint in this bias-variance trade-off.

3.2 Support Vector Machine

Support Vector Machine (SVM) is a versatile supervised learning algorithm applicable for both classification and regression tasks. For classification, it aims to establish a hyperplane that maximizes the margin between two classes, effectively separating them. In regression, it seeks to fit the data within a certain threshold and minimize the error, focusing on predicting continuous values.

SVM employs a kernel function to transform the data into a higher-dimensional space, making it easier to find a separating hyperplane when the data isn't linearly separable in the original space. One common choice is the Linear Kernel, which is used to linearly map the data into a higher-dimensional space. This mapping involves adding new features that are combinations of the original features, specifically, the products of pairs of original features, facilitating a linear separation even in complex scenarios. The choice of a linear kernel helps in cases where the relationship between the data points is expected to be linear or when the dataset is large and computational efficiency is a priority.

In SVM estimation, a key consideration is the penalty term. This parameter deftly balances between training error and the margin—the distance between the decision boundary (or hyperplane) and the nearest data points from either class. A larger penalty term prioritizes minimizing training error over having a wider margin. Essentially, the penalty term acts as a regulator for the model's bias and variance. For our purposes, we've set this parameter to 1, adhering to the default value[5].

3.3 Complete Subset Regression

This method, a relatively recent development, was introduced by Elliott, Gargano e Timmermann (2013). It presents a particularly appealing solution in high-dimensional settings. One of its main advantages is the exploration of all potential regression combinations among predictor variables. This comprehensive approach helps in addressing issues like multicollinearity.

However, there's an inherent challenge: with an increasing number of variables, the computational burden of estimating all possible combinations becomes immense. To address this challenge, we adopted the strategy outlined by Garcia, Medeiros e Vasconcelos (2017). This strategy begins with a pre-test where each predictor variable is estimated in a simple regression against the dependent variable. Based on this initial assessment, the top 12 predictors—ranked by their t-statistic values—are selected.

With this curated set, we proceed to estimate all feasible regression combinations involving 6 predictor variables. This means evaluating 924 potential regression models, given by ($\binom{12}{6}$). The final model prediction is then derived by averaging the forecasts from these selected regressions.

The logic behind the selection of 12 predictors hinges on emphasizing those variables which, on their own, exhibit a strong linear relationship with the dependent variable. This approach ensures a delicate equilibrium between predictor relevance and computational practicality. Further, by focusing on combinations of 6 variables, we're able to capture meaningful interactions among predictors. This offers a comprehensive, yet computationally feasible, exploration of the potential model landscape.

3.4 Random Forest

Random forests are created by training multiple decision trees on different subsets of the data, and then averaging the predictions of the trees using a bagging algorithm[6] (BREIMAN, 2001). Its primary goal is to reduce the variance and the issue of over-fitting, which is extremely prevalent when using deep decision trees. This method is able to model the non-linear relationships in the data, which can be a potential path for a better performance of this model in relation to the others.

[5] We experimented with various values for this parameter, yet the results showed negligible differences.

[6] Bagging is a method where multiple models are trained on random subsets of data and their predictions are aggregated, often improving accuracy and reducing overfitting.

The random forest method also requires choosing some hyperparameters[7]. The first is the number of trees in the forest, which we set as 1000. Another important parameter is the "mtry", which sets the number of randomly selected variables in each node, which in our case is equal to 10.

3.5 Evaluation metrics

The main method to evaluate the performance of the models is the root mean squared error (RMSE), which is defined as:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{3}$$

To make the comparasion between models easier, we will often show the RMSE in terms of a ratio. In this case, if a model have smaller errors than the benchmark, the ratio of this model will be smaller than 1, otherwise, the value will be greater than 1.

The Cumulative Sum of Squared Error Differences (CSSED) is utilized to assess the relative performance and stability of a forecast model against a benchmark. This measure helps in identifying the model's accuracy over time, allowing us to pinpoint when and how significantly the model's performance deviates from the expected benchmark. The smaller the CSSED, the closer the model's forecasts are to the actual values when compared to the benchmark model. Negative values indicates that the model is underperforming. We evaluate the stability of our forecast by graphically representing the CSSED between the model of interest and the benchmark model. Specifically, for a given model l at time t, the CSSED is calculated as follows:

$$\text{CSSED}_{l,t}^{h} = \sum_{i=R}^{t}\left(\left(y_t^h - \bar{y}_t^h\right)^2 - \left(y_t^h - \hat{y}_{l,t}^h\right)^2\right) \tag{4}$$

In this equation, y_t^h represents the actual value at forecast horizon h, $\bar{y}_t^h$ is the predicted value from the benchmark model, and $\hat{y}_{l,t}^h$ is the predicted value from model l. The CSSED effectively measures the cumulative difference in predictive accuracy between the model of interest and the benchmark, considering all predictions up to time t.

4 RESULTS AND DISCUSSION

This section presents the main results of the out-of-sample forecasting findings. Firstly, the *Aggregated results* subsection consolidates and provides the main results of the models compared to a simple benckmark. This broad perspective is followed by the *Comparison with Macroeconomic/Survey dataset*, where the models are compared with a macroeconomic dataset.

[7] We used the Ranger package in R (WRIGHT; WAGER; PROBST, 2020)

The temporal robustness of these findings is then evaluated in *Performance across time*. The subsequent subsection, *Encompassing tests*, provides the statistical tests comparing the Google trends model to the model with economic dataset. In *Targeting variables*, the focus shifts to show the specific categories that were selected by the Lasso. Finally, the section delves deep into *Case studies*, providing some look into some specific countries.

4.1 Aggregated results

Table 2 shows the results for all models and time horizons considered. Our sample includes a total of 32 countries, so the results in this table are summary statistics for these countries. Columns 1-5 show the average RMSE values of the models for time horizons 1, 3, 6, 9 and 12. In the first row, the values correspond to the RMSE of the moving average model used as benchmark. The values of the other models, CSR, Random Forest and SVM, are shown as ratios to the moving average model. Thus, results smaller than 1 indicate superiority of the tested models over the benchmark model. Similarly, columns 6 - 10 show the Mean Absolute Deviation (MAD) of the models for the same time periods. The same principle applies to the metric ratio as to the RMSE. Last two columns show the percentage of cases where the models in each row had the lowest RMSE and MAD, respectively. That is, this number indicates how often the model performed better than the others across country/horizon.

TABLE 2 – Comparison of Forecasting Methods across Different Metrics

	RMSE					MAD					Frequency of Outperformance	
	h = 1	h = 3	h = 6	h = 9	h = 12	h = 1	h = 3	h = 6	h = 9	h = 12	RMSE	MAD
Rolling-mean	**4.33**	2.54	1.83	1.56	1.41	**2.07**	1.17	0.86	0.76	0.72	19.38	20.62
CSR	1.07	0.98	0.91	0.85	0.80	1.06	0.93	0.88	0.83	0.76	5.62	20
Random Forest	1.02	**0.94**	**0.85**	**0.78**	**0.73**	1.03	**0.91**	**0.81**	**0.76**	**0.70**	**71.88**	**48.75**
SVM	1.20	1.06	0.98	0.90	0.86	1.21	1.04	0.95	0.85	0.81	3.12	10.62

Note: Values in the table represent respective metric outcomes for each forecasting method. **Bold** values indicate the best performance for the given metric and period. RMSE: Root Mean Square Error. MAD: Mean Absolute Deviation.

Two results from the table are very clear. The first is that the best model tested is the Random Forest. Its performance is superior for all horizons, both RMSE and MAD, except for the one-month horizon. This can be linked to the presence of non-linear interconnections between the variables, that Random Forest is able to capture (MEDEIROS; SCHÜTTE; SOUSSI, 2022). Second, the performance of the models estimated with Google Trends is significantly better compared to the benchmark the longer the estimation horizon is. The superiority of Random Forest over the other models is also confirmed by the percentage of cases where it was the best model: 71.88% and 48.75%, based on RMSE and MAD, respectively. This difference may tell that RF may reduce the incidence of very large errors in comparasion to the other two, thus performing better in RMSE metric. These results may suggest that this forecast exercise benefits from an approach that takes into account the non-linearity of statistical relationships between variables, which is the case with the Random Forest model.

The results discussed so far are only the average of the results of the individual countries. However, it might be interesting to analyze the distribution of the models' performance across countries and time horizons. Figure 2 shows the RMSE boxplot for the estimated models. Here, the RMSE values are given in absolute numbers, in contrast to the benchmark ratio in Table 2. The yellow triangle represents the mean RMSE of each model.

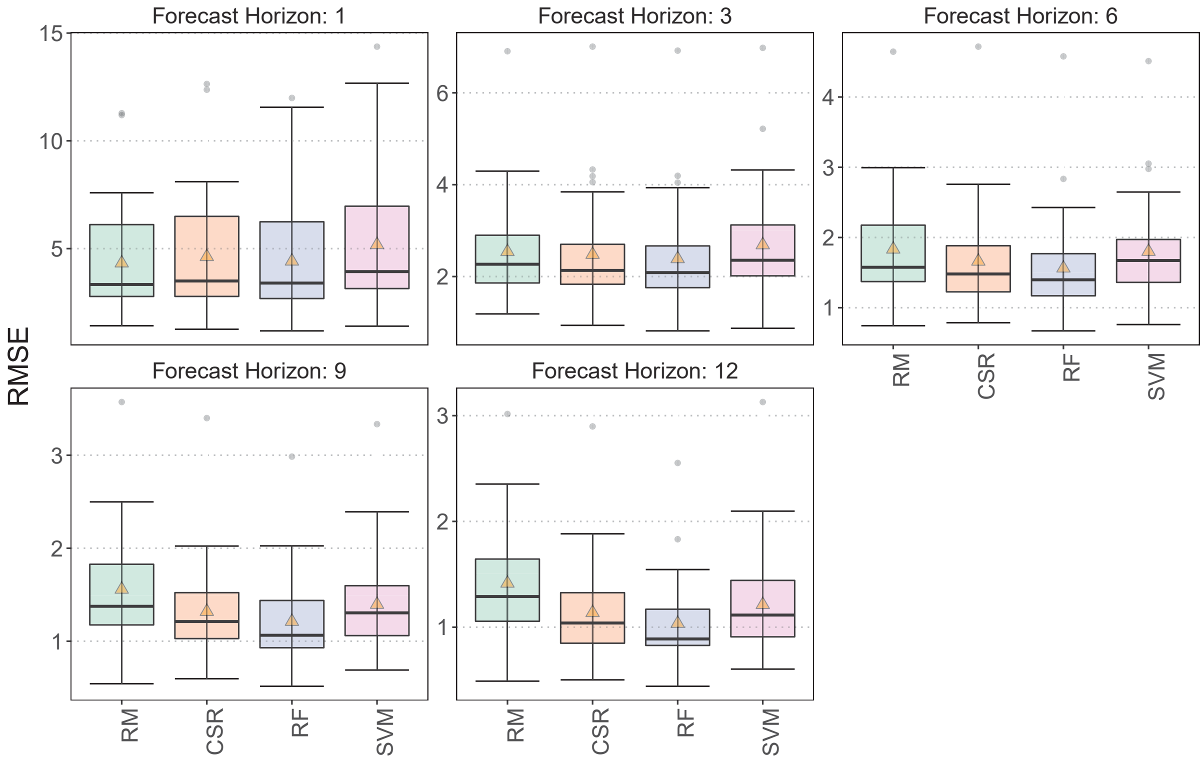


FIGURE 2 – Performance of each method by horizon

The RMSE boxplot provides some interesting clues to the predictive models. Even with a one-month time horizon, the RF and CSR models are competitive, with the latter having a very similar distribution to the Rolling-Mean (RM) model. In fact, both have a similar Q1 to the benchmark model. Moreover, the lower end of the distribution (Q1 - 1.5 * interquartile range) is smaller than in the moving average model. This means that the RMSE is slightly higher during at this horizon because a small group of countries drives the mean up. Another fact that emerges from the boxplot is that the CSR and RF models are superior from the 3-month horizon onwards, not only in terms of average but also in terms of dispersion. This relationship is more consistent the longer the estimation horizon. At the 12-month horizon, the Random Forest model has a very low dispersion compared to every other model.

The Figure 3 shows the errors distribution of the Random Forest model compared to the Rolling-Mean model by estimation horizon in detail. It can be seen that the RF model performs better at a 1-month horizon in many countries, reaching an RMSE ratio value of up to 0.8 in some cases. The curve for this time horizon is very close to a normal curve, but it loses this property the longer the forecast. The graph also shows that the frequency of over-performance of the RF model is close to 100% for the 9- and 12-month ahead estimations.

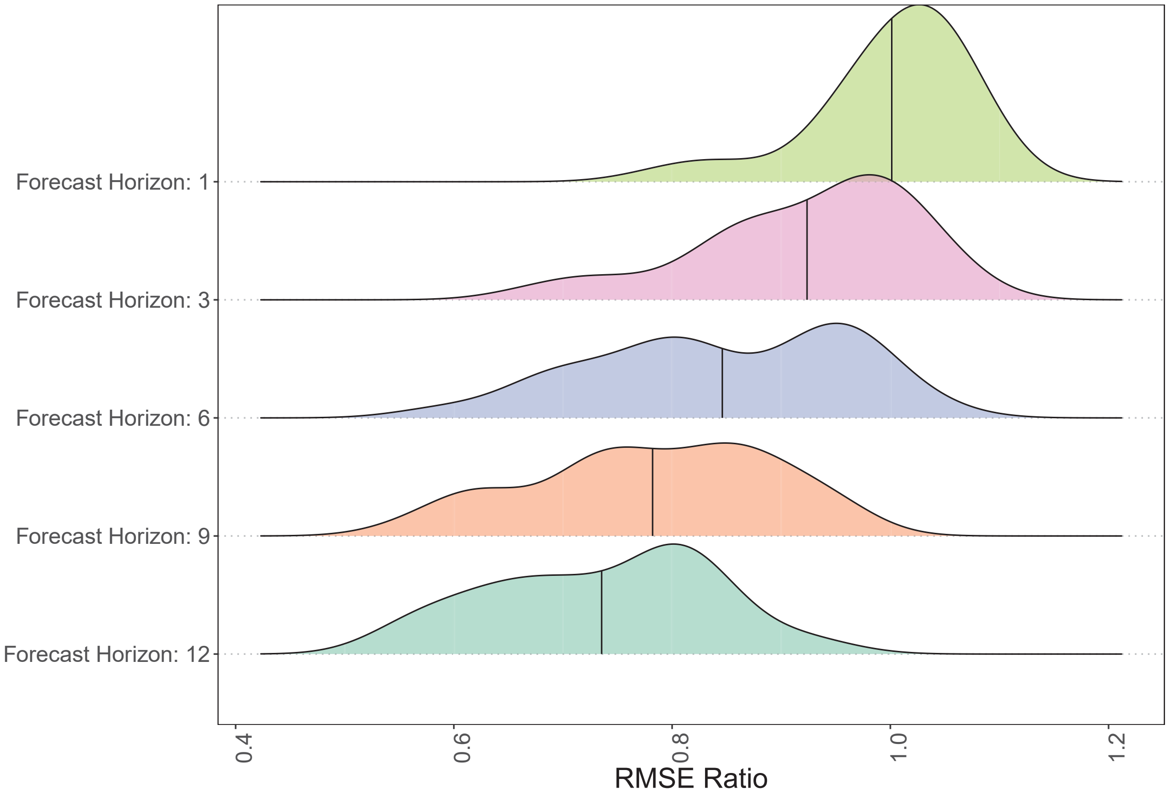


FIGURE 3 – Random Forest errors Distribution versus RM benchmark

4.2 Comparison with Macroeconomic/Survey dataset

As shown in the previous section, the performance of Google Trends data compared to a simple moving average was quite encouraging. However, it is interesting to analyze the performance of models using Google data compared to macroeconomic data, which are commonly used to predict unemployment in a country. In this way, we can see if this type of data contains additional information relevant to the prediction that is not included in the conventional variables. Similarly to Table 2, Table 3 shows the RMSE/MAD ratio in relation to the Rolling-Mean model. The Google Trends figures are exactly the same as those presented in the previous section. The difference lies in the inclusion of data from these same ratios for the macroeconomic/survey dataset. We also included a specification that combines both datasets within the targeted predictors pool to demonstrate the performance with all data.

For the Random Forest models applied to both datasets, the performance remains very similar. At shorter time horizons, the macroeconomic data edges ahead slightly, while at longer time horizons, the inverse is observed. Intuitively, one might expect macroeconomic data to reflect more 'structural' patterns and internet search data to capture 'momentary' fluctuations. Yet, given the closeness of the results, drawing such distinctions remains tentative. Something that stands out is the difference in the performance of the CSR and SVM models with Google

TABLE 3 – Comparison of Forecasting Methods: Google Trends vs Macroeconomic Datasets

	RMSE					MAD					Frequency of Outperformance	
	h = 1	h = 3	h = 6	h = 9	h = 12	h = 1	h = 3	h = 6	h = 9	h = 12	RMSE	MAD
Google Trends												
CSR	1.07	0.98	0.91	0.85	0.80	1.06	0.93	0.88	0.83	0.76	1.25	13.8
Random Forest	1.02	0.94	0.85	**0.78**	**0.73**	1.03	0.91	**0.81**	0.76	**0.70**	41.9	26.9
SVM	1.20	1.06	0.98	0.90	0.86	1.21	1.04	0.95	0.85	0.81	0.62	6.88
Macroeconomic dataset												
CSR	1.12	1.05	1.03	1	1.07	1.02	0.98	0.91	0.88	0.89	7.5	11.2
RF	**1.01**	**0.93**	**0.84**	**0.78**	0.77	**1.01**	**0.90**	**0.81**	**0.74**	0.72	**48.8**	**38.8**
SVM	1.40	1.37	1.40	1.30	1.44	1.26	1.10	1.05	0.94	0.95	0	2.5
Google Trends + Macroeconomic dataset												
CSR	1.09	0.99	0.91	0.85	0.83	1.06	0.96	0.87	0.82	0.78	-	-
RF	1.01	0.93	0.82	0.74	0.71	1.04	0.90	0.80	0.71	0.67	-	-
SVM	1.28	1.18	1.05	0.99	0.94	1.29	1.05	0.95	0.85	0.75	-	-

Note: Values in the table represent the RMSE/MAD ratio for each forecasting method using both Google Trends and Macroeconomic datasets in relation to RM benchmark. **Bold** values indicate the best performance for the given metric and period.

data. Both cases are well superior to their peers with economic data, regardless of the horizon of the dependent variable. Just as observed in the prior exercise, a disparity exists between the outperformance in RMSE and MAD metrics, which may be attributed to the Random Forest's mitigation of larger errors. As expected, incorporating both datasets together results in performance improvement compared to the best "single model", however the gains are only marginal.

Figure 4 shows the distribution of the RMSE ratio between the RF model with Google Trends data and the same RF with macroeconomic/survey dataset. Therefore, if this value is above 1, the model with economic data is performing better than the one with online search data.

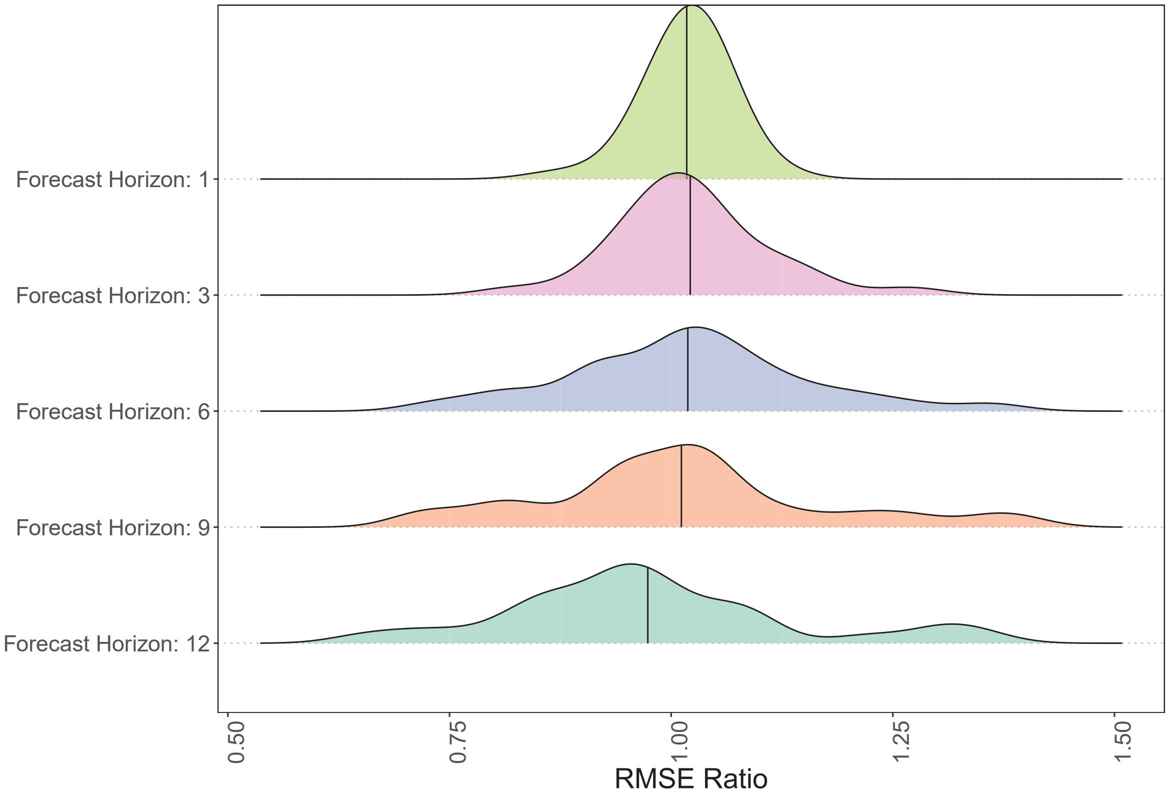


FIGURE 4 – Random Forest errors Distribution versus Economic dataset

In terms of distribution, there is a similar pattern to that found using the RM model as a benchmark. For the horizon of one month, there is an approximation of a normal curve. As the forecast horizon increases, the distribution flattens. That is, each model performs better in some countries. However, the mean remains more or less stable near 1. In general terms, the performance of the two datasets is very similar. On average, Google trends data models are at least as good as macroeconomic data models for every horizon.

