UNIVERSIDADE DE SÃO PAULO
FACULDADE DE ECONOMIA, ADMINISTRAÇÃO, CONTABILIDADE E ATUÁRIA
DEPARTAMENTO DE ECONOMIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ECONOMIA

GUSTAVO ROMERO CARDOSO

WHAT'S IN A HEADLINE? NEWS IMPACT ON THE BRAZILIAN ECONOMY

O QUE HÁ EM UMA MANCHETE? O IMPACTO DAS NOTÍCIAS NA ECONOMIA BRASILEIRA

SÃO PAULO
2023

Prof. Dr. Carlos Gilberto Carlotti Júnior
Reitor da Universidade de São Paulo

Profa. Dra. Maria Dolores Montoya Diaz
Diretora da Faculdade de Economia, Administração, Contabilidade e Atuária

Prof. Dr. Claudio Ribeiro de Lucinda
Chefe do Departamento de Economia

Prof. Dr. Mauro Rodrigues Junior
Coordenador do Programa de Pós-Graduação em Economia

GUSTAVO ROMERO CARDOSO

WHAT'S IN A HEADLINE? NEWS IMPACT ON THE BRAZILIAN ECONOMY

Dissertação apresentada ao Programa de Pós-Graduação em Economia do Departamento de Economia da Faculdade de Economia Administração, Contabilidade e Atuária da Universidade de São Paulo, como requisito parcial para obtenção do título de Mestre em Ciências.

Área de Concentração: Teoria Econômica

Orientador: Prof. Dr. Márcio Issao Nakane

Versão Corrigida

(versão original disponível na Biblioteca da Faculdade de Economia, Administração, Contabilidade e Atuária)

SÃO PAULO

2023

Catalogação na Publicação (CIP)
Ficha Catalográfica com dados inseridos pelo autor

Cardoso, Gustavo Romero
What's in a headline? News impact on the Brazilian economy / Gustavo Romero Cardoso – São Paulo, 2023.
75 p.

Dissertação (Mestrado) - Universidade de São Paulo, 2024.
Orientador: Márcio Issao Nakane

1. Notícias. 2. Dados Textuais. 3. Latent Dirichlet Allocation. 4. Ciclos Econômicos. I. Universidade de São Paulo. Faculdade de Economia, Administração, Contabilidade e Atuária. II. Título

ACKNOWLEDGEMENTS

My deepest gratitude goes to my family—my father Djalma, my mother Luciane, and my sister Marina—whose support and love have been the unwavering pillars of my journey. A special mention to my loyal companion, Margot, whose presence has consistently brought joy to my days.

A heartfelt thank you to Prof. Márcio Nakane, whose guidance, valuable insights, and advice were essential to the success of this research. His availability and support have been invaluable.

My friends from the 'Sala do 10' truly deserve special recognition for their friendship and constant support, even from a distance. I'm particularly thankful to Kovashikawa, for his computational contributions and brilliant ideas; to Robert, for his expertise in web scraping; and to Lin, for his invaluable assistance with my master's degree subjects and the insightful discussions we've shared. Freire's help with the revisions has been greatly appreciated. Roman, Chen, Toninho, and Micael, your fellowship and support have also been essential.

I am grateful to Professors Ricardo Avelino and Paula Pereda, whose classes allowed me not only to serve as a teaching assistant but also fostered significant academic growth.

This study was financed by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

"A fanatic is one who can’t change his mind and won’t change the subject."
Winston Churchill
"If you only read the books that others are reading, you can only think what others are thinking."
Haruki Murakami

ABSTRACT

The purpose of this dissertation is to explore the impact of textual data on the Brazilian economic cycle. Utilizing a dataset of articles from "Valor Econômico" newspaper spanning July 2011 to December 2022, we employ the topic model Latent Dirichlet Allocation (LDA) to transform this textual data into a series of monthly topic proportions. From this output, we have developed two news indices - NITVP and NLTM - each with distinct methodologies but sharing the objective of assessing the influence of news topics on asset prices. Adopting the identification strategy of Larsen and Thorsrud (2019), we incorporate these indices into a structural VAR model to differentiate between news and noise shocks and to analyze their effects on macroeconomic variables. Our results reveal that news shocks, as captured by the news indices, significantly impact both asset prices and a range of macroeconomic indicators. Both news and noise shocks are found to be crucial in explaining a considerable proportion of the variance in asset prices over short and long-term periods, underscoring the pivotal role of news information in market dynamics. Furthermore, our findings, when contrasted with previous studies like Beaudry and Portier (2006), affirm the effectiveness of employing news indices to identify news shocks, as opposed solely on asset prices. This dissertation contributes to the field of economic literature by showcasing the significance of textual data analysis in understanding economic cycles and highlighting the potential and importance of news as a resource for providing a comprehensive view of the economy.

Keywords: news, textual data, Latent Dirichlet Allocation, economic cycles
JEL: C8, C55, E32

RESUMO

O objetivo desta dissertação é investigar o impacto dos dados textuais no ciclo econômico brasileiro. Utilizamos um conjunto de dados de artigos do jornal "Valor Econômico", abrangendo de julho de 2011 a dezembro de 2022, empregamos o modelo de tópicos Alocação Latente de Dirichlet (LDA) para transformar esses dados textuais em uma série de proporções mensais de tópicos. A partir deste resultado, desenvolvemos dois índices de notícias - NITVP e NLTM - cada um com metodologias distintas, mas compartilhando o objetivo de avaliar a influência de tópicos de notícias sobre os preços de ativos. Adotando a estratégia de identificação de Larsen e Thorsrud (2019), incorporamos esses índices em um modelo VAR estrutural para diferenciar entre choques de notícias e de ruído e analisar seus efeitos sobre variáveis macroeconômicas. Nossos resultados revelam que os choques de notícias, capturados pelos índices de notícias, impactam significativamente tanto os preços de ativos quanto uma gama de indicadores macroeconômicos. Tanto os choques de notícias quanto de ruído são cruciais para explicar uma considerável proporção da variação nos preços dos ativos em períodos de curto e longo prazo, sublinhando o papel vital da informação das notícias na dinâmica do mercado. Além disso, nossas descobertas, quando contrastadas com estudos anteriores como Beaudry e Portier (2006), afirmam a eficácia do uso de índices de notícias para identificar choques de notícias, em oposição à dependência exclusiva dos preços de ativos. Esta dissertação contribui para o campo da literatura econômica ao destacar a importância da análise de dados textuais no entendimento dos ciclos econômicos e ressaltar o potencial e a importância das notícias como recurso para fornecer uma visão abrangente da economia.

Palavras-chave: notícias, dados textuais, Latent Dirichlet Allocation, ciclos econômicos

JEL: C8, C55, E32

LIST OF FIGURES

LIST OF TABLES

SUMMARY

1 INTRODUCTION

For macroeconomists and econometricians, one of the paramount tasks is to measure the state of the economy accurately. Over recent decades, advancements in econometric techniques have paved the way for improved nowcasting and forecasting models. These models are critical for anticipating economic cycles—identifying the onset of booms and periods of pessimism. Investors and policymakers, in particular, rely heavily on various indicators to make informed decisions. Yet, it is not solely the indicators crafted by survey companies or government agencies that matter. In an economy marked by complexity, all available information becomes vital for economic agents.

The economy operates in cycles, with periods of growth and contraction dictating the rhythm of economic progress and retreat. The adverse effects of recessions—layoffs, business closures, increased poverty, and heightened inequality—underscore the urgency for reliable predictive models. Typically, investors and policymakers turn to traditional indicators such as GDP, inflation, and unemployment rates to gauge economic health. However, the most telling signs of economic fluctuations may be reflected in higher-frequency sources before they appear in official data.

One such high-frequency source is news media, which can offer real-time insights into the state of the economy. For the purposes of this dissertation, we focus on textual data, such as that found in newspaper articles. This form of data is not only more accessible but also offers comprehensive coverage of a wide range of topics, including international trade, politics, and economics. It forms part of the information set that economic agents use to shape their expectations and decisions.