4.3 Performance across time

This section presents the performance of the Random Forest model using Google Trends data and compares it with the benchmark models, namely the Regression Model (RM) and the Random Forest (RF) using macroeconomic data, across different time periods. Specifically, Figures 5 and 6 display the 30th, 50th (median), and 70th percentiles of the Cumulative Sum of Squared Errors Difference (CSSED) relative to the RM and RF with macroeconomic data, respectively. These statistics are computed at each time point to provide a detailed temporal analysis of the models' performance differences.

Compared to the Rolling-Mean benchmark, from horizons 3-12, there is a relatively constant increase, which shows some consistency in the forecasting capacity of the model. This behavior is similar for the three percentiles. For the one-month horizon, poor accuracy in some countries (30th percentile) pulls average performance down. The performance compared to the model with economic data has practically constant behavior in relation to the median for horizons 3-9. In these three cases, the highest and lowest percentiles tend to hold their performances constant over time. For the 1-month horizon, the model with online search data rarely performs better than the one with economic data, and the median of the results appears to be constantly lower, even more so in the last months of the sample. Interestingly, for the horizon of 12 months ahead, the superior performance of Google Trends data occurs only from the last months of the sample.

A fact that is clear from both graphs is that the Covid-19 epidemic significantly altered the performance of the models. The effect is quite noteworthy for the Rolling-Mean benchmark model, which is most likely because of the labor market shock caused by the Covid outbreak, which reduced the predictive capacity of a model purely based on the history of the variable. As for the comparison with the model with economic data, the effects are more diverse. There is a meaningful impact on model error, but it appears to be positive in some countries and harmful in others.

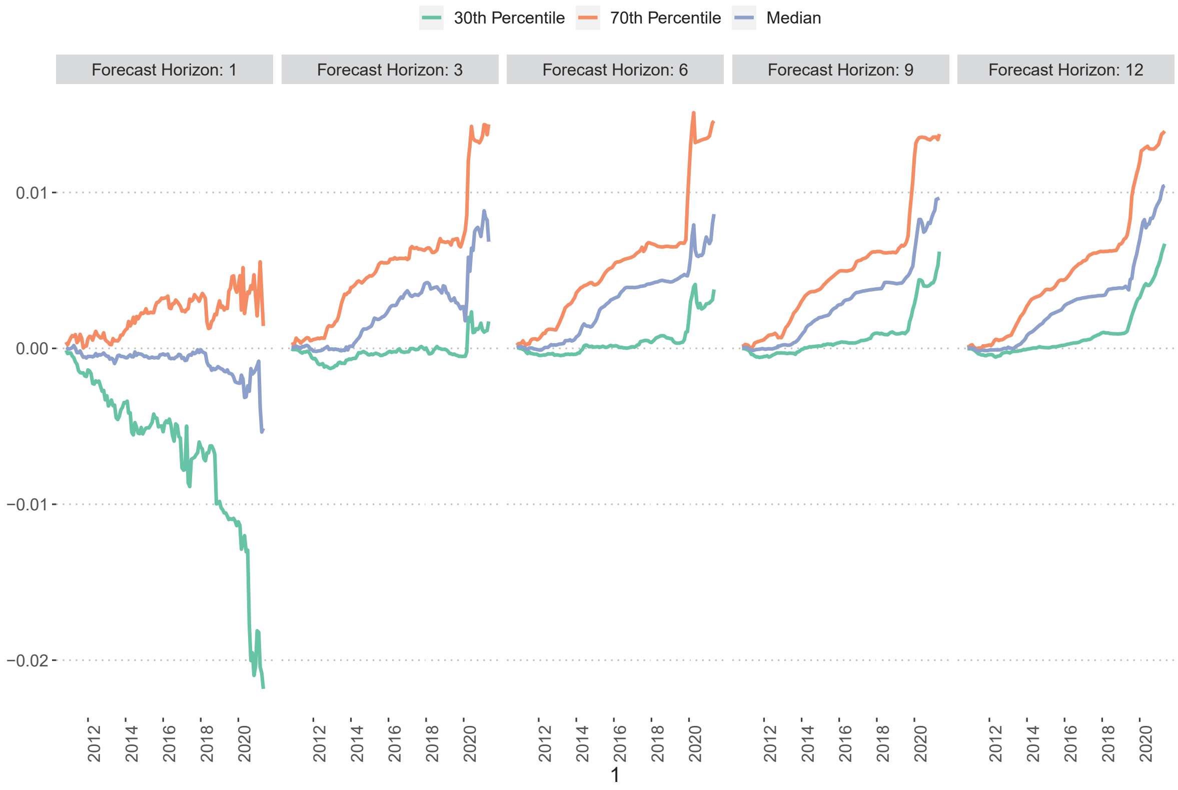


FIGURE 5 – RF Model versus RM benchmark - Cumulative Sum of Squared Errors Difference (CSSED)

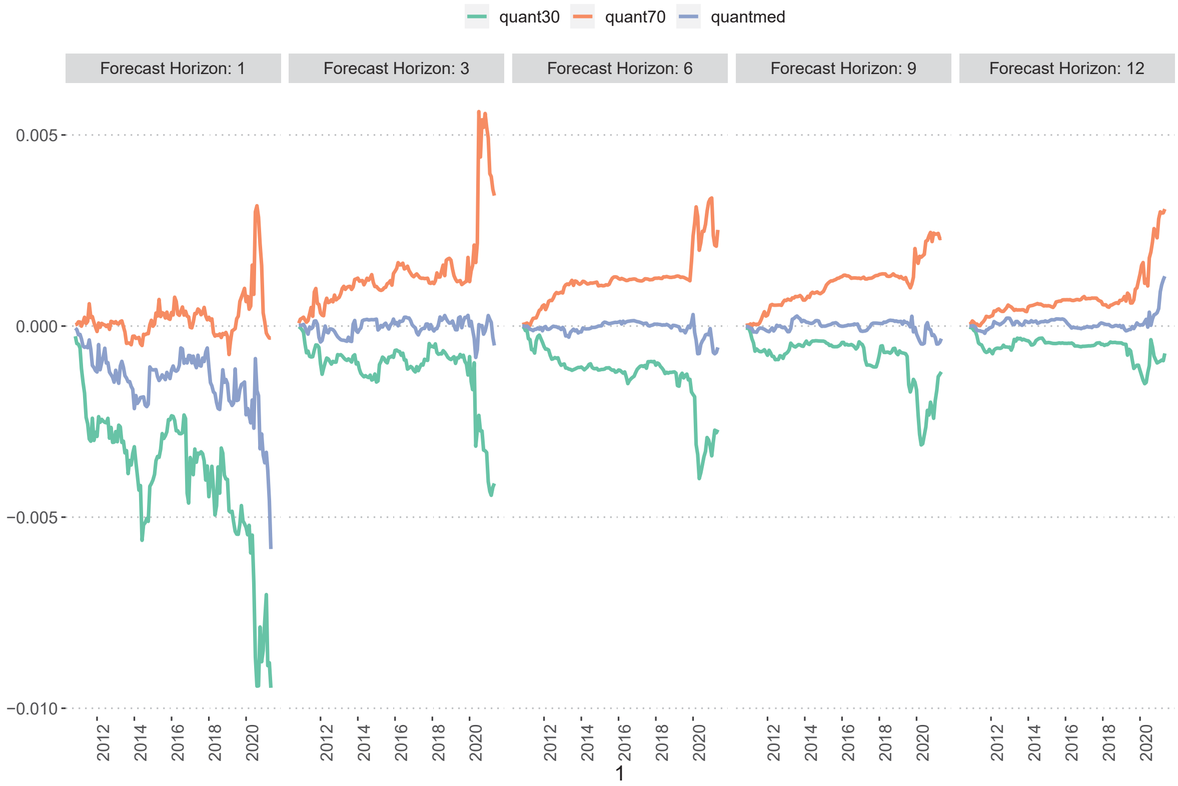


FIGURE 6 – RF Model versus Economic dataset - Cumulative Sum of Squared Errors Difference (CSSED)

4.4 Encompassing tests

The previous sections have shown that the performance of the Random Forest model with data from Google Trends and with macroeconomic data is relatively equivalent. To investigate this further, we use the Harvey, Leybourne e Newbold (1998) encompassing test. The test performs a combination of predictions with the model with Google Trends data ($\hat{y}^h_{g,t+h}$) and economic data ($\hat{y}^h_{m,t+h}$) to form unemployment (y^h_{t+h}).

$$y^h_{t+h} = \lambda^h_g \hat{y}^h_{g,t+h} + \lambda^h_m \hat{y}^h_{m,t+h} \tag{5}$$

Where $\lambda^h_g + \lambda^h_m = 1$. In the case that $\lambda^h_m = 1$, the benchmark model encompass the Google Trends model, so the latter does not contribute with any relevant forecasting information. On the other hand, a λ value above 1 means that the Google data contains information that is not present in the macroeconomic/survey data.

To test if λ^h_g is significant we use the MHLN statistic[8]. The null hypothesis is that $\lambda^h_g = 0$, so let $\hat{\varepsilon}^h_{g,t+h}$ and $\hat{\varepsilon}^h_{m,t+h}$ be the model specific forecast errors. We define $\hat{d}^h_{t+h} = \left(\hat{\varepsilon}^h_{m,t+h} - \hat{\varepsilon}^h_{g,t+h}\right)\hat{\varepsilon}^h_{m,t+h}$ to get $\mathrm{MHLN}^h = \frac{N_h+1-2h+N_h^{-1}h(h-1)}{N_h}\bar{d}^h\left(\hat{V}^h\right)^{1/2}$ where N_h is the number of out-of-sample observations, $\bar{d}^h$ is the average of $\hat{d}^h_{t+h}$, and $\hat{V}^h$ is the estimate of its variance. For this test, we used a 1% significance level.

[8] The modified version of Harvey, Leybourne e Newbold Test

Table 4 shows, for every forecast horizon, the mean values for λ_m and λ_g and the proportion of countries for each case that is defined based on the statistical significance of them. In the first scenario, λ_g is considered statistically significant while λ_m is not. In the second scenario, the situation is reversed. These scenarios suggest that either the Google Trends or the Macroeconomic dataset encompass the other. In the final two scenarios, both λ_g and λ_m are either found to be statistically significant or insignificant.

TABLE 4 – Results of the Encompassing Test

	Forecast Horizon				
	h = 1	h = 3	h = 6	h = 9	h = 12
Average λ_m	0.61	0.61	0.58	0.50	0.36
Average λ_g	0.39	0.39	0.42	0.50	0.64
λ_g is significant and λ_m is not	25.00%	34.38%	21.88%	34.38%	37.50%
λ_m is significant and λ_g is not	40.62%	31.25%	25%	18.75%	12.50%
λ_g and λ_m are significant	15.62%	9.38%	3.12%	0%	0%
λ_g and λ_m are not significant	18.75%	25%	50%	46.88%	50%

Notes: The values of λ_m and λ_g denote the average encompassing coefficients for forecasts using the Economic dataset and Google Trends, respectively. The percentages in the rows below represent the proportion of times each scenario occurs across different tests or datasets.

The first results indicate that none of the datasets are encompassed by the other, as the average weight assigned to each is close to 0.5. However, statistical significance tests show a clear and consistent pattern with previous results. The longer the estimation horizon, the greater the advantage of Google Trends data.

4.5 Targeting variables

Table 5 shows the top 15 selection frequencies for the Google Trends category using the targeting variables method. This calculation is based on the number of times each category was selected in an estimate (country window), so the categories are not necessarily always present in the same country for the entire sample.

It is clear that some categories are important in several countries, such as fuel prices, economic news, and financial markets. However, some categories were less prevalent than expected, including: "Jobs & Education", "Job Listings"and "Jobs", with 0.92%, 4.14% and 4.56%, respectively. This result is interesting as it may suggest that while the use of Google Trends is already established - in terms of usefulness for economic forecasting - the best method for selecting terms/categories is not yet clear. A definition according to purely theoretical or logical criteria may be sub-optimal in terms of information for forecasting. Given these findings, it is important to consider alternative approaches to selecting Google Trends terms and categories in order to maximize their usefulness for economic forecasting.

TABLE 5 – Presence of Various Google Trends Categories

Google trends categories	Presence (in %)
Fuel Economy & Gas Prices	53.20
Economy News	45.35
Financial Markets	37.42
News	33.3
Developer Jobs	31.79
Bankruptcy	31.04
Business News	28.47
Rail Transport	25.84
E-Commerce Services	25.64
Aviation	25.40
Credit Cards	25.35
Commercial Lending	25.26
Sports	24.86
Tobacco Products	23.82
Home Financing	23.44

Notes: The table lists the top 15 Google Trends categories by presence in the Targeting Variables. The categories are arranged in descending order based on the percentage presence.

4.6 Case studies

This section aims to show some specific cases. The choice of countries presented here was made to showcase a diverse range of countries from different regions and economic conditions. The selected countries include Germany, Slovakia and Turkey from Europe, Japan from Asia, and Mexico and USA from North America. This selection supplies a good range of country sizes, providing a more comprehensive understanding of its overall generalization and applicability.

As the Random forest model consistently outperforms its peers, we chose to only show the metrics for it in this section. The table 6 presents the Root Mean Square Error (RMSE) ratio between the Google trends model to both benchmark models, the rolling-mean (RM) and the economic dataset, for different countries and different forecast horizons (h). The ratio is calculated as the RMSE of the google trends model divided by the RMSE of the comparison model. Values under 1 are in bold, indicating that the target model outperforms the benchmark.

The table shows that the Google Trends RF model generally performs well against the Rolling-mean model, with a majority of the ratios being below 1 for most countries and forecast horizons. There is a clear pattern the longer the horizon of estimation, the better it is comparing to this benchmark.

TABLE 6 – Comparison of Forecasting Ratios of RF Google Trends model to benchmark Models in selected countries

	Ratio to RM					Ratio to economic dataset				
	h = 1	h = 3	h = 6	h = 9	h = 12	h = 1	h = 3	h = 6	h = 9	h = 12
Germany	**0.98**	**0.84**	**0.75**	**0.73**	**0.62**	1.13	1.27	1.36	1.40	1.22
Japan	1.06	1.04	1.04	**0.94**	**0.93**	1.04	1.17	1.27	1.24	1.29
Mexico	1.06	**0.97**	**0.93**	**0.87**	**0.82**	**0.95**	**0.94**	1.00	1.05	1.08
Slovakia	**0.86**	**0.67**	**0.59**	**0.56**	**0.56**	**0.88**	**0.82**	**0.74**	**0.73**	**0.75**
Turkey	1.04	**0.96**	**0.91**	**0.79**	**0.79**	1.02	1.02	**0.92**	**0.85**	**0.87**
USA	1.03	1.01	**0.99**	**0.84**	**0.85**	**0.97**	1.01	1.06	1.01	1.02

Notes: The table showcases ratios of forecasting performance of various countries relative to two reference models across multiple horizons. Ratios highlighted in bold indicate better forecasting performance relative to the respective reference model for that horizon.

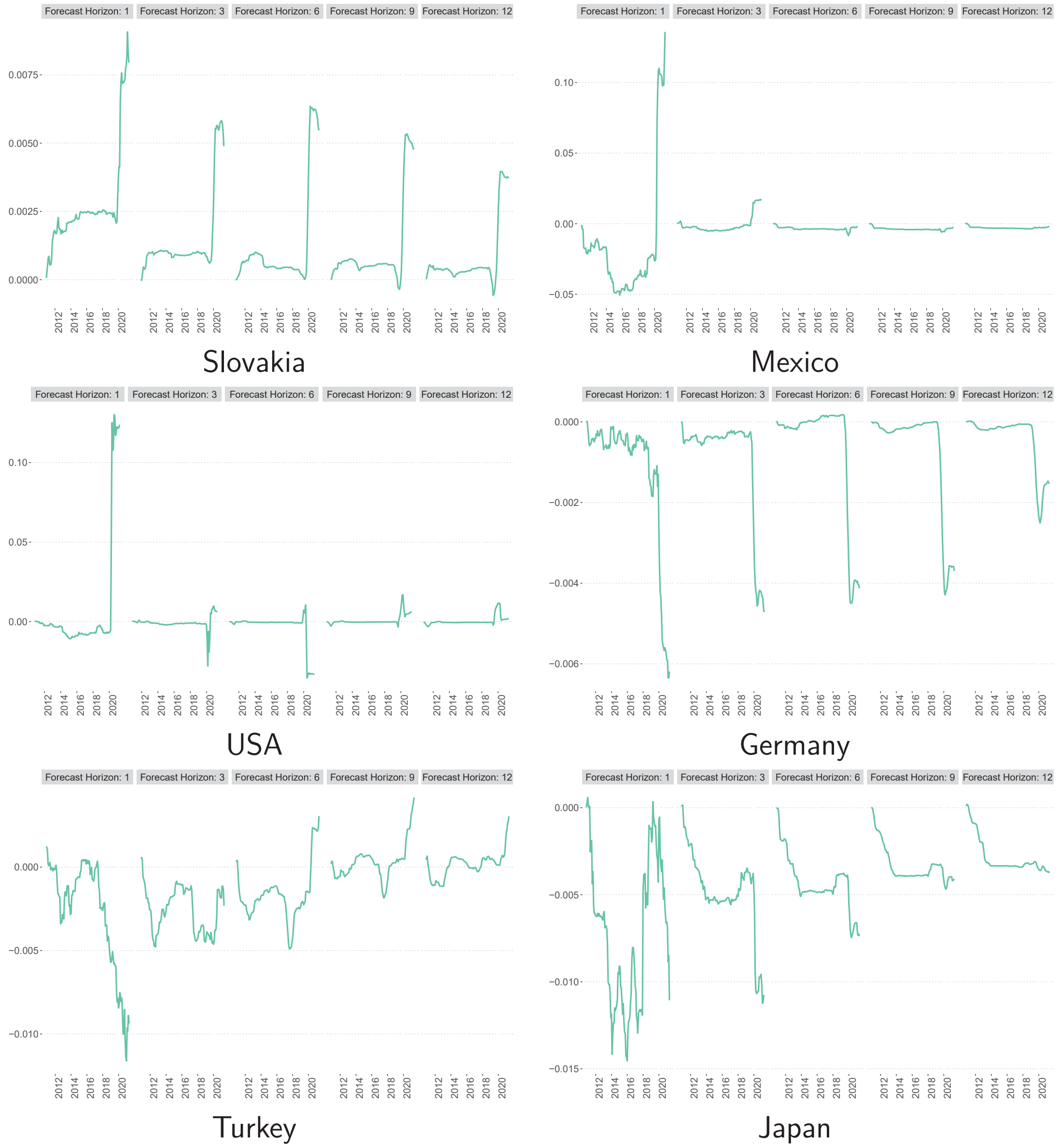
Additionally, the table also shows the comparison of the Google trends model to an economic dataset. In some cases, the Google Trends model outperforms the economic dataset, as indicated by the ratios being below 1. This is particularly notable in countries such as Mexico, Slovakia, and Turkey. However, in other cases, such as Germany and Japan, the Google Trends model does not perform as well against the economic dataset.

Figure 7 shows the relative performance (CSSED) of the Google Trends model in relation to macroeconomic data for these same countries. The graph reinforces, especially for Slovakia, Mexico, USA and Germany, the heavy influence of the pandemic on the model's performance. While in the first three there was a positive shock for virtually all estimation horizons, the model in Germany suffered a remarkable setback.

Complementing the analysis made in section 4.3, it would be expected that, due to an increase in people's use of search engines over time, the relative performance of Google Trends data compared to macroeconomic data would systematically improve. However, this did not happen. Performance varies greatly across countries and the reason for this should be investigated in future research.

4.7 The Role of Internet Usage in Forecasting gains

We demonstrate that data from Google Trends can have a relevant predictive factor for countries' unemployment rates, even outperforming macroeconomic datasets in some instances. However, as observed, this performance varies significantly between countries. One theoretical explanation for this relationship is the internet usage by the population. Countries with lower internet access might perform worse since fewer users would also reflect less use of the internet as a means to search for jobs, unemployment subsidies, or other searches through search engines.