News media, with its frequent updates and broad coverage, serves as a significant and valuable means to capture what drives the business cycle. It provides narratives and insights that are often challenging to quantify with traditional indicators, capturing the nuances of economic developments that numbers alone may miss (Baker et al.,2016). The central hypothesis of this paper is that news serves as a comprehensive resource for economic agents to grasp the broader economic landscape. This perspective aligns with the views presented by Larsen et al. (2021), who contend that news media is integral to societal functioning, acting as the principal conduit of information.

The application of textual data in economic analysis has garnered significant attention, with research increasingly turning to text mining techniques to both interpret and forecast economic phenomena. For instance, García (2012) investigated the sentiment expressed in New York Times news articles and its influence on asset prices during recessions, discovering that the sentiment captured in the news could help in predicting asset returns, particularly during economic downturns. Tetlock et al. (2008) employed a dictionary-based method to demonstrate how the prevalence of negative words in financial news could be indicative of future stock performance. Ellingsen et al. (2020) specifically add to the body of forecasting literature, illustrating how textual analysis can enhance the prediction accuracy for critical economic indicators such as

GDP, inflation, and unemployment.

The central bank communication theme is another area where textual data analysis has gained prominence. Hansen and McMahon (2016) combined dictionary methods with topic modeling to decipher the messages central banks convey to markets and the public, exploring the influence on macroeconomic and financial variables. Hansen et al. (2018) applied Latent Dirichlet Allocation (LDA) to study the effects of transparency in the Federal Open Market Committee's communications. Similarly, Apel and Grimaldi (2012) analyzed the sentiment and tone of Swedish central bank minutes to predict policy rate decisions.

In this body of literature, we see text mining's growing influence in dissecting the impact of news on real economies and market dynamics. Techniques like those used by Thorsrud (2018) to create a business cycle index and the investigations by Larsen and Thorsrud (2019) into news topics and their predictive power on economic fluctuations highlight this trend.

In this dissertation, the main contribution is the development of two monthly news indices inspired by Larsen and Thorsrud (2019), applied to Brazilian economic news from *Valor Econômico*. Using Latent Dirichlet Allocation (LDA), we organized a vast text archive of over 678,616 articles from July 2011 to December 2022. LDA allows us to transform this high-dimensional, unstructured data into discernible topics that reflect the content of the news, covering various economic narratives. If a specific topics predominates in a given period, it suggests its significance to the current and future economic landscape. Upon examining our results, we began by analyzing the output of the Latent Dirichlet Allocation (LDA), where we successfully identified 40 topics. These topics intuitively reflected a range of distinctive personalities and themes common in Brazilian news.

With the LDA structure, we developed two news indices: the News Index with Time-Varying Parameter (NITVP) and the News Index with Latent Threshold Model (NILTM). The former is elaborated with a variation from Larsen and Thorsrud (2019), utilizing a Time-Varying Parameter model instead of the Latent Threshold Model, which inspired our second index. With this foundation, we then developed structural vector autoregressive (SVAR) models to assess the impact of our indices on Brazilian economic variables, aiming to validate the newspaper as a crucial informant in economic development.

Following this, we applied the identification strategy of Larsen and Thorsrud (2019) within a structural VAR framework to assess the impact of news and noise shocks on key macroeconomic indicators. Our analysis revealed that these shocks, particularly when analyzed using the NILTM index, significantly influence the economy. This was especially noticeable in their contribution to the variance in asset prices, both in the short and long term. Additionally, these shocks demonstrated a considerable impact on various other macroeconomic variables, such as GDP proxy indicators and interest rates, over extended periods. Furthermore, we incorporated our news indices into a framework similar to that of Beaudry and Portier (2006), enabling a comparative analysis of identifying news shocks using both asset prices and the news indices. The results from this comparison suggest that the news indices offer a more effective alternative for capturing the influence of news on economic variables compared to asset prices.

The structure of this dissertation is methodically laid out to guide the reader through our comprehensive study. Chapter 2 presents an in-depth look at the data utilized in this research, offering a clear understanding of its scope and nature. Chapter 3 delves into the mechanics of the LDA model, explaining how this technique transforms raw textual data into meaningful economic indicators. In Chapter 4, we explore the construction of the news index, detailing the processes involved in turning qualitative news content into a quantitative economic analysis tool. Finally, Chapter 5 presents the results of our analysis, showcasing the predictive power of the news index on various aspects of the Brazilian economy.

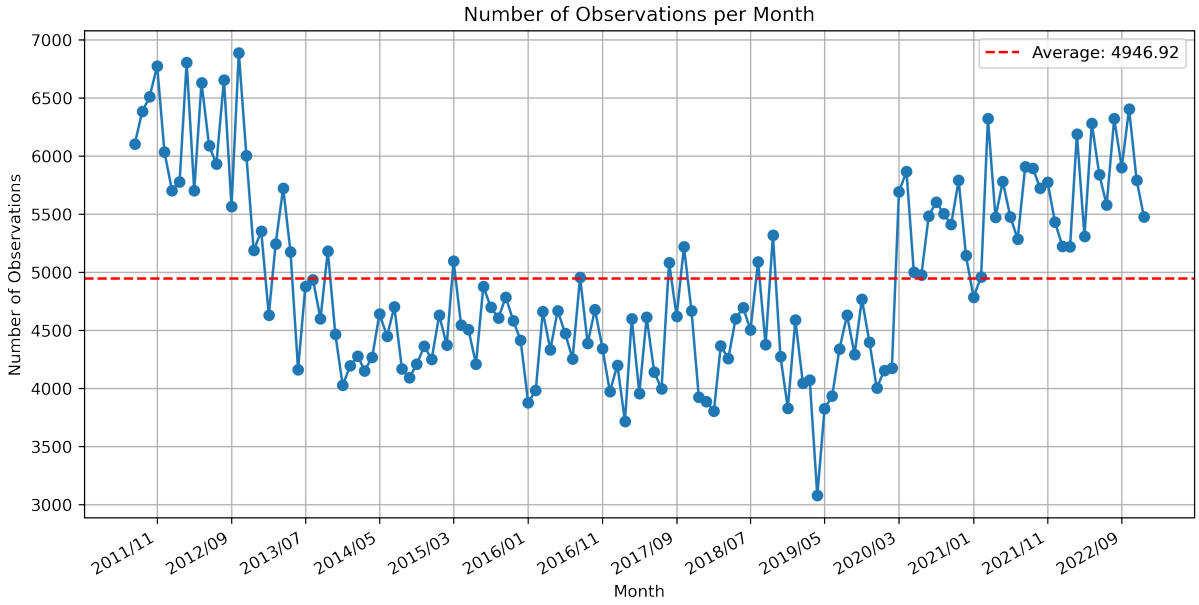
2 DATA

2.1 Textual Data

The initial phase of this research is centered around the collection of daily news articles. The data are textual, unstructured, and high-dimensional. In Brazil, there are several significant newspapers such as *'Estadão'*, *'Folha de São Paulo'*, *'O Globo'*, and *'Valor Econômico'*, which are rich sources of information[1]. Our interest lies particularly in economic news, hence our focus on *'Valor Econômico'*, a renowned economic newspaper in Brazil. This journal is available both in print and online, but our study concentrates on the online version.

Our dataset was sourced from the *Valor Econômico* website[2], where we gathered news articles and opinion columns. Using a web-scraping algorithm, we collected the titles, publication dates and times, full texts, and URLs of 678,616 articles published between July 25, 2011, and December 31, 2022[3]. Figure 1 provides a visual representation of the temporal distribution of article publications, categorizing the data by month.

Figure 1 – The Monthly Frequency of Published Articles



Note: Observations from July 2011 were excluded from this graphic due to the incomplete monthly dataset, which comprised only 888 articles.

The collection of articles encompasses a substantial volume of text, totaling over three hundred million words, averaging 471 words per article with approximately 4,946 articles published each month, not considering those from July 2011. This vast corpus presents computational challenges for statistical analysis. Nonetheless, we streamlined the dataset by focusing solely

[1] Available at: <https://www.poder360.com.br/economia/jornais-em-2021-impresso-cai-13-digital-sobe-6/>. Accessed on: October 19, 2023.