Slovakia

Mexico

USA

Germany

Turkey

Japan

FIGURE 7 – Cumulative Sum of Squared Errors Difference (CSSED) in selected countries

This subsection aims precisely to answer this. Here, we run a fixed effects panel model with the 32 countries for the Out-of-Sample period (Jan 2012-Jul 2021). Where the explanatory variable is the performance ratio of RMSE between the Google Trends model and the macroeconomic data benchmark usign the Random Forest model. The main independent variable is the Share of individuals using Internet in the country. Here, we aggregate this ratio for annual data to align with the other data we use in the estimation. We estimate over three time forcasting horizons: 1, 3, and 12 months. In doing so, we examine the relationship in the short, medium, and long term. The panel model is estimated as follows:

$$Y_{it} = \alpha + \beta_1 InternetUsage_it + \beta_2 \mathbf{X}_{it} + \epsilon_{it} \tag{6}$$

$$\text{Where: } Y_{it} = \frac{RMSE_{GoogleTrends,it}}{RMSE_{EconomicBenchmark,it}} \tag{7}$$

$InternetUsage_i$ represents the percentage of internet usage in country i and year t. α is the intercept, and β_1, β_2 are the coefficients for each explanatory variable. ϵ_{it} is the error term for country i at time t. $\mathbf{X}_{it}$ is a vector of control variables for country i at year t that contains the GDP per capita Growth, the Inflation level and the average of the Unemployment Rate for that year. These variables are crucial for isolating the effect of internet usage on the forecasting efficacy of Google Trends data.

TABLE 7 – Determining Factors of Google Trends Outperformance - Panel Data

	(1)	(2)	(3)
		h = 1	
Internet Usage	−0.045	−0.013	−0.180
	(0.130)	(0.136)	(0.156)
Δ GDP		0.001**	0.001**
		(0.0002)	(0.0002)
Inflation Level		−0.279	−0.267
		(0.235)	(0.234)
Unemployment Rate - Average			−1.014**
			(0.470)
		h = 3	
Internet Usage	−1.087***	−1.055***	−1.541***
	(0.323)	(0.327)	(0.369)
Δ GDP		−0.0002	−0.0004
		(0.001)	(0.001)
Inflation Level		−0.802	−0.142
		(1.600)	(1.601)
Unemployment Rate - Average			−3.169***
			(1.144)
		h = 12	
Internet Usage	−0.657**	−0.634**	−1.012***
	(0.311)	(0.315)	(0.356)
Δ GDP per capita		0.001	0.0003
		(0.001)	(0.001)
Inflation Level		−1.639	−1.127
		(1.536)	(1.544)
Unemployment Rate - Average			−2.460**
			(1.103)

Note: The table presents estimated coefficients with standard errors in parentheses. Coefficients represent the effect on the performance ratio of RMSE between Google Trends and the macroeconomic benchmark. Statistical significance is denoted as $p<0.1$; $p<0.05$; $p<0.01$. Results are shown for different forecasting horizons (1, 3, and 12 months) to illustrate the short, medium, and long-term effects of Internet Usage on predictive performance. Control variables included are GDP change, Inflation Level, and Average Unemployment Rate. The model is a fixed effects panel regression, accounting for individual country characteristics over time.

Table 7 shows the results of the estimations. For the shorter forecast horizon (h =

1), the coefficient for Internet Usage is not statistically significant, suggesting that at this immediate term, the proportion of internet users doesn't significantly impact the relative performance of Google Trends. However, as we look at the intermediate (h = 3) and longer (h = 12) horizons, the negative and statistically significant coefficients for Internet Usage indicate that higher internet penetration is associated with better performance of Google Trends in predicting unemployment rates relative to macroeconomic datasets. This suggests that in countries with higher internet usage, Google Trends might be capturing more accurate and timely signals relevant to unemployment, thereby relatively outperforming the traditional data sources especially as the forecast horizon lengthens. These finding are robust to the specifications including the control variables.

In general, these results demonstrate that there is an impact of internet usage on the performance of Google Trends data. It's worth mentioning that our sample is quite homogenous, predominantly consisting of wealthy countries where internet access is widespread. At the beginning of this data sample in 2012, the average proportion of people using the internet was 73.1%, increasing to 89.8% in 2021. This analysis would likely show an even greater difference if it included a broader range of countries, particularly those with varying levels of economic development and internet access.

5 FINAL REMARKS

In this paper, we show that Google trends data can be used as a valuable tool for predicting unemployment. Our results demonstrate that Google trends contains a significant amount of predictive information for unemployment and in some cases, even outperforms models using traditional economic data. This is especially true for longer estimation horizons. However, the performance varies between countries, highlighting the importance of considering the specific context when using internet-based data. We found that among machine learning methods tested, Random Forest performs better, which could be attributed to the role of non-linear relations. Furthermore, our results indicate that more direct Google trends queries may not always be optimal, and that valuable information may be hidden in queries that do not have a direct link to the target variable. Lastly, our findings reveal that the performance of Google Trends data in forecasting is linked to the percentage of people accessing the internet, suggesting that studies must consider the context of internet accessibility as a significant factor influencing forecasting power of internet data.

It is important to note that the set of countries used in this study is limited to OECD countries, which are among the wealthiest in the world and most from Europe. Therefore, it is not clear if the results obtained in this study can be generalized to developing countries. This is a limitation of our study and future research should investigate the use of Google trends data for predicting unemployment in a broader range of countries. Another limitation of this study is that we have used only one approach for utilizing Google trends data. There are many other ways to extract information from Google trends data, such as analyzing the volume of specific queries or using sentiment analysis techniques, which were not considered in this study. Investigating the potential of other methods for utilizing Google trends data for predicting unemployment and other economic indicators is a promising avenue for future research.

A significant shock was found in the performance of the model due to the Covid-19 pandemic. In some cases this change is positive and in others negative. It is not clear whether this is caused by a labor market shock or by systematic changes, such as the popularization of remote work. Future research could address the structural impacts of the pandemic on the performance of models with internet-based data.

Google Trends provides valuable aggregated data useful for forecasting, capturing trends over time and across various spaces and content. While it presents a conservative approach, in terms of type of data, it offers reliable and easily accessible insights. There are other sources of internet data, such as unstructured data from social networks, that could be explored in future research to understand their potential benefits and challenges in economic forecasting.

TABLE 8 – List of Categories Used in the Model

Category	id	Category	id	Category	id
Ent. Industry	612	Vehicle Brands	815	Business Edu.	799
Events	569	Fuel Econ.	1268	Corp. Lending	1160
Performing Arts	23	Body Hair Removal	144	Bus. News	784
Architecture	477	Bus. and Industrial	12	Mkt. Services	83
Autos	47	Agri. and Forestry	46	Aquaculture	747
Boats	1140	Food Prod.	621	Forestry	750
Vehicle Lic.	170	Ad. and Mkt.	25	Agricultural Equip.	748
Fuel Economy	1268	Commercial Lending	1160	Company News	1179
Unwanted Hair	144	Business Services	329	Office Supplies	95
Chemicals	288	Construction	48	Industrial Mat.	287
Printing and Publishing	1176	Professional Assoc.	1199	Transport and Logistics	50
Aviation	662	Distribution	664	Freight and Trucking	289
Import and Export	354	Mail Delivery	1150	Maritime	665
Rail Transport	666	Urban Transport	667	Computers	5
Computer Hardware	30	Consumer Electronics	78	Programming	31
Developer Jobs	802	Finance	7	Banking	37
Credit and Lending	279	Auto Financing	468	College Financing	813
Credit Cards	811	Home Financing	466	Insurance	38
Health Insurance	249	Food and Drink	71	Alcoholic Beverages	277
Restaurants	276	Medical Facilities	256	Doctors' Offices	634
Hospitals	250	Mental Health	437	Pharmacy	248
Occupational Health	644	Veterinarians	380	Gifts	99
Home and Garden	11	Home Appliances	271	Home Furnishings	270
Home Improvement	158	Internet and Telecom	13	Jobs and Education	958
Education	74	Jobs	60	Job Listings	960
Government	76	Public Finance	1161	Bankruptcy	423
Business Law	1272	Labor Law	701	Legal Services	969
Emergency Services	168	Social Services	508	Welfare	706
News	16	Economics	520	Real Estate	29
Real Estate Agencies	96	Shopping	18	Apparel	68
Footwear	697	Coupons	365	Price Comparisons	352
Ent. Media	1143	Luxury Goods	696	Tobacco Products	123
Sports	20	Travel	67	Bus and Rail	708
Hotels	179	Travel Agencies	1010		

Parte II

Ensaio 2 - The Twitter-Economic Nexus: Exploring the Correlation between Social Media News and Economic Indicators

The Twitter-Economic Nexus: Exploring the Correlation between Social Media News and Economic Indicators

Resumo

This paper investigates the correlation between Twitter news sentiment and economic indicators. Utilizing the Dirichlet Multinomial Mixture model for topic modeling and a financial-specific lexicon for sentiment analysis, it analyzes tweets from diverse brazilian news sources. The study correlates extracted themes and sentiments with key economic data, employing Lasso regression for topic relevance. Findings show a significant association between tweet sentiments and economic trends, particularly in stock markets and employment. Additionally, a new GDP nowcasting approach highlights the predictive value of sentiment indices, especially during negative GDP growth periods. This research underscores Twitter News' potential as a real-time, informative source for economic analysis, providing valuable insights for forecasting through tweets with news content.

1 INTRODUCTION

As the landscape of news consumption shifts dramatically with the rise of digital media, the use of textual data from news has been increasingly recognized as a potential source of economic data. The sentiment extracted from news stories provides quantifiable data on collective economic expectations, influencing and preceding market movements (TETLOCK, 2007). Baker, Bloom e Davis (2016) demonstrates how news-derived sentiment indicators can act as leading indicators for business cycles, capturing shifts in investor and consumer confidence that precede economic expansions and contractions. These shifts in sentiment are crucial for understanding business cycles as they reflect underlying economic sentiments before manifesting in actual economic activity (BARSKY; SIMS, 2012).

Several studies have shown that incorporating data extracted from news sources improves the accuracy of macroeconomic forecasts (ARDIA; BLUTEAU; BOUDT, 2019; BYBEE et al., 2020; TILLY; EBNER; LIVAN, 2021; ELLINGSEN; LARSEN; THORSRUD, 2022; BARBAGLIA; CONSOLI; MANZAN, 2023). This can be very useful as news is inherently real-time in nature, allowing for immediate access and utilization of information, a characteristic that is particularly valuable in the landscape of economic forecasting. However, much of the recent literature has focused on traditional news sources such as newspapers, despite being in digital form.

The emergence of social media has significantly transformed the method of both disseminating and consuming news, in addition to altering the speed at which information is transmitted (MATSA; SHEARER, 2018). Twitter[9], among other platforms, has become a primary channel for news propagation, utilizing its real-time updates and extensive reach to influence public opinion and decision-making processes worldwide. The vast amount of data produced on Twitter presents an opportunity for forecasting social and economic indicators, ranging from market patterns (BOLLEN; MAO; PEPE, 2011; NISAR; YEUNG, 2018), macroeconomic variables (ANTENUCCI et al., 2014; APRIGLIANO; EMILIOZZI et al., 2023; BOARETTO et al., 2023), and elections (TUMASJAN et al., 2011). The immediacy and brevity of platforms like Twitter make them particularly influential in shaping public sentiment and opinion. This shift is significant because the velocity and volume of information flow on social media can amplify sentiment signals, making them more volatile, but also potentially more reflective of real-time changes in public mood and expectations (STIEGLITZ; DANG-XUAN, 2013).

This paper aims to analyze the relationship between the general sentiment embedded in Tweets from news sources and economic and financial data in Brazil. Our methodology employs the Dirichlet Multinomial Mixture (DMM), a topic modeling technique particularly well-suited for the analysis of concise texts. For short texts, this model outperforms the more commonly used Latent Dirichlet Allocation (LDA) for clustering Qiang et al. (2020). Following this, using a economic-specific dictionary from Correa et al. (2017), sentiment analysis is applied to gauge the overarching sentiment within each tweet from our sample, subsequently aggregating these data into a cohesive time series. Utilizing a Lasso model, we automatically identify and extract topics that correspond with key economic variables, including Industrial Production, Unemployment Rate, Market Returns and Volatily, and Uncertainty Indexes. We further assert that, generally, the topics identified are conceptually related to each corresponding variable. In addition, we conduct a GDP nowcasting exercise, demonstrating that the sentiment indexes extracted from economic topics substantially enhance forecasting accuracy.

This study significantly extends existing economic forecasting literature in three key aspects. First, it builds upon Bybee et al. (2020) methodology by showing that topics derived from news source tweets can track related economic and financial variables, thereby enhancing out-of-sample forecasting power. Our second contribution to the use of news in forecasting literature is made by specifically incorporating tweets as a source. Twitter's real-time, easily accessible data offers a reliable and flexible source for this kind of study. Lastly, our approach of prioritizing topic sentiment over topic attention reveals to be more efficient in terms of selecting topics that are theoretical related to each economic variable in our sample.

The remainder of the paper is organized as follows: Section 2 outlines the data and methodology used in our study, including the data collection process, topic modeling using the DMM model, and Lasso regression analysis. Section 3 shows some exploration on the topics generated by our model. Section 4 presents the results of our analysis, demonstrating the

[9] In July 2023 the company changed its name to $\mathbb{X}$

relationship between the extracted topics and various economic variables, in addition to a GDP nowcasting exercise. Finally, Section 5 concludes the paper and highlights potential avenues for future research.

2 DATA AND METHODOLOGY

2.1 Data Collection

In order to investigate the correlation between economic variables and Twitter data, we first collected tweets from multiple user accounts belonging to various news sources, including newspapers, TV news stations, and magazines. These sources were chosen as they are known for providing up-to-date and reliable information on economic events, which could potentially influence the analyzed economic variables. The time frame for this data collection spans from January 2011 to December 2022. Prior to 2011, the availability of Twitter data is limited, and many news organizations had not yet established a presence on social media. The dataset comprises a total of 3,419,496 tweets.

Table 9 presents the selected user accounts, a brief description, and their respective number of followers at the end of the data collection period. This table aims to provide an overview of the sources used in our analysis and demonstrates the widespread reach of the included Twitter accounts.

We recognize that some inherent biases might arise from our data collection process, such as the selection of news sources or the chosen time frame. However, we believe that our diverse selection of news sources and the extensive time period should provide a representative sample of tweets discussing economic topics. We chose to select mainly general-topic sources, avoiding niche sources.

2.2 Topic modeling

To effectively assess whether economic variables are associated with sentiments expressed in news categories that are theoretically related with them, it is essential to define and distinguish these topics clearly. This requirement is addressed through the framework of Topic Modeling, a approach that systematically categorizes a large corpus of text into distinct topics. To identify the main topics discussed in the collected tweets, we employed the Dirichlet Multinomial Mixture (DMM) model, which is particularly suitable for short texts like tweets (YIN; WANG, 2014). The DMM model is a specialized type of generative probabilistic model, predicated on the assumption that each document (in this case, a tweet) is predominantly influenced by a single topic. This assumption aligns well with the characteristics of tweets, which are typically succinct and often concentrate on a solitary subject.

In the DMM model, each topic is represented as a probability distribution over a fixed vocabulary, and each document is modeled as a mixture of these topics (BLEI; NG; JORDAN,

TABLE 9 – Selected news Twitter accounts

User	News Source	Followers
@BandJornalismo	Band Jornalismo	0.8
@bbcbrasil	BBC News Brasil	3.4
@CBNoficial	Rádio CBN	0.4
@correio	Correio Braziliense	0.9
@Estadao	Estadão	7.5
@exame	Revista Exame	2.9
@folha	Folha de S.Paulo	8.8
@g1	G1.com	14.9
@GloboNews	GloboNews	5.7
@infomoney	InfoMoney	0.5
@jornalextra	Jornal Extra	1.1
@JornalOGlobo	Jornal O Globo	7.3
@portalR7	Portal R7.com	5.1
@radiobandnewsfm	Rádio BandNews FM	1.5
@RevistaISTOE	Revista ISTOÉ	1.9
@sbtjornalismo	SBT Jornalismo	0.5
@Terra	Portal Terra	3.0
@UOLNoticias	Portal UOL	5.7
@valoreconomico	Valor Econômico	2.6
@VEJA	Revista Veja	9.1

Note: The table lists selected Twitter accounts of news outlets relevant to Brazilian news coverage. Follower counts are denoted in millions and represent the number as of July 2023. These counts were sourced from Twitter's public metrics and are subject to change over time. The selection of news sources was based on their influence and coverage in the Brazilian media landscape.

2003). The DMM model consists of two main components: the document-topic distribution and the topic-word distribution. The document-topic distribution describes the probability of a document belonging to a specific topic, while the topic-word distribution represents the probability of a word appearing in a specific topic.

The DMM model employs a Dirichlet prior, which is a conjugate prior for the multinomial distribution (MINKA, 2000). This means that when the likelihood of the data is multinomial, the posterior distribution is also Dirichlet. The Dirichlet prior allows the model to capture uncertainty in the topic proportions and facilitates the estimation of model parameters using Bayesian inference.

The process of fitting the DMM model involves estimating the model parameters that maximize the likelihood of the observed data. This is achieved using the Expectation-Maximization (EM) algorithm, which iteratively refines the estimates of the model parameters until convergence is reached (DEMPSTER; LAIRD; RUBIN, 1977).

We used the coherence score (MIMNO et al., 2011) to determine the optimal number of topics for our analysis, which, in this case, was found to be 100. The coherence score was calculated using the 50 most relevant n-grams for each topic. This approach allowed us to effectively identify the main themes within the tweets and to avoid more topics than what is necessary, which is important to reduce the dimensionality of the text data for subsequent analysis.

2.3 Sentiment analysis

After obtaining the optimal number of topics, we assigned each tweet to its most probable topic. Then, it is crucial to acknowledge the multiplicity of methods available for text sentiment analysis. From simple rule-based systems utilizing predefined sentiment lexicons to machine learning models, the spectrum of tools varies widely. Despite the increasing sophistication of some approaches, simpler, rule-based systems retain their relevance due to their ease of implementation and interpretability.

In this study, we adopt a rule-based sentiment analysis method, utilizing a financial specific dictionary based on Correa et al. (2017). This lexicon is specifically tailored for financial stability, with keywords and phrases that are uniquely capable of capturing the nuances of sentiment in the financial domain. To accommodate the language of our data, the original English lexicon has been meticulously translated into Portuguese. Additionally, several terms, specifically relevant to the Portuguese financial context, have been integrated into this dictionary. Table 10 shows five examples for positive and negative terms from the dictionary.

TABLE 10 – Sample Lexicon for Financial Sentiment Analysis: Derived from Correa et al. (2017)

English Word	**Portuguese Translation**
Positive Terms	
favorable	favorável
enhance	melhorar
opportunity	oportunidade
effective	efetivo
buoyancy	flutuabilidade
Negative Terms	
illiquid	ilíquido
challenge	desafio
failure	falha
aggravating	agravante
overvalued	supervalorizado

Note: This lexicon is adapted from Correa et al. (2017) for sentiment analysis in the Portuguese financial domain. Terms are contextually relevant and may have broader or narrower meanings in general language use.

Therefore, in order to calculate the sentiment score of a tweet, we utilize a formula that evaluate the balance between positive and negative terms. In particular, the sentiment score of a tweet i is computed as:

$$\text{Sentiment Score}(i)(s_i) = \frac{\text{Positive Count}(i) - \text{Negative Count}(i)}{\text{len(Tokens)}(i)} \tag{8}$$

where Positive Count$_i$ and Negative count$_i$ represent the number of positive and negative words in tweet i, respectively, as per the financial lexicon employed. The denominator, len(tokens$_i$), stands for the total number of tokens in tweet i, therefore normalizing the sentiment score by the length of the tweet. This metric provides a normalized score, ranging from -1 (entirely negative sentiment) to 1 (entirely positive sentiment), enabling the quantification of the overall sentiment expressed in each tweet.

So, the mean sentiment index S_kt for a given time t is computed as follows:

$$S_{kt} = \frac{1}{N_m} \sum_{i=1}^{N_{kt}} s_i \tag{9}$$

where, S_{kt} is the mean sentiment index for topic k and time t, N_{kt} is the total number of tweets for topic k and time t, s_i is the sentiment score of the i-th tweet.

2.4 Attention metric

As a way of testing other possibilities for extracting economic information from news tweets, we will also use the attention to each topic. The attention to a particular topic reflects the prominence or relevance of that topic in the analyzed period, which may have implications for the relationship between Twitter data and economic variables. It is a method that has been utilized in some studies like Bybee et al. (2020), and it offers a different perspective from the sentiment analysis.

To quantify the attention to topic k at time t, we calculate the proportion of tweets assigned to that topic within the given time period. This can be represented mathematically as follows:

$$A_{kt} = \frac{n_{kt}}{\sum_{i=1}^{K} n_{it}} \tag{10}$$

where A_{kt} is the attention given to topic k at time t, n_{kt} is the number of tweets assigned to topic k in the time period t, and K is the total number of topics. By constructing time series for each topic's attention, we can examine how the prominence of these topics varies over time and explore their relationships with the selected economic variables.

2.5 Lasso Regression Analysis

Lasso (Least Absolute Shrinkage and Selection Operator) regression is a regularization technique that can be used for variable selection and regularization in linear regression models

(TIBSHIRANI, 1996). The objective of Lasso regression is to minimize the sum of squared errors with a constraint on the sum of the absolute values of the model parameters. This constraint causes some of the coefficients to shrink to zero, effectively excluding them from the model, and resulting in a sparse representation of the selected features.

To estimate a Lasso model with a fixed number of regressors, we can employ a procedure that involves iteratively fitting the Lasso regression with different shrinkage parameters until the desired number of non-zero coefficients is obtained. In our case, we will select 5 variables. The shrinkage parameter, also known as the regularization parameter, controls the amount of shrinkage applied to the coefficients. By adjusting this parameter, we can influence the number of variables selected by the Lasso model. To have a more refined way of identifying whether the selected variables are statistically related to the target variables, we report the p-values. This is obtained by the post-inference procedure from (TIBSHIRANI et al., 2016), using the *selectiveInference* R package.

The main goal of our Lasso regression analysis is to identify which topics are most strongly associated with each economic variable. To achieve this, we used the time series of topic frequencies obtained from the DMM model as the independent variables and the economic variables (Bovespa returns, Bovespa volatility, IPCA (inflation rate), Industrial Production, Unemployment Rate, EPU uncertainty index (BAKER; BLOOM; DAVIS, 2016), and FGV uncertainty index) as the dependent variables. In addition to the topic frequencies, we also included the lagged term of the dependent variable in the pool of potential independent variables to account for possible autocorrelation in the data.

3 RESULTS AND DISCUSSION

3.1 News Topics

In this section we present the output of the topic model analysis derived from our dataset. The main goal is to show the overall topic distribution, assuring that the topics are well built and can be used in the analysis. The name of the topics were defined based on the 20 most important words, which is built on the relative frequency of the word in the topic. We showcase a selection of topics and associated tweets derived from our dataset through topic model analysis in Table 11. The table provides a snapshot of the topic model, illustrating the coherence and thematic relevance across the topics identified. For an extensive list of topics and corresponding tweets, please refer to the appendix, which offers a broader view of our model's output.

TABLE 11 – Sample of Tweets from selected topics

User	**Tweet**	**Topic**
valoreconomico	Banco mundial prevê crescimento da China neste ano	GDP
correio	Ibovespa volta a fechar acima de mil pontos, dolar recua, Vale e Petrobras preço influenciaram resultado	Brazil Financial Market
radiobandnewsfm	Papa Francisco volta a rezar na praça São Pedro no Vaticano	Religion
g1	Milhares de estudantes universitários nos EUA passam fome, não sabem onde dormir, revela pesquisa	Education
GloboNews	Rio de Janeiro, escolas particulares poderão reabrir a partir de agosto	COVID Lockdowns

Note: This table is a sample that illustrates the coherence of identified topics.

First, we used the TF-IDF (Term Frequency–Inverse Document Frequency) method to transform the tweets into a matrix of TF-IDF features. For each topic, we then computed an average topic vector by combining and averaging the TF-IDF vectors associated with the tweets pertaining to that topic. We subsequently derived a similarity matrix based on the cosine similarity between these average topic vectors.

To make possible to visualize this high-dimensional data, we applied t-distributed Stochastic Neighbor Embedding (t-SNE) on the averaged topic vectors. Then, we used K-means to cluster the topics into five distinct groups. These clusters were visualized using a scatter plot, where the points, representing topic vectors, were colored according to their associated cluster.

The clusters were given names that are meaningful about the topics they contained: "News Highlights", "Justice", "Politics & Economy", "Lifestyle & Culture", and "Incidents & Tragedies". Figure 8 provides a representation of the topic distribution and their relationships within the data, as it shows the distances between the topics.

The process of clustering is essentially an illustrative exercise, demonstrating the significant correlation between closely situated topics. For instance, the topics "International Financial Markets"and "Brazil Financial Market"nearly coincide in their scatter plot positions, indicating a strong similarity. A closer inspection of the "Politics & Economy"cluster reveals an interesting dichotomy: subjects leaning towards politics are predominantly on the right, whereas those skewed towards the economy are mainly on the left. It is noteworthy, however, that certain topics such as "Country Risk,""Legislation,"and "Trade Balance"straddle the boundaries of their respective clusters, indicating that they could easily belong to adjacent clusters. Following the clustering of topics, we examined their temporal distribution. Figure 9 illustrates the Ratio

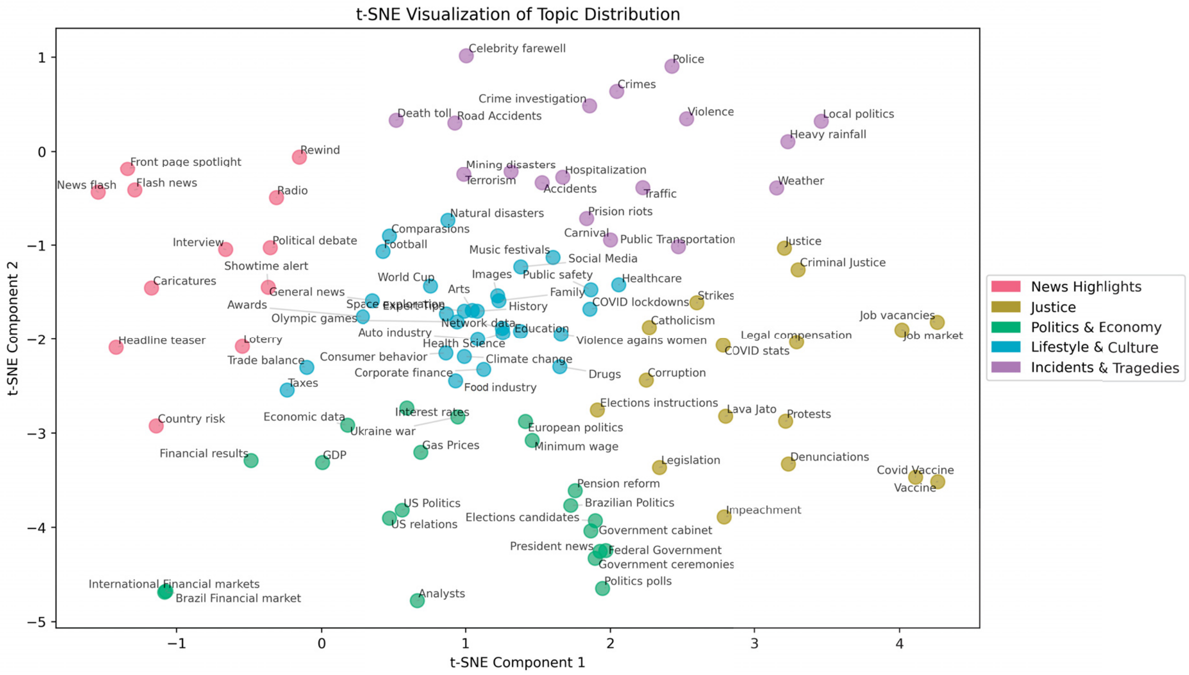


FIGURE 8 – Cluster visualization of the news topics
Note: Topics are named after the representative words defined by the model.

of tweets in each cluster over time.

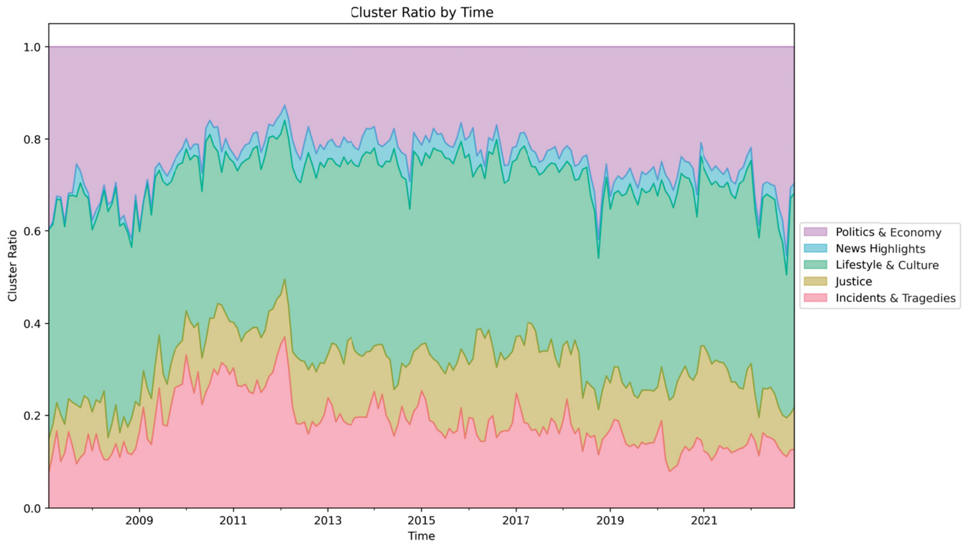


FIGURE 9 – Clusters attention share by month

The temporal analysis of the clusters reveals notable dynamics. While the majority of clusters maintain a consistent presence over the observed period, the “Politics & Economy"cluster exhibits a discernible upward trend, signifying its growing prominence. Conversely, the “Incidents & Tragedies"cluster demonstrates a downward trajectory, indicating a decrease in its presence. Despite these shifts, the “Lifestyle & Culture"cluster stands out as the most prominent throughout the time.

3.2 Visual Assessment of Sentiment Index for GDP

The potential value of sentiment indices in economic analysis can be visually assessed through the example of the sentiment index derived from the topic ‘GDP’. GDP is a comprehensive measure of a country’s economic activity and is widely recognized as a primary indicator of economic health. By focusing on GDP-related sentiment, we can directly correlate public sentiment with the overall economic performance and cycles. Figure 10 depicts the sentiment index alongside marked recession periods in Brazil. A striking pattern emerges from the graph: periods of recession correspond to pronounced dips in the sentiment index related to GDP. This visual correlation highlights the alignment between negative sentiment and economic downturns in Brazil. The low sentiment readings during recessionary times offer an indication that the topics and the sentiment analysis extracted from them may carry data information of the underlying economic conditions.

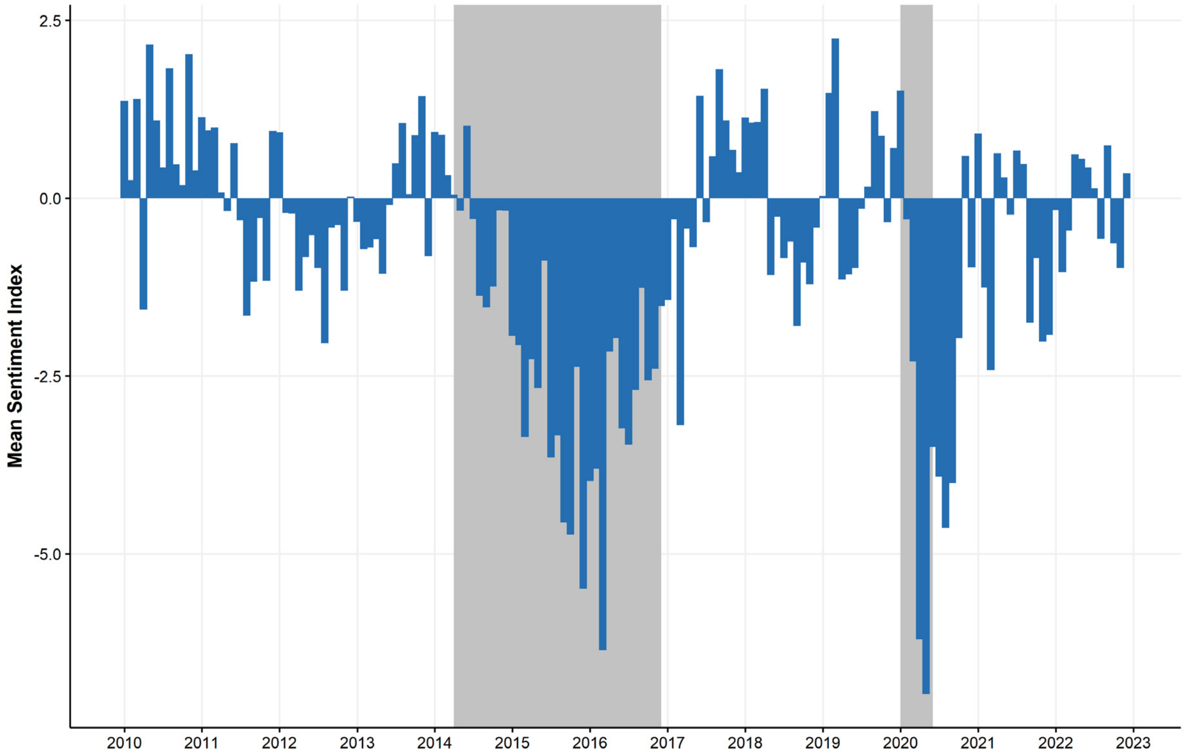


FIGURE 10 – Mean Sentiment index for Brazilian GDP.
Note: Recessions are marked as grey shaded area

3.3 Sentiment

Table 12 shows the results for two important macroeconomic variables, Employment and Industrial Production. In regards to employment, the selected topics - ’GDP’, ’Job vacancies’, ’US Politics’, ’Political debate’, and ’Corruption’ - are reflective of core facets of an economy. It is insightful to observe that both ’GDP’ and ’Job vacancies’ sentiment have a direct correlation with employment. This is in line with economic theory, where an increase in GDP would typically

TABLE 12 – Estimation Results for Macroecconomic Target Variables

Target	Topic	Coefficient	p-value	RMSE/AR(1)*
Employment growth	GDP	7.97	0.01	In-sample: 0.96
	Job vacancies	1.34	0.09	
	US Politics	-3.01	0.07	
	Political debate	-0.53	0.06	Out-of-sample: 1.69
	Corruption	0.12	0.26	
Industrial Production growth	Consumer Behavior	23.05	0.00	In-sample: 0.89
	General news	-4.53	0.01	
	Healthcare	4.52	0.26	
	Trade balance	3.70	0.33	Out-of-sample: 0.94
	Social Media	3.31	0.27	

Note: Metrics for forecasting evaluation are presented as ratios to a first-order autoregressive model (AR(1) benchmark model). In this context, the AR(1) model serves as a simple predictive model where future values are essentially a function of their immediate past values with some constant drift. Ratios below 1 indicate that the model in question is outperforming the AR(1) benchmark, signifying a more accurate or efficient predictive capability. Conversely, ratios above 1 suggest underperformance. These comparative metrics provide a standard to judge the relative efficacy of different forecasting approaches or variables within the context of this study.

correlate with a rise in employment levels, while job vacancies might be seen as an indicator of demand in the labor market. The negative coefficients associated with US Politics and Political Debate indicates a more complex interplay. This potentially indicates the influence of political decisions, debates, and conversations surrounding labor market policies and conditions.

For Industrial Production, the topics that emerged as significant were ‘Consumer Behavior’ and ‘General news’. Consumer Behavior has a strong positive coefficient, highly significant at the 1% level, implying that shifts in consumer sentiment and behavior have a direct and pronounced impact on industrial production. This aligns with the economic understanding that consumer demand drives production. In contrast, the negative coefficient for the General News topic, significant at the 5% level, implies that positive sentiment in general news correlates with a decrease in industrial production. This inverse relationship might reflect specific economic conditions or policy decisions favoring other sectors over industrial production, or it could be a statistical artifact stemming from our limited time period. The remaining topics — Healthcare, Trade Balance, and Social Media—do not demonstrate significant relationships with industrial production in this analysis, as evidenced by their p-values greater than 0.05.

Table 13 displays the same exercise for two stock markets variables, Returns and Volatility of Ibovespa, the composite index for Brazilian Stock Market. For Returns, the feature ‘Brazil Financial Market’ naturally stands out with a strong positive correlation. This suggests that as positivity in the news about Brazil’s and International financial market increases, we see a corresponding increase in Returns. This aligns with the fundamental financial theory that market sentiment often directly influences market returns.

When considering Volatility, the most relevant feature is ‘Interest rates’ with a negative

TABLE 13 – Estimation Results for Financial Target Variables

Target	Feature	Coefficient	p-value	RMSE/AR(1)
Returns	Brazil Financial Market	19.86	0.01	In-sample: 0.91
	Public safety	-4.79	0.07	
	Public Transport	-3.28	0.11	
	International Financial Markets	2.91	0.08	Out-of-sample: 0.92
	Minimum Wage	-2.77	0.20	
Volatility	Interest rates	-18.87	0.00	In-sample: 0.83
	Gas Prices	-14.07	0.02	
	Food industry	-5.47	0.09	
	Climate change	-4.29	0.10	Out-of-sample: 1.03
	Brazilian Politics	-3.80	0.16	

Note: Metrics for forecasting evaluation are presented as ratios to a first-order autoregressive model (AR(1) benchmark model). In this context, the AR(1) model serves as a simple predictive model where future values are essentially a function of their immediate past values with some constant drift. Ratios below 1 indicate that the model in question is outperforming the AR(1) benchmark, signifying a more accurate or efficient predictive capability. Conversely, ratios above 1 suggest underperformance. These comparative metrics provide a standard to judge the relative efficacy of different forecasting approaches or variables within the context of this study.

correlation. Theoretically, this could be explained by the impact of interest rates on financial market stability; positive news on interest rates might suggest stability, reducing market volatility. The negative correlation with 'Gas Prices' could be interpreted as positive sentiment around gas prices signaling market stability and thus lower volatility. An argument that contributes to the importance that this variable has in the Brazilian financial market is the economic weight of Petrobras in the Ibovespa index, which is close to 10%. Similarly, news on the 'Food industry' might be reflective of the larger agricultural sector and its impact on the economy, where positive sentiment could indicate stability and hence lower volatility.

Table 14 shows the estimations for two different metrics of Economic Uncertainty, the FGV and EPU indexes. For the FGV index, the negative coefficients for GDP, General News, and Consumer Behavior, all significant at the 1% level, may signal that positive sentiment in these areas lead to a decrease in economic uncertainty. Specifically, a robust GDP might signal economic stability, while positive General News sentiment could reflect broader confidence in the economy. Similarly, favorable Consumer Behavior may indicate increased consumer confidence, all contributing to a reduction in perceived economic uncertainty. These relationships underscore the interconnected nature of various economic indicators and sentiment measures in shaping the perception of economic uncertainty within the Brazilian market context. The significance of European Politics underscores the influence of international political dynamics on domestic economic uncertainty.

Regarding the EPU index, another measure of economic uncertainty, these topics selected by the Lasso estimations also demonstrate negative correlations, implying that positive sentiment in these areas correlates with reduced economic uncertainty. It is notable that 'GDP'

TABLE 14 – Estimation Results for Uncertainty Target Variables

Target	Feature	Coefficient	p-value	RMSE/AR(1)
FGV index	GDP	-16.71	0.00	In-sample: 0.80
	General news	-12.97	0.00	
	Consumer Behavior	-11.86	0.00	
	European Politics	-9.80	0.00	Out-of-sample: 2.43
	Economic data	-1.67	0.58	
EPU index	GDP	-22.41	0.00	In-sample: 0.88
	Political debate	-7.80	0.01	
	Climate change	-5.85	0.06	
	Russia relations	-1.76	0.04	Out-of-sample: 1.01
	Healthcare	-1.00	0.06	

Note: Metrics for forecasting evaluation are presented as ratios to a first-order autoregressive model (AR(1) benchmark model). In this context, the AR(1) model serves as a simple predictive model where future values are essentially a function of their immediate past values with some constant drift. Ratios below 1 indicate that the model in question is outperforming the AR(1) benchmark, signifying a more accurate or efficient predictive capability. Conversely, ratios above 1 suggest underperformance. These comparative metrics provide a standard to judge the relative efficacy of different forecasting approaches or variables within the context of this study.

topic once again stands out as a key feature, underscoring its importance in signaling economic health. The negative relationship with Political Debate, significant at the 5% level, highlights the sensitivity of the EPU index to political discourse. This may reflect how heated or divisive political debates can lead to increased uncertainty regarding economic policies and regulations.

3.4 Attention

In this subsection, we extend our exercise slightly and examine the potential of an alternative metric, known as the attention metric. This alternative lens quantifies the proportion of news dedicated to a certain topic, effectively measuring the level of attention that topic is receiving from the media. Table 15 shows the selected topics for the same variables that were used before. Topics that are statistically significant at the 5% confidence level have their text in bold.

TABLE 15 – Selected Topics for Different Target Variables

Target Variable	Topics selected
Employment	Interest rates, US Politics, Pension reforms, Analysts, Local politics
Industrial Production	**Images**, **Strikes**, **Minimum wage**, **Corporate finance**, World cup
Returns	Comparasions, Vaccines, Covid Stats, **Interest rates**, Music Festivals, Interviews
Volatility	**Interest rates**, Consumer behavior, Taxes, Local politics, **Political debate**
FGV index	**Interest rates**, Pension reform, Presidential news, Corruption, Traffic
EPU index	Corruption, Social Media, Family, Impeachment, Space exploration

Note: Bolded topics indicate statistically significance at 5% level.

When comparing the relevance of the topics selected through the attention metric against those identified through sentiment analysis, the theoretical linkages are clearly stronger in the case of the latter. The topics chosen based on attention do not align as closely with our economic target variables as those chosen based on sentiment. This difference is not insignificant and shows that the use of our sentiment-based topic selection may be a better approach than using attention for this sample.