[2] Available at: <https://valor.globo.com/>. Acessed on: October 19, 2023.

[3] In total, 2.2 GB of raw textual data was extracted.

on the titles and introductory paragraphs of each article[4]. We operate under the assumption that these elements encapsulate the most critical content of the news. This strategic reduction significantly enhances computational efficiency for subsequent model applications. Following this modification, the average word count per document has been reduced to around 57 words.

2.2 Preprocess and Cleaning

After condensing the full articles to just the titles and the first paragraph, our next step was to further preprocess the textual data, aiming to decrease its dimensionality as is customary in NLP research.

Initially, we segmented the news text into sentences. Following this, we removed punctuation, special characters, and excessive whitespace, and standardized all text to lowercase. Subsequently, we tokenized these sentences into individual words.

To further reduce the text's dimensionality, we eliminated a predefined list of stop words—words that do not contribute to a meaningful interpretation or significance to the news content. Examples include articles, prepositions, and other common words such as *'de', 'a', 'o', 'em', 'com', 'foi', 'ele'* [5]. Given the presence of English-language articles within the dataset, a corresponding list of English stop words was also employed[6]. In addition, we removed numerical data and geographical locators, specifically city names associated with the publication of the news. This entailed a targeted removal process where city names were excluded if they were positioned as the initial word in the first paragraph. For illustrative purposes, a singular example of the data cleaning and preprocessing steps is presented in Appendix A. This example outlines the sequence of actions taken to transform the raw data into a format suitable for analysis.

Through these steps, we achieved a reduction to an average of 28 words per document. However, at this stage, our analysis was limited to individual words, ignoring the potential significance of word pairs. Recognizing this, we constructed a bi-gram model to capture meaningful combinations like *"banco central"* (central bank), where the joint term holds a distinct meaning from the individual words *"banco"* (bank) and *"central"* (central).

In the final phase of cleaning our dataset, we filtered out terms that appeared in fewer than 0.1% of the articles, thereby focusing on vocabulary that bears more weight across the corpus. This approach ensures our base is primed for computational analyses. After the text data was cleansed, we took two critical steps in preparation for model estimation. First, we created a dictionary to map out the vocabulary present in our documents. Subsequently, we constructed the *corpus* — a term-document frequency matrix to represent the distribution of terms across the documents.[7]

---

[4] We consider each news item as a combination of its title and the first paragraph, concatenating these elements to form a unified text block for analysis.

[5] The Portuguese stop-words translate to 'of/from', 'to/the', 'the', 'in/on', 'with', 'was/went', and 'he', respectively.

[6] Despite being a Brazilian-language news platform, English-language articles were also identified within the *Valor Econômico* domain.

[7] A formal definition of *corpus* is provided in Chapter 3.

2.3 Hard Data

In the upcoming sections, we will detail the creation of two monthly news indices using our textual data. Alongside these indices, we will also utilize traditional structured data, collected from various sources on a monthly basis.

From the Central Bank of Brazil, we have sourced several key economic indicators: the interest rate, and the Central Bank Economic Activity Index (IBC-BR). The interest rate here refers to the monthly capitalization of the daily Selic rate. The IBC-BR indicator, normalized to 100 in 2002, is seasonally adjusted by the source.

Data from the *'Instituto Brasileiro de Geografia e Estatística'* (IBGE) include inflation (IPCA), the industrial production indicator (IPI), the number of employed individuals, average income, and the Nominal Sales Revenue Index in retail trade. The IPCA inflation series, a consumer price index for all items, is normalized to 100 as of December 1993. The Production Indicator is standardized to 100 as of June 2019. Employment numbers and average incomes are derived from the *'Pesquisa Nacional por Amostra de Domicílios Contínua'* (PNADC), with respective codes 6320 and 6387. Retail sales data come from the *'Pesquisa Mensal de Comércio'* (PMC) under code 8882. All these series are seasonally adjusted using the X-13 ARIMA-SEATS package developed by the United States Census Bureau.

Asset prices, specifically the IBOVESPA (IBOV) index, were obtained from Yahoo Finance. The IBOV, Brazil's benchmark stock index, comprises a dynamic mix of about 60 to 70 Brazilian companies, reflecting a broad spectrum of the Brazilian economy.