Nevertheless, it is worth noting that the topic of 'interest rates' demonstrated an interesting pattern. It consistently showed up as a noteworthy predictor across multiple economic indicators. This might be attributable to the influential role of interest rates in shaping macroeconomic conditions, which would naturally garner media focus. It is important to mention that these two ways of extracting information —attention and sentiment— are not necessarily competitors. While the attention metric quantifies the level of media focus on a particular topic, sentiment analysis evaluates the tone and feelings expressed towards that topic. They measure two different aspects and can offer complementary insights.

3.5 GDP weekly tracker

To show the usefulness of the sentiment index extracted from the news topics, we built a GDP weekly tracker. For this, we follow the approach of Aprigliano, Emiliozzi et al. (2023). Three different datasets were used to make the tracker. First, what we call a "benchmark"dataset (X_b), that contains some typical variables used in this kind of exercise, such as electricity consumption and road traffic. Table 16 shows the full list of variable for our benchmark dataset. Then, the other two estimations will be done adding sentiment (X_s) and attention (X_a) indexes from the topics that were grouped in the "Politics & Economy"cluster.

First, for each dataset, we disaggregate monthly variables into weekly frequency using Chow-Lin method (CHOW; LIN, 1971). Then, we extract the first two principal components of the whole dataset and take the 13-week moving average from these two components, PC_{MA}, to match the quarterly frequency of the GDP growth, Y_t. After that, we estimate the following regression:

$$Y_t = \alpha_t + B_t PC_{\text{MA},t} + \epsilon_t \tag{11}$$

Then, we are able to calculate the weekly tracker using the coefficients from 11. Our model is trained with a rolling window of 280 weeks, starting at January 2009. Our out-of-sample period goes from October 2015 until December 2022. This approach is designed to strike an optimal balance between having a sufficiently long training set at the outset and an adequate period for evaluation. Table 17 shows the out-of-sample performance of the weekly trackers with economic topics data in relation to benchmark dataset using the Root Mean Squared Forecasting Error (RMSFE) as a metric. Given the actual values Y_t and the forecasted values $\hat{Y}_t$ for $t = 1, \ldots, T$, the RMSFE is calculated as:

$$\text{RMSFE} = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(Y_t - \hat{Y}_t)^2} \tag{12}$$

TABLE 16 – Variables in the benchmark dataset

Variable	Full Description	Source
BTS (Manufacturing) - Confidence	Confidence indicators reflecting the sentiment and expectations of manufacturers from the Business Tendency Surveys (Manufacturing)	OECD
BTS (Retail Trade) - Employment	Expected employment trends in the retail trade sector based on the Business Tendency Surveys (Retail Trade)	OECD
Consumer Opinion Surveys - Confidence	Confidence indicators gauging consumer sentiment and outlook from Consumer Opinion Surveys	OECD
BTS (Manufacturing) - Capacity Utilization	Data on the extent of capacity utilization in manufacturing facilities from the Business Tendency Surveys (Manufacturing)	OECD
Energy Load	Total electrical energy demand data	ONS
Toll Plaza Traffic Volume	Traffic volume data at toll plazas	ANTT
Monthly Services Survey	Revenue Index from the Monthly Services Survey encompassing the whole service sectors	IBGE

Definitions and Acronyms: BTS (Business Tendency Surveys), OECD (Organisation for Economic Co-operation and Development), ONS (National Electric System Operator), ANTT (National Land Transport Agency), IBGE (Brazilian Institute of Geography and Statistics)

In Table 17, the relative RMSFE is compared across different economic scenarios. The table divides the RMSFE into three categories: full sample, negative GDP growth, and positive GDP growth. Values below 1 indicates that the model outperforms the benchmark. For the full sample, the RMSFE value for attention dataset is 1.02, and for the sentiment dataset is 0.84, which is statistically significant according to the Diebold-Mariano test (DIEBOLD; MARIANO, 2002). During periods of negative GDP growth, the value for attention and sentiment datasets are 0.904 and 0.770, respectively, with the sentiment of the news topics being significant as per the DM test. In contrast, during periods of positive GDP growth, the attention data rises to 1.17, and the sentiment indexes, which is again significant in the DM test, is accentuated at 0.951. This shows that the use of the sentiment variables from the news topics enhances the performance in nowcasting the GDP systematically. Also, these performance gains are larger in periods when the GDP is decreasing.

Figure 11 provides a comparative analysis of the Real GDP with the forecasts generated by the benchmark, attention, and sentiment models. Visually, the prediction patterns of the benchmark and sentiment models appear to align closely. The attention model, on the other hand, diverges slightly from this common trend, particularly in regular, non-crisis periods. Both the attention and sentiment models present less extreme predictions during Covid crisis period, reflecting a more suited response to the crisis.

Figure 12 showcasing the Cumulated Sum of Squared Errors Difference (CSSED) in relation to the benchmark further corroborates this observation, with the sentiment model

TABLE 17 – Nowcasting performance of news-based indicators

	Attention	Sentiment
Full Sample	1.02	**0.84****
Negative GDP	0.904*	**0.770*****
Positive GDP	1.17	**0.951**

Note: Values are relative RMSE in comparasion to economic dataset benchmark. Bolded values indicates the best model performance. Significance levels are indicated by stars: *, **, and *** representing 10%, 5%, and 1% levels respectively, based on a one-sided Diebold-Mariano test against the benchmark model.

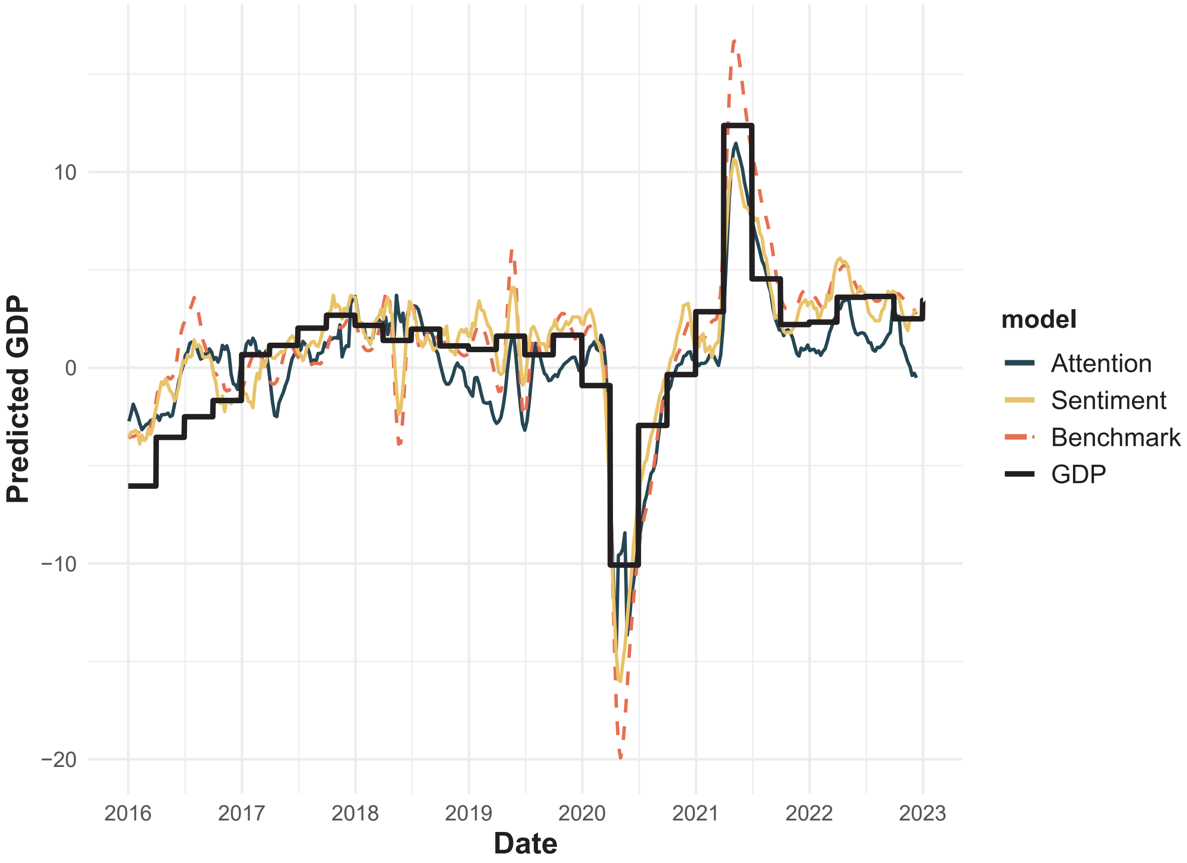


FIGURE 11 – GDP and model predictions

greatly outperforming the benchmark during the crisis period. These findings suggest that the inclusion of the sentiment data from the news topics enhances the GDP weekly tracker, especially in periods of fast shocks when usual hard data may be lagging or inadequate in fully demonstrating the status of the economy.

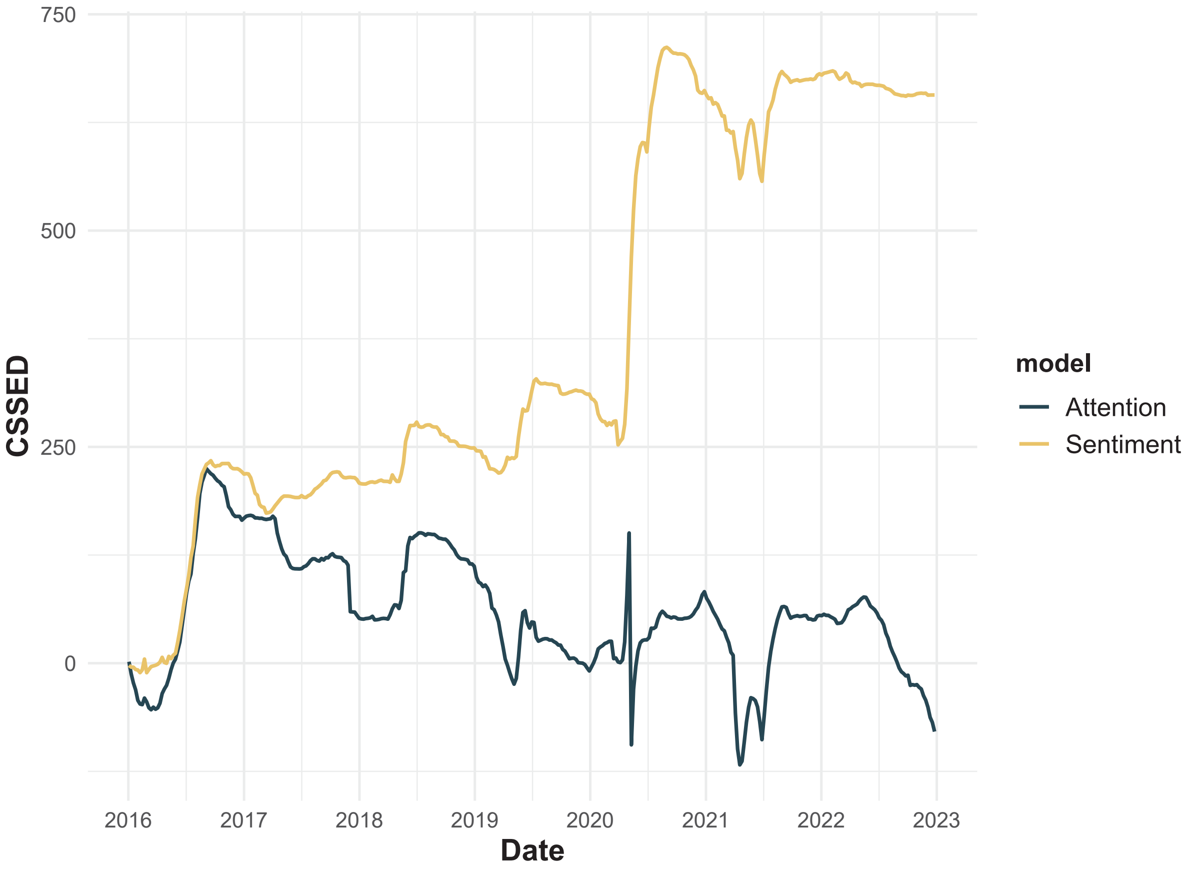


FIGURE 12 – Cumulative sum of squared errors difference (CSSED) of the models

4 FINAL REMARKS

In this study, we investigated the use of Twitter data from news sources for economic forecasting. Our methodological approach integrated sentiment analysis, Dirichlet Multinomial Mixture (DMM) modeling, and Lasso regression. The use of the DMM model proved effective in distilling coherent topics from the vast array of tweets, while Lasso regression helped in selectively pinpointing the most relevant themes for predicting various economic indicators. The findings showed that sentiment from specific topics carry relevant information for related economic variables. Also, our GDP nowcasting exercise showed that this sentiment indicators can enhance forecasting in out-of-sample settings. This is particularly valuable for the 'Forecasting with News' literature, highlighting Twitter's advantages as a news source due to its timely availability, flexibility, and ease of access.

However, our research has some limitations. The use of Twitter offers significant advantages, but it also presents certain drawbacks. While other studies in this field work with full-length texts, Twitter is typically constrained to a limited number of characters[10]. This can limit the amount of information conveyed and potentially omit subtle nuances. Finally, an important aspect is the methodological approach to text data. While our study employs a relatively simple dictionary-based technique for sentiment extraction, future research could enhance information extraction with more advanced Natural Language Processing techniques.

In conclusion, the study contributes to the growing body of research that recognizes the importance of digital media, especially social media as news source, in economic analysis. It opens up possibilities for future research to delve deeper into this intersection, exploring other platforms and innovative methodologies. The insights from this study are particularly relevant for policymakers, economists, and market analysts who are increasingly looking towards digital data sources for more timely and nuanced understanding of economic phenomena.

[10] Although the 280-character limit has been removed, the platform is predominantly known for its brief text format.

APPENDIX

TABLE 18 – Comprehensive Table of Topics Accompanied by Representative Tweet Examples

User	Tweet	Topic
bbcbrasil	americana sobrevive apos horas galho atravessado pescoco atingida passear	Accidents
CBNoficial	ouca comentario pereira	Analysts
folha	mortes eternizou beleza criatividade arte	Arts
folha	cientistas britanicos desenvolvem robo faz receita campeao assista	Auto industry
bbcbrasil	criticos apostas consideram democracia vertigem	Awards
correio	ibovespa volta fechar acima mil pontos dolar recua vale petrobras preco influenciaram resultado	Brazil Financial market
bbcbrasil	circunstancias fazem ambiguo cunha	Brazilian Politics
GloboNews	rio janeiro escolas particulares poderao reabrir partir agosto	COVID lockdowns
valoreconomico	casos crescem indicam repique	COVID stats
folha	charge alexandra publicada folha deste sabado quer ver outras charges acesse humor opiniao	Caricatures
g1	escolas fazem ensaio geral fim semana antes carnaval	Carnival
radiobandnewsfm	papa francisco volta rezar praca sao pedro vaticano dois meses local ficou fechado causa hoje acenou catolicos pediu pessoas mantenham	Catholicism
correio	nelson trajetoria marcada luta contra	Celebrity farewell
bbcbrasil	qualidade educacao freia desenvolvimento brasil revista britanica afirma	Climate change
folha	destinos sul sudeste sao preferidos brasileiros verao	Comparasions
exame	quase dobro lucro	Consumer behavior
valoreconomico	familia deixa conselho santos brasil	Corporate finance
valoreconomico	abre acao contra lei terceirizacao	Corruption
g1	morre policial nova york vitima disparo quarto meses	Country risk
exame	painel eua recomenda uso vacina	Covid Vaccine
BandJornalismo	policia civil marca julho reproducao simulada morte gravida romeu	Crime investigation
RevistaISTOE	mulher anos homicidio culposo posse ilegal arma fogo omissao cautela guarda	Crimes
GloboNews	presos chegam exames corpo delito curitiba	Criminal Justice
radiobandnewsfm	sobe numero mortos fortes chuvas cidade interior sao paulo duas pessoas ainda estao desaparecidas	Death toll

Continued on next page

TABLE 18 – continued from previous page

User	Tweet	Topic
UOLNoticias	felipe moura brasil tentar fugir sendo	Denunciations
exame	africa sul volta autorizar venda alcool tabaco	Drugs
valoreconomico	federal soma melhor agosto desde	Economic data
g1	estudantes universidade federal fazem provas	Education
exame	falou nao quer vice respeito	Elections candidates
bbcbrasil	brasileiros vao urnas hoje eleicoes precedentes	Elections instructions
GloboNews	bom ficar olho eleicoes presidenciais siria	European politics
exame	inseguro momento certo realizar	Expert Tips
folha	livro inspirou quarto nao explora vitima afirma autora	Family
exame	pedido filho bolsonaro travar investigacao desgasta governo	Federal Government
bbcbrasil	tartaruga foge encontrada meses metros casa	Financial results
folha	destaques desta	Flash news
correio	brasil area refinaria pode comprometer saude futuros moradores	Food industry
folha	comeca segundo tempo japao classificando quartas final vivo	Football
jornalextra	bom dia viram capa extra hoje acompanhem ultimas noticias domingo paz	Front page spotlight
valoreconomico	banco mundial preve crescimento china neste ano	GDP
g1	tarifas onibus intermunicipais voltam valores antigos	Gas Prices
correio	iluminacao natal esplanada foto	General news
folha	relator aposta ciro nogueira casa civil avanco voto impresso congresso	Government cabinet
JornalOGlobo	presidente bolsonaro deixou residencia embaixador londres participar exequias rainha recepcao organizada ministro exterior britanico	Government ceremonies
radiobandnewsfm	jornal venha gente fique dentro principais assuntos noticiario	Headline teaser
g1	criacao cordeiros feita uso medicamentos convencionais	Health Science
radiobandnewsfm	hospital luz vila mariana zona sul sao paulo instala camara estacionamento armazenar corpos ate liberados sepultamento desocupem leitos unidade tratamento intensivo	Healthcare
g1	chuva forte deixa familias desalojadas vassouras sul rio	Heavy rainfall
bbcbrasil	britanicos usaram humor desafiar hitler alemanha	History
correio	filho baleado pai norte transferido hospital particular	Hospitalization

Continued on next page

TABLE 18 – continued from previous page

User	Tweet	Topic
g1	veado perde adotado rebanho ovelhas inglaterra	Images
UOLNoticias	vice amazonas faz armacao contra governador tenta exonerar secretario seguranca madrugada	Impeachment
exame	video especialistas citam oportunidades renda fixa juro alto	Interest rates
valoreconomico	bancos aceleram alta tarifas	International Financial markets
radiobandnewsfm	inovando camisa branca estilo jornalista estilista apresentadora debora ensinam inovar duvidas mande	Interview
g1	empresas setor veem crescer apos capacitacao	Job market
g1	atento seleciona vagas dois estados	Job vacancies
g1	advogada explica acontece policia conclui inquerito	Justice
correio	delacoes devem divulgadas	Lava Jato
correio	mulher anos recebera pensao morte pai ate conquistar cargo decisao baseada lei	Legal compensation
valoreconomico	senado vota reforma trabalhista acompanhe	Legislation
CBNoficial	sao paulo	Local politics
Terra	fez aposta pode pagar ate milhoes sorteio hoje loteria	Loterry
exame	pagar juros beneficios atrasados	Minimum wage
g1	milhares estudantes universitarios eua passam fome nao onde dormir revela pesquisa	Mining disasters
BandJornalismo	banana aposentada sao paulo apresenta nova tecnologia recuperacao jogadores	Music festivals
radiobandnewsfm	segundo autoridades chances replicas sismo mapa abaixo mostra regiao onde terremoto registrado	Natural disasters
folha	nao curtiu novo milhoes votos rever alteracoes folha	Network data
RevistaISTOE	nova confira	News flash
correio	toquio retira parte revezamento tocha vias	Olympic games
valoreconomico	relator reforma administrativa carreiras excepcionalidade alguns beneficios tambem precisam assumir	Pension reform
BandJornalismo	casal preso diversos animais silvestres dentre onca litoral sao paulo	Police
UOLNoticias	freitas garcia participam debate amanha junto outros candidatos governo sao	Political debate
g1	caca voto indefinido leia	Politics polls
folha	gasolina vira trocadilho setembro faixas rio brasilia	President news

Continued on next page

TABLE 18 – continued from previous page

User	Tweet	Topic
g1	nao perfil alguem sorrateiro espera crianca delegada mato grosso sul	Prision riots
g1	manifestacoes bloqueiam vias raposo	Protests
exame	sabia voce pode receber noticias dia logo manha	Public Transportation
Terra	associacao pede retirada anuncio cerveja	Public safety
folha	advogado reducao maioridade penal criaria marginais	Radio
g1	volvo primeiras impressoes	Rewind
g1	onibus equipe volei tomba avenida grande	Road Accidents
valoreconomico	cidade poupada pior cenario nevasca jornal eua	Showtime alert
jornalextra	mostra fratura rosto delegado agredido atletas argentinos rio	Social Media
Terra	modulo comeca descida direcao	Space exploration
folha	sindicato funcionarios aderiram	Strikes
exame	libera ate pessoas	Taxes
GloboNews	policia procura origem arma atirador usou ataque campinas homem matou cinco pessoas catedral	Terrorism
exame	brasil milhoes outubro	Trade balance
correio	motoristas avancam dificuldade principalmente chegada viaduto santa maria	Traffic
correio	acordado membros comissao nevada surpresa	US Politics
correio	mundo negociacoes programa nuclear ira nao progridem	US relations
bbcbrasil	relatos dao conta tropas russas trancadas poroes subsolos lutar ucrania	Ukraine war
g1	segundo eua partir abril todas doses estarao disponiveis postos vacinacao	Vaccine
folha	corregedoria investiga assalto comerciante dentro delegacia salto	Violence
folha	igreja anglicana criancas devem explorar identidade genero	Violence agains women
g1	conquista titulo ano quente registrado	Weather
CBNoficial	quatro campo mostrado irregularidade desde inicio eliminatorias	World Cup

Parte III

Ensaio 3 - Forecasting inflation using twitter data

Forecasting inflation using Twitter data

Resumo

This research explores the use of Twitter for real-time economic forecasting in Brazil, focusing on inflation. We hypothesize that Twitter-based indicators, derived from tweets about price changes, correlate with actual inflation rates. The study aims to construct indicators of consumer sentiment on inflation, linking social media discourse with economic data. This approach is grounded in the growing field of using unstructured data for economic forecasting, particularly from social media. Our methodology involves selecting specific keywords related to price changes and applying topic modeling, enhanced by a dictionary-based method using bi-grams and tri-grams associated with inflation. We compare our model with established inflation survey-based forecasts and evaluate the predictive power of tweets from the general public versus news sources. Our findings indicate that this data source encompasses valuable insights pertinent to contemporary inflation trends. Furthermore, employing this data yields superior performance benefits over survey-based expectations, particularly for projections over longer periods.

1 INTRODUCTION

In an era where social media platforms wield significant influence on public opinion and perception, utilizing these platforms to assess economic conditions offers a new perspective in economic forecasting. The real-time nature of responses and the wide spectrum of viewpoints found on platforms such as Twitter constitute an underutilized source of data. Given the propensity of Twitter users to share their immediate experiences, analyzing Twitter feeds may be significantly effective for measuring current economic conditions and expectations. For example, the frequent tweets about price changes may offer a real-time window into inflationary trends. These observations, reflecting personal encounters with rising or falling costs, can provide a timely and direct measure of economic fluctuations, potentially serving as a reliable indicator for assessing the current state of inflation. As inflation is a key indicator of economic health and a focal point in policy decisions, understanding its nuances through direct public discourse offers a novel approach to capturing its impacts and trends.

In this research, we aim to explore Brazilian tweets to construct real-time indicators of consumer sentiment regarding inflation within the country. The central hypothesis posits that Twitter-based indicators, derived from tweets, may be correlated with the actual inflation rates. This study seeks to establish a link between social media discourse and economic indicators, potentially offering a real-time method of gauging consumer perceptions of inflation. The aim is to identify and utilize novel data types from these discourses that could significantly enhance the precision of inflation forecasts.

While assessing whether the Twitter-based data can provide complementary, additional information, enhancing nowcasting predictive power, we also test how well indicators derived from the general public compare to the ones that only use tweets from news sources. As a benchmark, we use the Focus Survey inflation expectations to measure a baseline indicator. The findings could provide valuable insights for economists and policy makers in understanding and responding to inflationary trends in a more timely and nuanced manner.

In recent years, the use of unstructured data for economic forecasting has emerged as a dynamically growing field within the literature. Numerous papers have indicated that including information from news sources enhances macroeconomic forecasts (ARDIA; BLUTEAU; BOUDT, 2019; BYBEE et al., 2020; TILLY; EBNER; LIVAN, 2021; ELLINGSEN; LARSEN; THORSRUD, 2022; BARBAGLIA; CONSOLI; MANZAN, 2023). Another expanding approach is the use of Central Banks communication for the same task (DRÄGER; LAMLA; PFAJFAR, 2016; LIMA; GODEIRO; MOHSIN, 2021; FERREIRA et al., 2021; HUBERT; LABONDANCE, 2021; LIN et al., 2023). The application of social network data for forecasting is particularly notable in the context of financial market forecasting (CHEN et al., 2014; SHEN; URQUHART; WANG, 2019; JIAO; VEIGA; WALTHER, 2020). In relation to social media data, specifically from Twitter, in the context of inflation forecasting, our work builds upon the findings of Angelico et al. (2022) and Boaretto et al. (2023).

In our study, we aim to select specific keywords that are closely associated with price changes and expected inflation. This selection is crucial to ensure that the tweets we analyze are directly relevant to our research objectives. Once the relevant tweets are identified, we apply a topic modeling approach to analyze and categorize them. This method is similar to the one employed in Angelico et al. (2022), allowing us to systematically filter themes and patterns related to inflation within the tweets.

Additionally, we enhance our analysis with a dictionary-based method. This involves the use of manually labeled bi-grams and tri-grams that are specifically chosen for their association with price changes. The integration of this method helps us to capture the specific language and expressions that are often used in the context of discussing inflation and price variations. This combined approach of targeted keyword selection, topic modeling, and the use of a tailored dictionary of bi-grams and tri-grams, equips us with a comprehensive framework for accurately analyzing and forecasting inflation trends based on social media data.

Our research aims to enrich the field by adding new dimensions to the existing work done in Angelico et al. (2022) e Boaretto et al. (2023). We integrate the concept of nowcasting into our approach, utilizing Twitter data to monitor inflation in real-time. This method complements the approaches taken in the two referenced papers, where the focus was predominantly on forecasting inflation expectations. By analyzing current inflation trends as they happen, our approach provides an additional layer of insight, enabling us to capture a more immediate and accurate understanding of inflationary changes, alongside the forward-looking predictions established in previous research.

In handling the succinct and unique nature of Twitter texts, we employ the Dirichlet Multinomial Mixture (DMM) topic model. This model is particularly suited for short texts, outperforming the more common Latent Dirichlet Allocation (LDA) for clustering in this setting Qiang et al. (2020). By utilizing the DMM model, we can interpret the fast-paced and condensed information on social media more accurately, thereby expecting to enhance the precision of our inflation forecasts.

Another key aspect of our research is the comparative analysis between different types of Twitter data. By examining tweets from news accounts alongside general tweets, we can evaluate the relative effectiveness of these different data sources in forecasting inflation. This comparison allows us to understand the nuances and predictive capabilities of various Twitter content types, further refining our forecasting methodology. In this study, we diverge from the methodology of Boaretto et al. (2023) by conducting a comparative analysis of two distinct subsets derived from the same platform, Twitter, contrasting our approach with their comparison of Twitter data against entire newspaper articles.

This essay is organized as follows: Section 2 describes the criteria for data extraction from Twitter, the preprocessing steps that we have done and the methodology that we built to extract relevant inflation indicators from the tweets. Section 3 we show how our Twitter-based indicators correlate with the Focus Survey forecasting for inflation. In Section 4 we conduct a forecasting exercise with our indicators. Finally, Section 5 brings the final remarks of the paper.

2 DATA AND METHODOLOGY

2.1 Data Collection

In our study, we used Twitter data, spanning from January 2010, up to December 2022, to analyze consumer sentiments on inflation in Brazil. We selected 2010 as the start date because data from earlier periods do not exhibit significant relevance on this social media platform. During this period, we extracted all the tweets that contained at least one word from a predefined list of terms related to inflation and economic trends. The full list and the English translation is shown in table 19. The selection of terms was carefully designed to encompass a wide range of relevant expressions, while ensuring the feasibility of extracting and processing all the data. The complete dataset obtained from our extraction process comprises 85.4 million tweets.

In the phase of data preprocessing for our Twitter analysis, we undertook a series of systematic steps to refine and prepare the dataset for in-depth examination. Initially, we focused on removing duplicates, ensuring that each tweet in our dataset was unique and that

TABLE 19 – List of Keywords Used for Data Extraction

Keyword in Portuguese	English Translation
Inflação	Inflation
Preços	Prices
Preço	Price
Caro	Expensive
Barato	Cheap
Contas	Bills
Custo	Cost
Oferta	Offer
Deflação	Deflation
Gasolina barata	Cheap Gasoline
Gasolina cara	Expensive Gasoline

Note: This table lists keywords used for filtering economic data from Tweets in Portuguese. The keywords were selected based on their relevance to price dynamics and were similar to the ones chosen by Angelico et al. (2022).

our analysis would not be skewed by redundant data. Following this, we employed tokenization, breaking down the tweets into individual words or tokens, which allowed for more granular analysis of the text. We then normalized the data, converting all text to lower case to maintain consistency across the dataset. Lastly, we removed stopwords, which are common words that offer little meaning for our analysis. These steps were essential in transforming the raw Twitter data into a clean, structured format.

2.2 Topic Model

Following Angelico et al. (2022), a critical step in our methodology involved the use of the Dirichlet Multinomial Mixture (DMM) topic model to effectively reduce noise within the Twitter dataset [11]. The primary objective of this process was to isolate and focus on tweets specifically relevant to economic conditions, particularly those related to price changes. This filtration was essential because, even that our dataset contains only a set of predetermined keywords, a significant proportion of tweets do not pertain to economic matters and thus could introduce substantial noise. This targeted approach ensured that our analysis was concentrated on tweets that are most likely to yield meaningful insights into public sentiment and perspectives regarding price changes, crucial for the objectives of our research.

Given the brevity and focused nature of tweets, the DMM model was selected for topic classification. Unlike models that assume multiple topics per document, like the most used LDA (Latent Dirichlet Allocation), DMM assigns each document to a single topic, making it

[11] Due to the computationally intensive nature of the DMM model, our initial analysis was conducted on a random sample of 10 million tweets from our database. This sample size was chosen to balance computational feasibility with representativeness of the overall dataset. Following the successful application of the model on this subset, we then extended the DMM model's predictions to categorize the topics of the remaining tweets in our dataset.

appropriate for short-text data like tweets (YIN; WANG, 2014). We used the coherence score to determine that the optimal number of topics was 50. The full list of topics and the top-10 words associated with them are displayed in table 25 in the Appendix.

Among the identified topics, we specifically selected clusters 2, 6, and 34 from the 50 identified topics, as these clusters demonstrated a significant economic relevance to our analysis of price movements. Cluster 2 encompassed discussions related to fuel and energy prices, a key indicator of economic shifts. Cluster 6 addressed broader conversations around inflation and market trends, directly relating to the economic condition. Cluster 34 provided insights into the socio-political discourse surrounding economic policies and their impact on inflation and prices. The selection of these clusters was grounded in their ability to capture diverse yet economically pertinent aspects of the price-related discourse on Twitter, thus aligning with our research's objective to understand the multifaceted nature of public sentiment on inflation. After this filter process, we reduce the total number of tweets from 85.4 Millions to 2.4 Millions.

Tweets classified under other topics were removed in further analysis. Most of these were related to specific product offers, commercial promotions, or different non-economic contexts of 'price', which likely introduced too much noise for the precise economic focus of our study. While these topics might provide insights into general pricing discourse, they do not contribute to the nuanced understanding of economic conditions, market sentiments, or policy impacts that are central to our analysis. Eliminating these tweets ensured a concentrated examination of the economic discourse pertinent to price movements and public sentiment, minimizing the potential noise and enhancing the potential forecasting power of our indicators.

2.3 Sentiment Analysis

In our sentiment analysis, we employed a dictionary approach to systematically categorize bigrams and trigrams extracted from tweets. This methodology involved manually labeling the 2000 most frequent n-gram based on its contextual relevance to economic conditions, specifically focusing on price movements. N-grams indicative of rising prices were assigned a '+1' to reflect an inflationary sentiment, while those suggesting falling prices were labeled with '-1', capturing deflationary trends. Neutral or unrelated n-grams were designated with a '0', acknowledging their presence but excluding them from the economic sentiment analysis. This systematic approach to n-gram categorization is encapsulated in the formula shown as Equation 13, where s_i is the sentiment score of a n-gram.

$$s_i = \begin{cases} 1 & \text{if n-gram indicates rising prices} \\ -1 & \text{if n-gram indicates falling prices} \\ 0 & \text{otherwise (neutral or unrelated)} \end{cases} \tag{13}$$

Table 20 displays the list of the top ten n-grams that indicate price rises and falls. On the left side, the n-grams associated with price increases are listed, with 'price increase'

TABLE 20 – First 10 most frequent N-Grams by occurrences indicating positive and negative Price Changes (with English Translations)

Price Increases		*Price Decreases*	
N-Gram (Translation)	**Count**	**N-Gram (Translation)**	**Count**
aumento preço (price increase)	55326	queda preço (price drop)	40312
alta preço (high price)	31978	reducao preço (price reduction)	35159
alta preços (high prices)	28317	reduz preço (reduces price)	25962
aumento preços (price increase)	23257	queda preços (price drop)	23628
fica caro (becomes expensive)	17876	petrobras reduz preço (Petrobras reduces price)	17369
aumenta preço (raises price)	15954	reduzir preço (reduce price)	15975
eleva preço (raises price)	13231	reducao preços (price reduction)	12864
reajuste preços (price adjustment)	11847	reduz preço gasolina (reduces gasoline price)	10814
caro apartir (expensive from)	11195	reducao preço gasolina (gasoline price reduction)	8607
aumento preço gasolina (gasoline price increase)	9650	precos baixos (low prices)	8327

Note: This table showcases the ten most prevalent N-Grams associated with price escalations and reductions, extracted from a corpus of tweets categorized under the 'Economy' topic by our topic model spanning from 2011 to 2022. The counts denote the frequency of each N-Gram within the dataset. Translations are provided for non-Portuguese speakers to understand the context of each N-Gram.

and 'high price' being the most frequent, occurring 55,326 and 31,978 times, respectively. On the right side, the table shows the n-grams indicative of price decreases, led by 'price drop' and 'price reduction', which appear 40,312 and 35,159 times.

The next phase involved analyzing the sentiment at the tweet level. For each tweet, we identified the occurrence of n-grams that had been manually labeled as '+1' or '-1'. In cases where a tweet contained multiple n-grams indicating price changes, the specific sentiment of the tweet was determined by calculating the average of these n-gram sentiment values, as shown in Equation 14.

$$\text{Inflation Sentiment (Tweet)} = \frac{1}{N}\sum_{i=1}^{N} s_i \tag{14}$$

2.4 Temporal Aggregation

In our effort to build a sentiment forecasting index, we deploy three different aggregation methods to consolidate tweet sentiments into meaningful time series indicators. Each method aggregates individual tweet sentiments into a daily score, followed by further processing to analyze trends over different timeframes. The three approaches are described below:

- **Net Daily Sentiment - Moving Average:** This method starts by aggregating the sentiment scores of individual tweets for each day to derive a net daily sentiment score. It reflects the overall sentiment of the public for that day, considering both positive and negative sentiments. The net daily sentiment, denoted as $s_{net,daily}$, is the sum of individual tweet sentiments divided by the total number of tweets for the day:

$$s_{net,daily} = \frac{\sum_{i=1}^{N} s_i}{N}$$

where s_i is the sentiment score of the i^{th} tweet, and N is the total number of tweets for that day. This daily score is then subjected to a moving average over chosen periods (30, 60, 90, and 120 days) to observe sentiment trends over time.

- **Mean Sentiment - Moving Average:** This method also involves aggregating daily tweet sentiments. However, it focuses specifically on the average sentiment value, rather than the net sentiment. The mean daily sentiment, $s_{mean,daily}$, is calculated using the same formula as the net daily sentiment. This average is then smoothed using a moving average over the selected timeframes, providing a different perspective on the sentiment trends.

- **Net Daily Sentiment - Exponentially Weighted Moving Average:** This method, while initially aggregating daily tweet sentiments equals the first method, employs an Exponentially Weighted Moving Average for smoothing. The EWM approach places more weight on recent data, making it more responsive to new trends. The calculation for EWM, s_{ewm}, is as follows:

  $$s_{ewm,t} = \alpha \cdot s_t + (1 - \alpha) \cdot s_{ewm,t-1}$$

  Here, s_t is the daily sentiment score, and α is the smoothing factor, determining the weight given to more recent data.

2.5 Alternative Approach: News Sources Subset

In complementing our primary methodology, we introduce an alternative approach that narrows its focus to Twitter data exclusively from news sources. This subset analysis targets tweets originating from reputable and verified news agencies in Brazil. The central objective of this approach is to investigate how narratives and sentiments concerning inflation and economic trends, as conveyed by news sources, might provide more precise and timely forecasts in comparison to the broader public discourse observed in our main dataset.

For this approach, we have compiled a comprehensive dataset encompassing all tweets posted by a select group of news sources, spanning the same January 2010 to December 2022 timeframe as our main analysis. This dataset was compilled by identifying verified Twitter accounts of leading national news organizations, newspapers, and television stations, with a focus on their credibility, reach, and relevance to Brazilian economic news. The full list is displayed in table 21. Contrasting with our main approach, which was restricted to tweets containing predefined keywords, this subset includes the entire corpus of tweets from the selected news sources. This broader scope aims to capture a more holistic view of the economic narrative as presented by these agencies.

Despite the larger volume of tweets in this subset, we maintain methodological consistency by employing the same Dirichlet Multinomial Mixture (DMM) topic model used

TABLE 21 – Selected news Twitter accounts

User	News Source	Followers
@BandJornalismo	Band Jornalismo	0.8
@bbcbrasil	BBC News Brasil	3.4
@CBNoficial	Rádio CBN	0.4
@correio	Correio Braziliense	0.9
@Estadao	Estadão	7.5
@exame	Revista Exame	2.9
@folha	Folha de S.Paulo	8.8
@g1	G1.com	14.9
@GloboNews	GloboNews	5.7
@infomoney	InfoMoney	0.5
@jornalextra	Jornal Extra	1.1
@JornalOGlobo	Jornal O Globo	7.3
@portalR7	Portal R7.com	5.1
@radiobandnewsfm	Rádio BandNews FM	1.5
@RevistaISTOE	Revista ISTOÉ	1.9
@sbtjornalismo	SBT Jornalismo	0.5
@Terra	Portal Terra	3.0
@UOLNoticias	Portal UOL	5.7
@valoreconomico	Valor Econômico	2.6
@VEJA	Revista Veja	9.1

Note: The table lists selected Twitter accounts of news outlets relevant to Brazilian news coverage. Follower counts are denoted in millions and represent the number as of July 2023. These counts were sourced from Twitter's public metrics and are subject to change over time. The selection of news sources was based on their influence and coverage in the Brazilian media landscape.

in our main analysis. This consistency is crucial for ensuring comparability and coherence in our study. The model, already trained for the main dataset, will be applied to predict and categorize the topics of the news sources' tweets. This approach allows us to filter and focus on tweets specifically relevant to economic conditions and trends. Furthermore, in our alternative approach focusing on news sources, it is essential to mention that the temporal aggregation of data will follow the same methodology as described in our primary analysis. This ensures consistency in the way we handle and interpret data across both datasets, allowing for accurate comparisons and coherent conclusions.

3 CORRELATION WITH FOCUS SURVEY

3.1 Full sample

In this section, we delve into a comparative analysis of our Twitter-based inflation indicators with the Focus forecasting survey, which serves as a benchmark for inflation expectations in Brazil. The Focus survey, compiled by the central bank, reflects the consensus forecast from financial market professionals and is widely regarded as a reliable predictor of future inflation.

To align our analysis with the timing of the Focus survey, we concentrate on the data corresponding to the first day of the month following the reference month—effectively positioning our analysis at t-1. This approach allows us to evaluate the immediacy and predictive quality of our indicators since nowcasting at t-1 provides an assessment of inflation expectations just before the release of official data. By focusing on t-1, we aim to capture the most current sentiment and expectations, offering a near-real-time perspective on inflation trends. We used the 60-day MA for the first two indicators and the alpha = 0.1 for the third. As our indicator is in real-time, we use the data for the exact same day. [12]

The figure 13 illustrates a side-by-side comparison between the Twitter-based inflation indicators and the Focus survey nowcasting data for inflation, standardized and plotted over the same time scale. The first figure displays the regular dataset while the second figure focuses on the subset of tweets from news sources. In both plots, the x-axis represents the time frame from January 2011 to the present, while the y-axis shows the standardized value of the indicators, facilitating a direct comparison between the different data series. Observations reveal that the Twitter-based indicators exhibit fluctuating patterns, sometimes closely mirroring the trends of the Focus survey. These variations offer a visual understanding of the indicators' performance and their potential to reflect inflation sentiment accurately.

To quantify the relationship observed in the plots, a regression analysis was conducted with the Focus survey inflation nowcasting as the dependent variable and the Twitter-based indicators as the independent variables. The results are shown in table 22. A positive coefficient indicates that the Twitter indicator moves in tandem with the Focus survey nowcasting, suggesting that the sentiment expressed on Twitter is reflective of inflation expectations. By examining the plots and regression results, we can infer the potential of Twitter-based inflation indicators to serve as a real-time gauge of inflation sentiment.