Additionally, we sourced another Gross Domestic Product (GDP) proxy from the *'Instituto Brasileiro de Economia'* (FGV-IBRE), known as the *'Monitor do PIB'* (GDP-M). This series too has been seasonally adjusted using the X-13 ARIMA-SEATS method.

We have developed two distinct monthly Total Factor Productivity (TFP) measures: one adjusted and the other unadjusted for capital utilization. Our methodology relies on *'Monitor do PIB'* (GDP-M) as our output measure (Y), alongside the number of workers (L) to represent labor, and capital stock (K) for capital input. For the monthly capital stock, we refer to the series estimated by Júnior and Cornelio (2020). To account for capital utilization (u), we use the indicator published by the *'Confederação Nacional da Indústria'* (CNI).

The two TFP measures are formulated as follows:

1. Unadjusted for capital utilization: $TFP = log\left(\frac{Y_t}{K_t^{\alpha} L_t^{1-\alpha}}\right)$

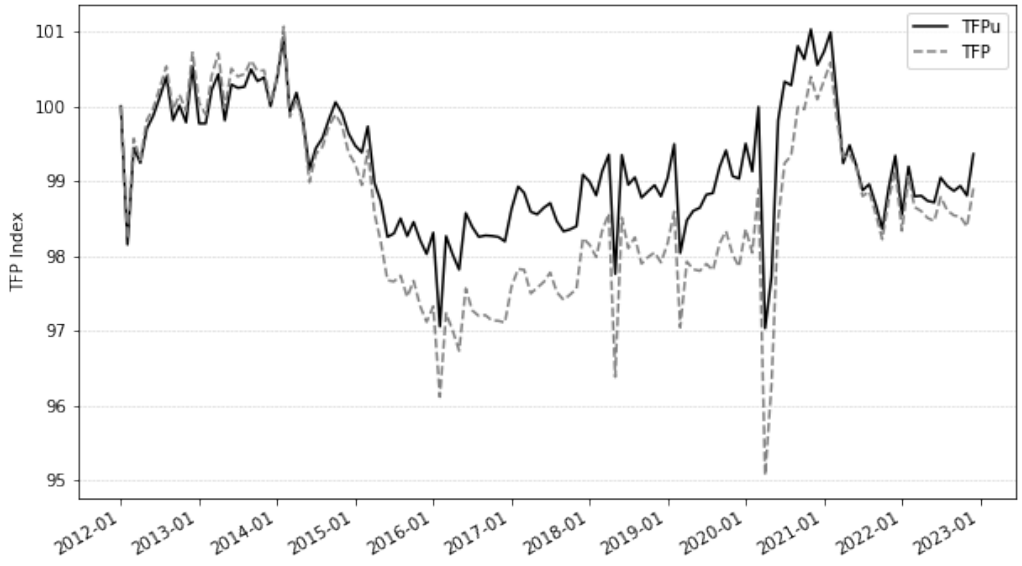
2. Adjusted for capital utilization: $TFPu = log\left(\frac{Y_t}{(u_t K_t)^{\alpha} L_t^{1-\alpha}}\right)$

In both formulas, α represents the output elasticity of capital. Following the methodology of Gomes et al. (2003) applied to the Brazilian economy, we set α at 0.4 for both measures.

The TFP measures cover the period from January 2012 to December 2022. This timeframe is chosen as it aligns with the availability of data from the PNADC, which started in January

2012. Figure 2 presents both measures constructed for TFP, with the indices normalized to a baseline value of 100 in January 2012.

Figure 2 – Total Factor Productivity Index (January 2012 = 100)



In the next chapter, we will turn our attention to the textual data. We will discuss how to convert the rich narrative content of text into interpretable data using Latent Dirichlet Allocation (LDA). The chapter will cover the essentials of the model's estimation and demonstrate how it can be used to create a structured monthly series from our textual data. This step is crucial as it forms the foundation for subsequent analysis, allowing us to translate the vast amount of unstructured information into meaningful economic insights.