3.2 News subset

In the news subset analysis, we apply the same methodology as in the full sample to indicators derived solely from tweets of news accounts. This subset provides a narrative concentrated on economic news, which is expected to give a different perspective on inflation sentiment. In Figure 14, the sentiment indicators derived from the news subset are juxtaposed with the Focus survey's nowcasting of inflation. The plot demonstrates periods where the news-sourced sentiment indexes tracks the survey nowcasting closely, mainly in more recent times. However, in relation to the general indicators, they are more volatile, probably due to a smaller sample.

The results of the regression estimation is displayed in 23. Regression analysis of the news subset against the Focus survey data yields slightly higher R^2 and correlation values, implying that the sentiments expressed by news outlets may serve as a more precise

[12] It is crucial to emphasize that the Focus survey inherently carries a publication delay of one week. Consequently, our analysis juxtaposes a real-time indicator, derived from the immediacy of Twitter data, with information that will only become publicly available a week later.

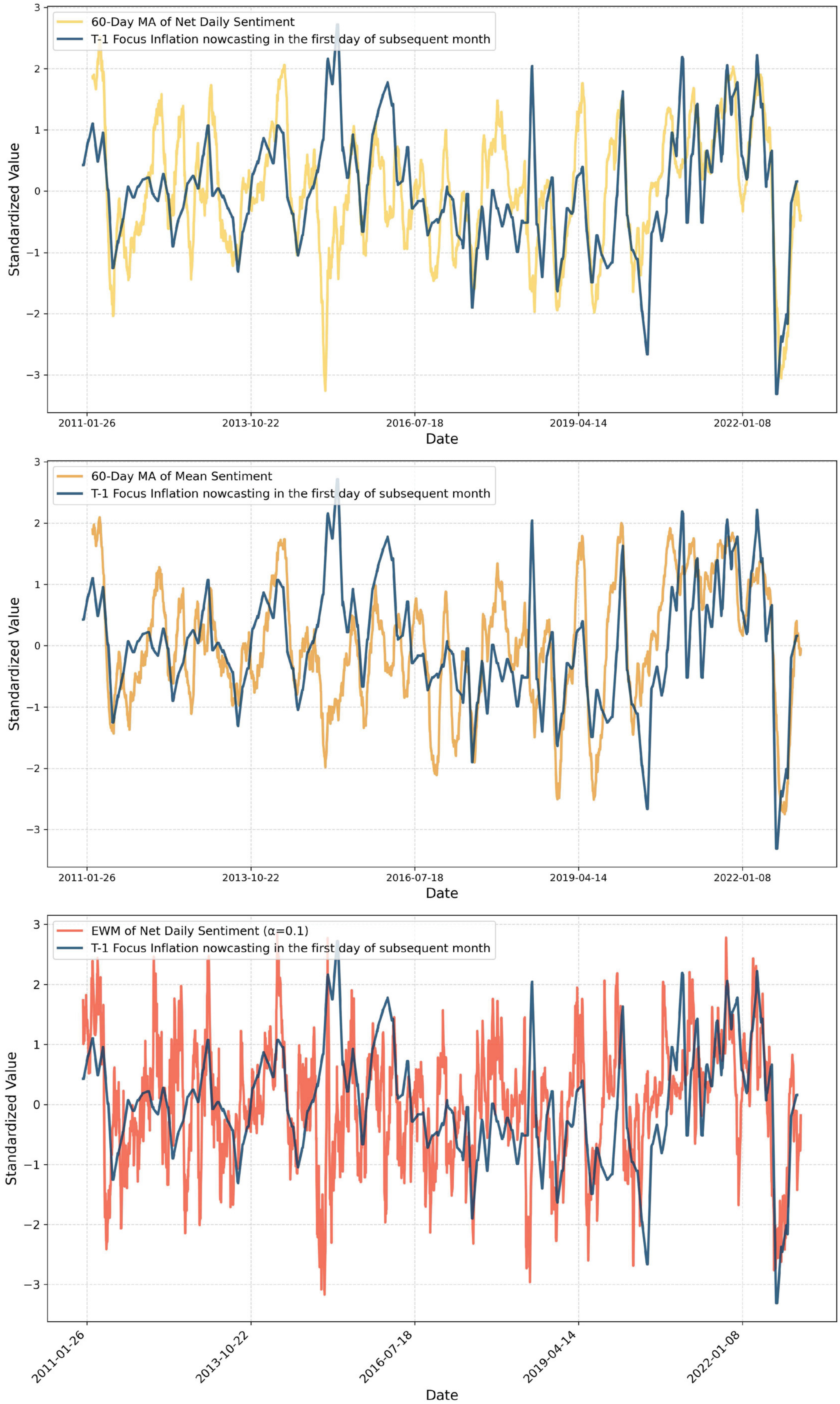


FIGURE 13 – Inflation expectations calculated with Twitter - General Twitter Sample

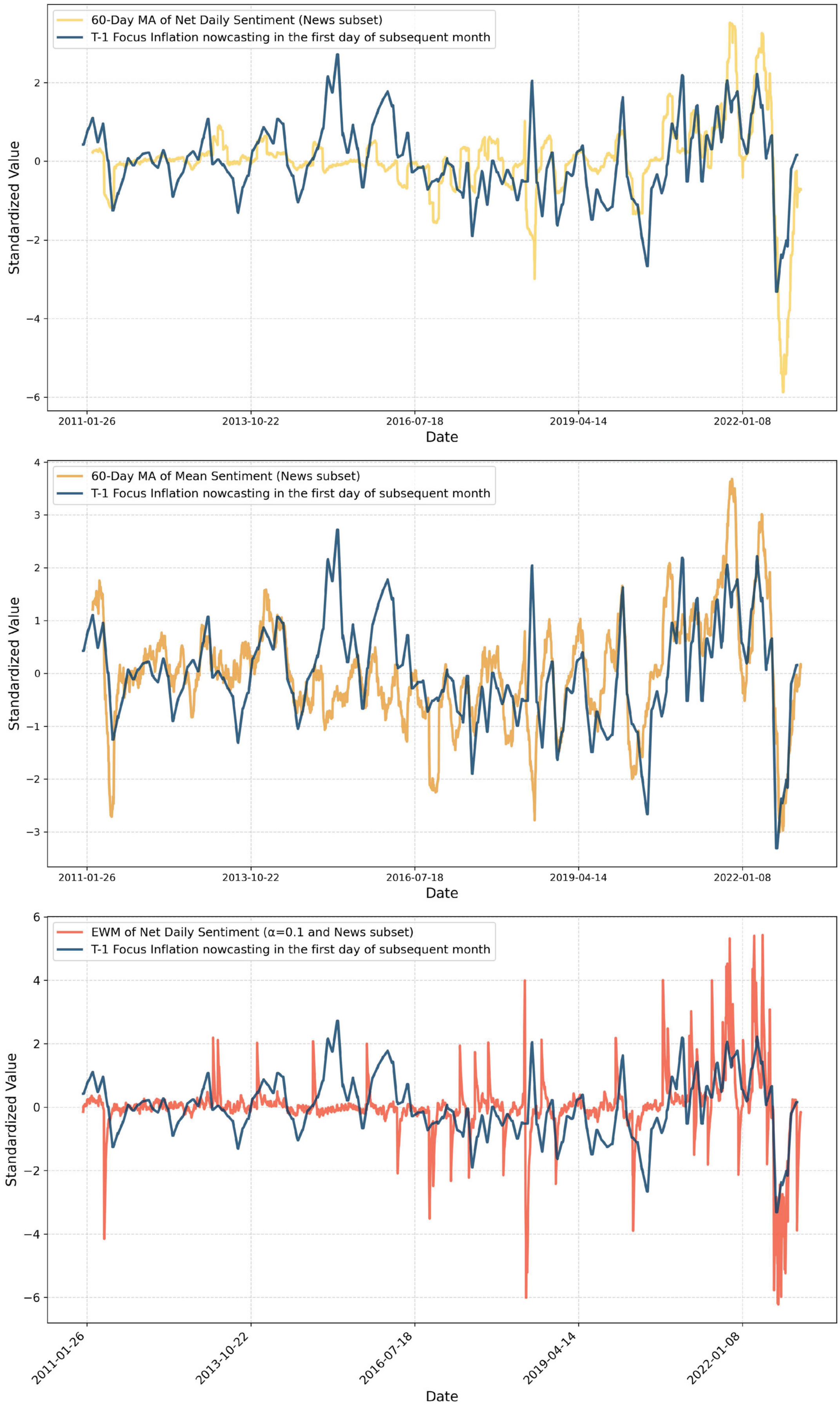


FIGURE 14 – Inflation expectations calculated with Twitter - News Subset

TABLE 22 – Regression Results for Inflation Indicators Using General Twitter Sample

	Coefficient	SE	N	R^2	Correlation
Net Daily Sentiment - Moving Average					
30-day MA	0.132	0.028	140	0.141	0.375
60-day MA	0.160	0.026	140	0.215	0.463
90-day MA	0.149	0.026	140	0.193	0.439
120-day MA	0.131	0.026	140	0.156	0.395
Mean Sentiment - Moving Average					
30-day MA	0.123	0.027	140	0.131	0.361
60-day MA	0.157	0.026	140	0.214	0.462
90-day MA	0.152	0.026	140	0.204	0.451
120-day MA	0.133	0.026	140	0.161	0.401
Net Daily Sentiment - Exponentially Weighted Moving Average					
$\alpha =$ 0.1	0.147	0.026	140	0.195	0.441
$\alpha =$ 0.3	0.128	0.024	140	0.171	0.413
$\alpha =$ 0.5	0.112	0.024	140	0.138	0.371

Note: This table represents regression outcomes for various twitter-based inflation indicators, with the dependent variable being the Focus survey nowcasting on the first day of the subsequent month. 'SE' denotes standard errors, 'N' the sample size, 'R^2' the coefficient of determination, and 'Correlation' the correlation coefficient between the indicator and the Focus survey nowcasting.

proxy for inflation sentiment. The closer correlation observed in the plot and the regression results highlights the potential of leveraging news accounts on Twitter for real-time economic sentiment analysis. Thus, analogous to the full sample, the indices constructed from this subset demonstrate a capability to measure inflation to a certain extent.

TABLE 23 – Regression Results of Inflation Indicators from News Subset

	Coefficient	SE	N	R^2	Correlation
Net Daily Sentiment - Moving Average					
30-day MA	0.145	0.025	140	0.198	0.444
60-day MA	0.167	0.024	140	0.258	0.507
90-day MA	0.150	0.025	140	0.209	0.457
120-day MA	0.142	0.026	140	0.182	0.426
Mean Sentiment - Moving Average					
30-day MA	0.146	0.026	140	0.187	0.433
60-day MA	0.169	0.024	140	0.259	0.508
90-day MA	0.164	0.025	140	0.241	0.491
120-day MA	0.152	0.025	140	0.212	0.460
Net Daily Sentiment - Exponentially Weighted Moving Average					
$\alpha = 0.1$	0.139	0.025	140	0.185	0.429
$\alpha = 0.3$	0.114	0.025	140	0.129	0.358
$\alpha = 0.5$	0.156	0.024	140	0.230	0.479

Note: This table represents regression outcomes for various news-twitter-based inflation indicators, with the dependent variable being the Focus survey nowcasting on the first day of the subsequent month. 'SE' denotes standard errors, 'N' the sample size, 'R^2' the coefficient of determination, and 'Correlation' the correlation coefficient between the indicator and the Focus survey nowcasting.

4 OUT-OF-SAMPLE FORECASTING

In this section, we show the effectiveness of indices based on Twitter data for predicting inflation as measured by the IPCA (Broad Consumer Price Index) over various time horizons. Given the nature of our indices, they are expected to be more suited for nowcasting - real-time forecasting - rather than for longer time horizons. Nevertheless, we will also test their performance for 3, 6, 9, and 12-month horizons to assess whether the extracted data contains additional information useful for longer-term forecasts. For this exercise, we employed an expanding window approach beginning in June 2016 to estimate the models. This strategy aims to balance the trade-off between an adequate initial length of the training set and a sufficient evaluation period.

For forecasting purposes, linear regressions will be estimated in the following format to predict the monthly cumulative IPCA (π_t):

$$\pi_t = \mu + \phi \text{Focus}_{t-h|t} + \beta \text{Twitter}_t + \epsilon_t \tag{15}$$

In this equation, the term $\text{Focus}_{t-h|t}$ represents the median of the Focus survey's forecast for IPCA available at time t for the prediction horizon h. The term Twitter_t corresponds to the Twitter-based indicator variable and ϵ_t is the error term at time t, accounting for unexplained variation in the IPCA.

Additionally, as a benchmark, we will estimate a Focus bias correction model, similar to the approach used by Boaretto et al. (2023):

$$\pi_t = \alpha + \beta \text{Focus}_{t|t-h} + u_t \tag{16}$$

The results are measured according to the Root Mean Squared Error (RMSE) ratio relative to the RMSE of the median Focus predictions for the same horizon. To determine whether one model consistently provides better predictions than the other, we employ the Diebold-Mariano test, using the Focus bias correction as a benchmark for comparison. By contrasting the models with Twitter-based indicators against this benchmark, the DM test helps us ascertain if there are statistically significant improvements in forecast accuracy offered by them.

Table 24 presents the out-of-sample forecasting results for models using Twitter-based indicators over time horizons ranging from 3 to 12 months. The Focus Bias Correction model serves as a baseline, demonstrating consistent performance across all time horizons with RMSE ratios close to 1, varying from 0.993 to 1.002. This model's stability provides a benchmark for assessing the impact of Twitter data on economic forecasting.

The Net Daily Sentiment models, particularly those with shorter moving averages like 30-day and 60-day MAs, show significant improvements in nowcasting. However, for longer-term

TABLE 24 – Out-of-Sample Results for Twitter-Based Indicators

Model	Nowcasting	3 Months	6 Months	9 Months	12 Months
Focus Benchmark					
Focus	1	1	1	1	1
Focus Bias Correction	0.933	0.9338	0.994	1.002	1.002
Twitter Indicators - Full sample					
Net Daily Sentiment (30-day MA)	0.926	0.964**	0.993	0.971	0.926**
Net Daily Sentiment (60-day MA)	0.927	0.978	0.994	0.969	0.926*
Net Daily Sentiment (90-day MA)	0.936	0.988	0.996	0.970	0.927
Net Daily Sentiment (120-day MA)	0.938	0.990	0.992	0.967	0.925
Mean Sentiment (30-day MA)	0.907**	0.946**	0.980	0.953*	0.901***
Mean Sentiment (60-day MA)	0.917	0.957	0.978	0.942*	0.893**
Mean Sentiment (90-day MA)	0.926	0.970	0.975	0.934*	0.887**
Mean Sentiment (120-day MA)	0.928	0.974	**0.967**	**0.924***	**0.881****
Net Daily Sentiment (EWM - $\alpha = 0.1$)	0.920	**0.956*****	0.991	0.978	0.930**
Net Daily Sentiment (EWM - $\alpha = 0.3$)	0.937	0.962***	0.992	0.983	0.937**
Net Daily Sentiment (EWM - $\alpha = 0.5$)	0.943	0.967**	0.992	0.983	0.938**
Twitter Indicators - News Only					
Net Daily Sentiment (30-day MA)	0.934	0.970	1.008	0.992	0.945
Net Daily Sentiment (60-day MA)	0.947	0.992	1.006	0.992	0.946
Net Daily Sentiment (90-day MA)	0.962	1.006	1.014	0.998	0.951
Net Daily Sentiment (120-day MA)	0.959	1.012	1.013	0.998	0.952
Mean Sentiment (30-day MA)	**0.905***	0.949	0.980	0.960	0.912
Mean Sentiment (60-day MA)	0.940	0.968	0.979	0.955	0.906
Mean Sentiment (90-day MA)	0.940	0.981	0.981	0.955	0.905
Mean Sentiment (120-day MA)	0.937	0.987	0.981	0.954	0.905
Net Daily Sentiment (EWM - $\alpha = 0.1$)	0.946	0.968	0.997	0.983	0.945
Net Daily Sentiment (EWM - $\alpha = 0.3$)	0.951	0.959**	0.993	0.982	0.947
Net Daily Sentiment (EWM - $\alpha = 0.5$)	0.946	0.974	1.002	0.988	0.946

Note: This table presents metrics for the out-of-sample period from June 2016 to December 2022. The models were estimated using an expanding window approach. The best-performing model for each horizon is highlighted in bold. Significance levels are indicated by stars: *, **, and *** representing 10%, 5%, and 1% levels respectively, based on a one-sided Diebold-Mariano test against the Focus Bias Correction benchmark model, excluding Twitter data.

forecasting, the 120-day moving average in the Mean Sentiment model stands out. With an RMSE ratio of 0.928 at the initial period, this model shows a distinct proficiency in predicting economic trends over extended periods, as evidenced by its progressively decreasing RMSE ratios, reaching 0.881 at the 12-month horizon. This trend suggests that while shorter moving averages are useful in capturing immediate shifts in sentiment, longer moving averages like the 120-day MA in the Mean Sentiment model provide a more stable and reliable indicator for longer-term economic trends. The longer moving average smoothens out short-term fluctuations and provides a clearer picture of the underlying trend, making it particularly valuable for forecasting horizons extending beyond 9 months.

When focusing solely on 'News only' content in the Net Daily Sentiment and Mean Sentiment models, the results exhibit a mixed trend. While maintaining nowcasting capabilities comparable to the baseline Focus Bias Correction model, their effectiveness in longer-term forecasting is reduced. This is evident as these models do not show statistical significance in the Diebold-Mariano (DM) test across any horizon. The lack of statistical significance suggests that when limited to news content, the predictive power of Twitter-based sentiment indicators may not be robust enough to significantly outperform the baseline model, particularly for forecasts extending beyond the short term. This could be attributed to higher volatility in these indicators, possibly due to the influence of a smaller number of tweets, as indicated in Figure 14. Therefore, while these models are effective in capturing real-time shifts in sentiment for nowcasting, their reliability decreases for longer-term forecasting.

Overall, the primary takeaway from this section is that our Twitter-based indicators demonstrate robust performance and enhance the predictive capabilities of focused forecasts. Nevertheless, our findings are somewhat more conservative when juxtaposed with a closely related exercise by Boaretto et al. (2023), especially regarding their high-dimensional model's efficacy over extended periods. This difference is particularly noticeable in the performance of long-term forecasts. This suggests that while our methodology is effective to extract useful data from Twitter, it may benefit from more advanced techniques.

5 FINAL REMARKS

This paper has illustrated the feasibility of leveraging tweets to extract pertinent information for inflation forecasting, marking a significant stride in economic analysis. The indicators we developed demonstrate a good correlation with the Focus Bulletin in the context of inflation nowcasting. More notably, our out-of-sample estimation results have indicated substantial enhancements in forecasting accuracy compared to the survey-based expectations. This improvement is pronounced across various forecasting horizons, with a remarkable distinction in the context of longer-term projections.

An interesting insight from our research is the superior performance of indicators derived from the general public, encompassing a diverse and unrestricted sample, as opposed to those solely based on accounts from established news outlets. The reason for this could be partly attributed to the smaller sample of news tweets related to price changes, which led to higher overall variance.

Future studies might look into more sophisticated and detailed techniques to make better use of this type of textual data. With the recent advancements in Natural Language Processing, there's a real opportunity to extract more refined information from social media data, which could enhance economic forecasting. Additionally, integrating machine learning algorithms with NLP techniques could offer new dimensions in data analysis, leading to more accurate and insightful economic models.