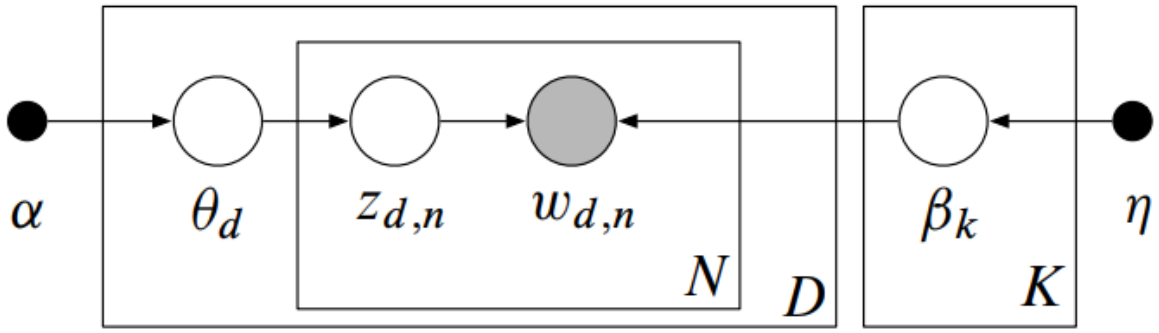
3 LATENT DIRICHLET ALLOCATION

To interpret the words collected from the news, we will use a topic model, Latent Dirichlet Allocation (LDA), to reduce the dimensionality of the text data into something interpretable. The LDA model is an unsupervised probabilistic topic modeling[1] and is one of the most popular topic models in the Natural Language Processing (NLP) literature, introduced by Blei et al. (2003).

The idea behind LDA is straightforward: documents are viewed as mixtures of topics, and these topics are treated as mixtures of words. Consequently, LDA's goal is to provide a concise description by uncovering topics from a collection of documents. Following the approach of Blei et al. (2003), we can describe LDA using the subsequent notation:

- A *corpus* is a collection of M documents denoted by D = $(\mathbf{w_1}, \mathbf{w_2}, ..., \mathbf{w_M})$.
- A *document* is a sequence of N words denoted by **w** = $(w_1, w_2, ..., w_n)$, where w_n represents the n^{th} word in the sequence.
- A *word* is an item from a vocabulary indexed by $1, ..., V$. V denotes the number of unique words across the entire sample.
- The *topic* k is a distribution over a fixed vocabulary for each $k \in \{1, ..., K\}$, with K representing the number of topics.

Figure 3 – LDA: Graphical Model



Note: Figure available at: <https://www.cs.columbia.edu/~blei/talks/Blei_VI_tutorial.pdf>. Acessed on: November 21, 2023.

Figure 3 represents LDA as a graphical model. The hyper-parameters α and η control the proportion parameter for θ_d and the topic parameter for β_k, respectively, where $d \in$ {1,...,D}

[1] In machine learning, an unsupervised model refers to an algorithm that is trained on data without prior labeled information. Instead, these models discover patterns and relationships directly from the training data.

and $k \in \{1,..., K\}$. The θ_d represents the per-document topic proportions for each document d[2], and β_k stands for the distribution over the vocabulary for each topic k[3]. Both θ_d and β_k follow a Dirichlet distribution with prior parameters α and η, respectively[4]. The variable $z_{d,n}$ assigns a topic to the n^{th} word in document d. The observed words for document d are represented by w_d, where $w_{d,n}$ is the n^{th} word in that document. It's worth noting that the arrows, for instance, indicate that the per-word topic assignment depends on the per-document topic proportions.

A crucial aspect to understand is that a document can encompass multiple topics, and each document possesses its unique per-document topic proportion (θ_d). Every article shares the same set, K, of potential topics but exhibits these topics in varying proportions.

With this structure, the joint distribution of latent and observed variables is:

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D}) = \prod_{i=1}^{K} p(\beta_i) \prod_{d=1}^{D} p(\theta_d) \prod_{n=1}^{N} p(z_{d,n}|\theta_d) p(w_{d,n}|\beta_{1:K}, z_{d,n}) \tag{3.1}$$

To solve this inferential problem, we apply Bayes' Theorem to compute the posterior distribution of the hidden variables, given the observed documents[5].