APPENDIX

TABLE 25 – Clusters created with DMM Topic Model and the Top-10 words associated with each one

Cluster Number	Words
0	desconto, cupom, cupom desconto, ganhe, compra, compras, site, off, codigo, voce
1	adquira, conheca, precos, produtos, dinheiro, usar, diversos, site, adquira usar, precos adquira
2	gasolina, preco, barata, petrobras, gasolina barata, precos, preco gasolina, combustiveis, diesel, gas
3	desconto, nao, pra, cupom, dar, vou, cupom desconto, dar desconto, ter, desconto pra
4	preco, vip, expresso, vip expresso, bigfollow, desconto, tambem, hoje, assine, assine tambem
5	preco, nao, caro, precos, pra, barato, bem, vale, acho, sao
6	inflacao, precos, preco, alta, mercado, indice, queda, sobe, ano, maior
7	preco, paga, pago, caro, ler, preco pago, alto, saber, saber ler, ter
8	desconto, preco, receba desconto, receba, email, surpresa, desconto surpresa, surpresa email, receba desconto surpresa, desconto surpresa email
9	preco, paga, preco paga, barato, momentos, saudade, saudade preco, viver, inesqueciveis, momentos inesqueciveis
10	preco, nao, vida, nao preco, tudo, precojusto, coisas, nessa, valor, algumas
11	desconto, gratis, apenas, frete, apenas frete, preco, frete gratis, apenas frete gratis, promocao, gratis promocao
12	caro, nao, amigo, caro amigo, voce, barato, pra, dia, bem, bom
13	desconto, ate, ate desconto, produtos, compra, coletiva, compra coletiva, produtos ate, servicos, produtos ate desconto
14	oferta, amazon, desconto, oferta amazon, cupom, promocao, confira, cupons, cupomdedesconto, super
15	preco, nao, gasolina, pra, cara, caro, precos, preco gasolina, tudo, agora
16	precos, preco, desconto, oferta, ate, sao, via, caro, vagas, dia
17	caro, custa, pode, custa caro, amor, gasolina, alcool, acaba, rapido, caro acaba
18	preco, desconto, apenas, oferta, acesse, oferta desconto, lcdbr, apenas acesse, vale, vale preco
19	barato, nao, pra, caro, comprar, sai, vou, bem, ainda, ate
20	barato, bala, comprado, unidade, curiosidade, comprado barato, seguidor, sabia, voc, unidade seguidor comprado
21	hoje, vip, desconto, planos, planos vip, desconto hoje, estao, estao desconto, estao desconto hoje, beleza
22	preco, hoje, valor, metade, metade preco, nada, tudo, preco tudo, hoje metade preco, hoje metade
23	estao, twitter, desconto, beleza, hoje, estao desconto, galera, vip, hoje galera, beleza planos
24	caro, nao, pra, pagar, pagar caro, ter, voce, caro pra, barato, gente
25	preco, bom, nao, vida, nao preco, vida nao, vida nao preco, bom vida, bom vida nao, perfil
26	precos, preco, melhores, voce, melhor, qualidade, aqui, confira, produtos, melhor preco
27	preco, nao, pagar, nao preco, vida, voce, alto, pra, tudo, ter
28	preco, comprar, nao, desconto, pra, barato, caro, comprei, black, bom
29	desconto, ate, desconto ate, hoje, ate hoje, desconto ate hoje, cupom, cupom desconto, bigfollow, bigfollow desconto
30	preco, nao, nao preco, ver, dia, acordar, bom, pra, casa, amigos
31	barato, preco, caro, bom, nao, pra, aqui, alguem, lugar, vinho
32	desconto, ate, oferta, voce, dia, preco, nao, promocao, hoje, pra
33	preco, caro, tudo, beber, amar, beber caro, nada, tudo preco, valor, mundo
34	inflacao, nao, precos, preco, governo, brasil, pra, bolsonaro, aumento, pais
35	caro, pra, nao, tao, tao caro, tudo, vou, comprar, caro pra, demais
36	caro, nao, jogador, preco, barato, pra, jogador caro, oferta, time, bom
37	oferta, promocao, promocao oferta, preco, desconto, link, confira, pode, moda, estoque
38	desconto, ate, ate desconto, liquidacao, promocao, frete, gratis, loja, oferta, aproveite
39	preco, nunca, pago, preco pago, pago nunca, preco pago nunca, amada, nunca amada, verdade, pago nunca amada
40	desconto, seguidores, hoje, preco, bigfollow, aqui, planos, desconto planos, hoje aqui, planos seguidores
41	preco, youtube, video, barato, video youtube, gostei, gostei video, gostei video youtube, precos, brasil
42	preco, ingresso, precos, ingressos, nao, caro, barato, pra, show, jogo
43	desconto, ate, ate desconto, more, learn, learn more, account, the, twitter, because violates the
44	caro, desconto, siga, gringo, gringo caro, baladas, desconto melhores, desconto melhores baladas, melhores baladas, melhores
45	desconto, oferta, promocao, desconto oferta, cupom, relampago, oferta relampago, promocao cupom, desconto promocao, relampago desconto
46	preco, nao, precos, caro, barato, oferta, pra, sao, mercado, brasil
47	desconto, codigo, uber, cupom, use, ganhe, use codigo, primeira, ganhe desconto, ate
48	oferta, emprego, oferta emprego, seguidores, compre, trabalho, barato, bigfollow, empleo, milhares
49	preco, pra, nao, precos, ver, vou, comprar, ver preco, fazer, aqui

Note:

This table presents the results of a topic modeling analysis conducted on Twitter data, focusing on conversations related to price movements. Each cluster number identifies a distinct topic, with the listed words representing the most frequent and representative terms associated with that topic.

CONCLUSION

This essay focused on the use of internet-derived data for economic forecasting. In this context, we evaluated the use of data from Google Trends and Twitter, the latter involving a set of tweets from news sources and the general public. Regarding Google Trends data, it was demonstrated that they can perform better than the same models using traditional economic data for predicting unemployment rates. A key point is that the performance of the model with Google Trends data improves with greater internet usage in the population. Regarding news tweets, we showed that the sentiment of topics extracted from Twitter related to each variable correlates with the variables themselves. Furthermore, we demonstrated that they contain relevant information for forecasting. Lastly, we showed that tweets related to price changes can hold important information for future inflation expectations, especially over longer horizons. Overall, this thesis contributes by providing more evidence to the literature that internet data can contain information not found in traditional data and can improve forecasting models.

Future research in this area has several possible directions. First, methods dealing with unstructured data in more complex ways than those handled here, such as Transformers, could yield even more significant results. Another important point is the exploration of data from other websites and social networks, like Instagram, which has a larger and less niche audience than Twitter. Comparisons between different data sources from internet can be conducted to highlight the differences between them. Further evaluations can be carried out to assess whether the relationship between internet usage and the performance of models based on internet data holds true for sources other than Google Trends.

REFERÊNCIAS

AARONSON, Daniel et al. Forecasting unemployment insurance claims in realtime with Google Trends. **International Journal of Forecasting**, Elsevier B.V., n. 40, 2021. ISSN 01692070. DOI: `10.1016/j.ijforecast.2021.04.001`. Disponível em: <`https://doi.org/10.1016/j.ijforecast.2021.04.001`>. Citado 2 vezes nas páginas 14, 15.

ADU, Williams Kwasi; APPIAHENE, Peter; AFRIFA, Stephen. VAR, ARIMAX and ARIMA models for nowcasting unemployment rate in Ghana using Google trends. **Journal of Electrical Systems and Information Technology**, SpringerOpen, v. 10, n. 1, p. 1–16, 2023. Citado 1 vez na página 15.

ANGELICO, Cristina et al. Can we measure inflation expectations using Twitter? **Journal of Econometrics**, Elsevier, v. 228, n. 2, p. 259–277, 2022. Citado 4 vezes nas páginas 64, 66.

ANTENUCCI, Dolan et al. **Using social media to measure labor market flows**. [S.l.], 2014. Citado 1 vez na página 41.

APRIGLIANO, Valentina; ARDIZZI, Guerino; MONTEFORTE, Libero. Using payment system data to forecast economic activity. **60th issue (October 2019) of the International Journal of Central Banking**, 2019. Citado 1 vez na página 20.

APRIGLIANO, Valentina; EMILIOZZI, Simone et al. The power of text-based indicators in forecasting Italian economic activity. **International Journal of Forecasting**, Elsevier, v. 39, n. 2, p. 791–808, 2023. Citado 2 vezes nas páginas 41, 53.

ARDIA, David; BLUTEAU, Keven; BOUDT, Kris. Questioning the news about economic growth: Sparse forecasting using thousands of news-based sentiment values. **International Journal of Forecasting**, Elsevier, v. 35, n. 4, p. 1370–1386, 2019. Citado 2 vezes nas páginas 40, 64.

AYAT, Leila; BURRIDGE, Peter. Unit root tests in the presence of uncertainty about the non-stochastic trend. **Journal of Econometrics**, Elsevier, v. 95, n. 1, p. 71–96, 2000. Citado 1 vez na página 18.

BAI, Jushan; NG, Serena. Forecasting economic time series using targeted predictors. **Journal of Econometrics**, Elsevier, v. 146, n. 2, p. 304–317, 2008. Citado 1 vez na página 20.

BAKER, Scott R; BLOOM, Nicholas; DAVIS, Steven J. Measuring Economic Policy Uncertainty. **The Quarterly Journal of Economics**, v. 131, n. 4, p. 1593–1636, 2016. Citado 2 vezes nas páginas 40, 46.

BANTIS, Evripidis; CLEMENTS, Michael P; URQUHART, Andrew. Forecasting GDP growth rates in the United States and Brazil using Google Trends. **International Journal of Forecasting**, Elsevier, v. 39, n. 4, p. 1909–1924, 2023. Citado 1 vez na página 14.

BARBAGLIA, Luca; CONSOLI, Sergio; MANZAN, Sebastiano. Forecasting with economic news. **Journal of Business & Economic Statistics**, Taylor & Francis, v. 41, n. 3, p. 708–719, 2023. Citado 2 vezes nas páginas 40, 64.

BARSKY, Robert B; SIMS, Eric R. Information, animal spirits, and the meaning of innovations in consumer confidence. **American Economic Review**, American Economic Association, v. 102, n. 4, p. 1343–1377, 2012. Citado 1 vez na página 40.

BLEHER, Johannes; DIMPFL, Thomas. Knitting Multi-Annual High-Frequency Google Trends to Predict Inflation and Consumption. **Econometrics and Statistics**, Elsevier, 2021. ISSN 2452-3062. Citado 1 vez na página 14.

BLEI, David M; NG, Andrew Y; JORDAN, Michael I. Latent Dirichlet Allocation. **Journal of Machine Learning Research**, v. 3, p. 993–1022, 2003. Citado 1 vez na página 42.

BOARETTO, Gilberto et al. What news and social media tell us about future inflation?, 2023. Citado 6 vezes nas páginas 41, 64, 65, 75, 77.

BOLLEN, Johan; MAO, Huina; PEPE, Alberto. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. In: 1. PROCEEDINGS of the international AAAI conference on web and social media. [S.l.: s.n.], 2011. v. 5, p. 450–453. Citado 1 vez na página 41.

BORUP, Daniel; CHRISTENSEN, Bent Jesper et al. Targeting predictors in random forest regression. **International Journal of Forecasting**, Elsevier, 2022. Citado 3 vezes nas páginas 20, 21.

BORUP, Daniel; SCHÜTTE, Erik Christian Montes. In search of a job: Forecasting employment growth using google trends. **Journal of Business Economic Statistics**, Taylor Francis, p. 1–15, 2020. ISSN 0735-0015. Citado 6 vezes nas páginas 14, 15, 18, 20.

BREIMAN, Leo. Random forests. **Machine learning**, Springer, v. 45, n. 1, p. 5–32, 2001. Citado 1 vez na página 22.

BULLIGAN, Guido; MARCELLINO, Massimiliano; VENDITTI, Fabrizio. Forecasting economic activity with targeted predictors. **International Journal of Forecasting**, Elsevier, v. 31, n. 1, p. 188–206, 2015. Citado 1 vez na página 20.

BULUT, Levent. Google Trends and the forecasting performance of exchange rate models. **Journal of Forecasting**, Wiley Online Library, v. 37, n. 3, p. 303–315, 2018. ISSN 0277-6693. Citado 1 vez na página 14.

BYBEE, Leland et al. **The structure of economic news**. [S.l.], 2020. Citado 4 vezes nas páginas 40, 41, 45, 64.

CHEN, Hailiang et al. Wisdom of crowds: The value of stock opinions transmitted through social media. **The Review of Financial Studies**, Oxford University Press, v. 27, n. 5, p. 1367–1403, 2014. Citado 1 vez na página 64.

CHOI, Hyunyoung; VARIAN, Hal. Predicting the present with Google Trends. **Economic record**, Wiley Online Library, v. 88, p. 2–9, 2012. Citado 1 vez na página 14.

CHOW, Gregory C; LIN, An-loh. Best linear unbiased interpolation, distribution, and extrapolation of time series by related series. **The review of Economics and Statistics**, JSTOR, p. 372–375, 1971. Citado 1 vez na página 53.

CORREA, Ricardo et al. Constructing a dictionary for financial stability, 2017. Citado 2 vezes nas páginas 41, 44.

DEMPSTER, Arthur P; LAIRD, Nan M; RUBIN, Donald B. Maximum Likelihood from Incomplete Data via the EM Algorithm. **Journal of the Royal Statistical Society: Series B (Methodological)**, v. 39, n. 1, p. 1–38, 1977. Citado 1 vez na página 43.

DIEBOLD, Francis X; MARIANO, Robert S. Comparing predictive accuracy. **Journal of Business & economic statistics**, Taylor & Francis, v. 20, n. 1, p. 134–144, 2002. Citado 1 vez na página 54.

DIEBOLD, Francis X; SHIN, Minchul. Machine learning for regularized survey forecast combination: Partially-egalitarian lasso and its derivatives. **International Journal of Forecasting**, Elsevier, v. 35, n. 4, p. 1679–1691, 2019. Citado 1 vez na página 15.

DILMAGHANI, Maryam. The racial 'digital divide'in the predictive power of Google trends data for forecasting the unemployment rate. **Journal of Economic and Social Measurement**, IOS Press, v. 43, n. 3-4, p. 119–142, 2018. ISSN 0747-9662. Citado 1 vez na página 14.

DÖPKE, Jörg; FRITSCHE, Ulrich; PIERDZIOCH, Christian. Predicting recessions with boosted regression trees. **International Journal of Forecasting**, Elsevier, v. 33, n. 4, p. 745–759, 2017. Citado 1 vez na página 15.

DRÄGER, Lena; LAMLA, Michael J; PFAJFAR, Damjan. Are survey expectations theory-consistent? The role of central bank communication and news. **European Economic Review**, Elsevier, v. 85, p. 84–111, 2016. Citado 1 vez na página 64.

ELLINGSEN, Jon; LARSEN, Vegard H; THORSRUD, Leif Anders. News media versus FRED-MD for macroeconomic forecasting. **Journal of Applied Econometrics**, Wiley Online Library, v. 37, n. 1, p. 63–81, 2022. Citado 2 vezes nas páginas 40, 64.

ELLIOTT, Graham; GARGANO, Antonio; TIMMERMANN, Allan. Complete subset regressions. **Journal of Econometrics**, Elsevier, v. 177, n. 2, p. 357–373, 2013. Citado 1 vez na página 22.

FERREIRA, Leonardo Nogueira et al. **Forecasting with VAR-teXt and DFM-teXt Models: exploring the predictive power of central bank communication**. [S.l.]: Banco Central do Brasil, 2021. Citado 1 vez na página 64.

FRIEDMAN, Jerome; HASTIE, Trevor; TIBSHIRANI, Robert et al. **The elements of statistical learning**. [S.l.]: Springer series in statistics New York, 2001. v. 1. Citado 1 vez na página 20.

GARCIA, Márcio G P; MEDEIROS, Marcelo C; VASCONCELOS, Gabriel F R. Real-time inflation forecasting with high-dimensional models: The case of Brazil. **International Journal of Forecasting**, Elsevier, v. 33, n. 3, p. 679–693, 2017. Citado 1 vez na página 22.

HARVEY, David I; LEYBOURNE, Stephen J; NEWBOLD, Paul. Tests for forecast encompassing. **Journal of Business & Economic Statistics**, Taylor & Francis, v. 16, n. 2, p. 254–259, 1998. Citado 1 vez na página 30.

HUBERT, Paul; LABONDANCE, Fabien. The signaling effects of central bank tone. **European Economic Review**, Elsevier, v. 133, p. 103684, 2021. Citado 1 vez na página 64.

JIAO, Peiran; VEIGA, Andre; WALTHER, Ansgar. Social media, news media and the stock market. **Journal of Economic Behavior & Organization**, Elsevier, v. 176, p. 63–90, 2020. Citado 1 vez na página 64.

KOPOIN, Alexandre; MORAN, Kevin; PARÉ, Jean-Pierre. Forecasting regional GDP with factor models: How useful are national and international data? **Economics Letters**, Elsevier, v. 121, n. 2, p. 267–270, 2013. Citado 1 vez na página 21.

LIMA, Luiz Renato; GODEIRO, Lucas Lúcio; MOHSIN, Mohammed. Time-varying dictionary and the predictive power of FED minutes. **Computational Economics**, Springer, v. 57, p. 149–181, 2021. Citado 1 vez na página 64.

LIN, Jianhao et al. Real-time macroeconomic projection using narrative central bank communication. **Journal of Applied Econometrics**, Wiley Online Library, v. 38, n. 2, p. 202–221, 2023. Citado 1 vez na página 64.

MATSA, Katerina Eva; SHEARER, Elisa. News use across social media platforms 2018. **Pew Research Center**, v. 10, 2018. Citado 1 vez na página 41.

MEDEIROS, Marcelo C; PIRES, Henrique F. The proper use of google trends in forecasting models. **arXiv preprint arXiv:2104.03065**, 2021. Citado 1 vez na página 18.

MEDEIROS, Marcelo C; SCHÜTTE, Erik Christian Montes; SOUSSI, Tobias Skipper. Global inflation forecasting: Benefits from machine learning methods. **Available at SSRN 4145665**, 2022. Citado 1 vez na página 24.

MEDEIROS, Marcelo C; VASCONCELOS, Gabriel F R et al. Forecasting inflation in a data-rich environment: the benefits of machine learning methods. **Journal of Business & Economic Statistics**, Taylor & Francis, v. 39, n. 1, p. 98–119, 2021. Citado 1 vez na página 15.

MIHAELA, Simionescu. Improving unemployment rate forecasts at regional level in Romania using Google Trends. **Technological Forecasting and Social Change**, Elsevier, v. 155, p. 120026, 2020. ISSN 0040-1625. Citado 2 vezes nas páginas 14, 15.

MIMNO, David et al. Optimizing Semantic Coherence in Topic Models. In: PROCEEDINGS of the 2011 Conference on Empirical Methods in Natural Language Processing. [S.l.: s.n.], 2011. P. 262–272. Citado 1 vez na página 43.

MINKA, Thomas P. Estimating a Dirichlet distribution. **Technical report, Microsoft Research**, 2000. Citado 1 vez na página 43.

NACCARATO, Alessia et al. Combining official and Google Trends data to forecast the Italian youth unemployment rate. **Technological Forecasting and Social Change**, Elsevier, v. 130, p. 114–122, 2018. Citado 1 vez na página 15.

NAGAO, Shintaro; TAKEDA, Fumiko; TANAKA, Riku. Nowcasting of the US unemployment rate using Google Trends. **Finance Research Letters**, Elsevier, v. 30, p. 103–109, 2019. ISSN 1544-6123. Citado 1 vez na página 14.

NG, Serena. Boosting recessions. **Canadian Journal of Economics/Revue canadienne d'économique**, Wiley Online Library, v. 47, n. 1, p. 1–34, 2014. Citado 1 vez na página 15.

NISAR, Tahir M; YEUNG, Man. Twitter as a tool for forecasting stock market movements: A short-window event study. **The journal of finance and data science**, Elsevier, v. 4, n. 2, p. 101–119, 2018. Citado 1 vez na página 41.

QIANG, Jipeng et al. Short text topic modeling techniques, applications, and performance: a survey. **IEEE Transactions on Knowledge and Data Engineering**, IEEE, v. 34, n. 3, p. 1427–1445, 2020. Citado 2 vezes nas páginas 41, 65.

SERMPINIS, Georgios et al. Inflation and unemployment forecasting with genetic support vector regression. **Journal of Forecasting**, Wiley Online Library, v. 33, n. 6, p. 471–487, 2014. Citado 1 vez na página 15.

SHEN, Dehua; URQUHART, Andrew; WANG, Pengfei. Does twitter predict Bitcoin? **Economics letters**, Elsevier, v. 174, p. 118–122, 2019. Citado 1 vez na página 64.

SIMIONESCU, Mihaela; CIFUENTES-FAURA, Javier. Forecasting national and regional youth unemployment in Spain using google trends. **Social Indicators Research**, Springer, v. 164, n. 3, p. 1187–1216, 2022. Citado 1 vez na página 15.

STIEGLITZ, Stefan; DANG-XUAN, Linh. Emotions and information diffusion in social media—sentiment of microblogs and sharing behavior. **Journal of management information systems**, JSTOR, p. 217–247, 2013. Citado 1 vez na página 41.

TETLOCK, Paul C. Giving content to investor sentiment: The role of media in the stock market. **The Journal of finance**, Wiley Online Library, v. 62, n. 3, p. 1139–1168, 2007. Citado 1 vez na página 40.

TIBSHIRANI, Robert. Regression shrinkage and selection via the lasso. **Journal of the Royal Statistical Society: Series B (Methodological)**, Wiley Online Library, v. 58, n. 1, p. 267–288, 1996. Citado 3 vezes nas páginas 15, 21, 46.

TIBSHIRANI, Ryan J et al. Exact post-selection inference for sequential regression procedures. **Journal of the American Statistical Association**, Taylor & Francis, v. 111, n. 514, p. 600–620, 2016. Citado 1 vez na página 46.

TILLY, Sonja; EBNER, Markus; LIVAN, Giacomo. Macroeconomic forecasting through news, emotions and narrative. **Expert Systems with Applications**, Elsevier, v. 175, p. 114760, 2021. Citado 2 vezes nas páginas 40, 64.

TUMASJAN, Andranik et al. Election forecasts with Twitter: How 140 characters reflect the political landscape. **Social science computer review**, Sage Publications Sage CA: Los Angeles, CA, v. 29, n. 4, p. 402–418, 2011. Citado 1 vez na página 41.

WOLOSZKO, Nicolas. Tracking activity in real time with Google Trends. OECD, 2020. Citado 2 vezes nas páginas 14, 16.

WRIGHT, Marvin N; WAGER, S; PROBST, P. Ranger: A fast implementation of random forests. **R package version 0.12**, v. 1, 2020. Citado 1 vez na página 23.

YIN, Jiaming; WANG, Jian. A dirichlet multinomial mixture model-based approach for short text clustering. **Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining**, p. 233–242, 2014. Citado 2 vezes nas páginas 42, 67.