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D}|w_{1:D}) = \frac{p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D})}{p(w_{1:D})} \tag{3.2}$$

To adress the inference challenge in the LDA model, we utilize the Online Variational Inference method developed by Hoffman et al. (2010) to approximate the posterior distribution. This method reinterprets the inference task as an optimization problem rather than a sampling one. As highlighted by Assuncion et al. (2009), online LDA provides a more efficient approach for analyzing large data collections.

As previously discussed, to fit the model, three input parameters are required, given the observed words from our collection of documents: α, η, and K. Each of these parameters is a single value. The prior specification for α and η follows the proposal by Griffiths and Steyvers (2004):

$$\alpha = \frac{50}{K} \tag{3.3}$$

$$\eta = \frac{200}{V} \tag{3.4}$$

Where V represents the unique words and K denotes the number of topics. The authors suggest fixing α and η and then examining the implications of varying the number of topics,

[2] At the document level, we have θ_d as a vector of dimension $1 \times K$, where each column represents the proportion of topic k in that document d. We can interpret θ as a matrix of dimension $M \times K$.

[3] At vocabulary level, we represent β_k as a vector with dimensions $1 \times V$, where each column represents the proportion of a specific word from the vocabulary in topic k. We can view β as a matrix of dimension $K \times V$.

[4] In essence, the intuition behind the parameters is as follows: A higher value of α suggests that a document is likely to contain a mixture of many topics. Analogously, a larger value of η indicates that each topic is likely to be composed of a broader set of words.

[5] The numerator is the joint distribution defined in Equation (3.1), which can be computed. The denominator is the marginal probability of the observed data; however, it is intractable.

K. Given α and η, our challenge is to determine the appropriate value for K. To evaluate the model, we create a grid of topics ranging from 10, increasing by increments of 10, up to 80, and compute the perplexity score at each increment. Blei et al. (2003) proposed using the perplexity score to gauge the model's quality. The perplexity score is designed to measure a model's capability to make predictions for documents that were not included in its training set. A lower perplexity score signifies enhanced model performance. When subjected to evaluation on a test set comprising M_{test} documents, the perplexity score is calculated as:

$$\text{perplexity}(D_{\text{test}}) = \exp\left\{-\frac{\sum_{d=1}^{M_{\text{test}}} \log p(\mathbf{w}_d)}{\sum_{d=1}^{M_{\text{test}}} N_d}\right\} \tag{3.5}$$

where M_{test} represents the size of the test corpus, specifically, the number of documents within the test corpus. N_d denotes the size of the d^{th} document, referring to the number of words in that document d. Meanwhile, $p(\mathbf{w}_d)$ indicates the probability that the set of word $\mathbf{w}_d$ is generated within the document.

Table 1 – Perplexity Scores for Different Number of Topics*

Topics (K)	10	20	30	40
Perplexity Score	2654.19	2541.57	2536.77	**2462.39**
Topics (K)	50	60	70	80
Perplexity Score	2477.07	2508.22	2510.43	2533.45

* The perplexity scores are computed using the α and η priors from Griffiths and Steyvers (2004) for each value of K.

We use approximately the first 80% of the documents from the data as the training set and the remaining 20% as the test set. The training set contains 542,248 articles with news up to December 31, 2020.[6] The test set comprises 136,368 articles. Referring to Table 1, we observed that the number of topics, denoted as $\mathbf{K = 40}$, yielded the lowest perplexity score, indicating it as the best model for classifying unknown documents outside of the training set

To achieve a more dependable perplexity score, we use the 5-Fold cross-validation technique, which systematically (and randomly) divides the corpus of documents into 5 distinct subsets. For each k, we run 5 separate models. In each model, we use a different subset as the test set, with the remaining four subsets combined to form the training set. This approach ensures that every subset gets a turn as the test set across the five models. The perplexity score for each k is then computed as the mean of the scores from these 5 model configurations. Different from the first analysis, we create a grid of topics ranging from 20, increasing by increments of 20, up to 100.

As illustrated in Table 2, when $\mathbf{K = 60}$, the model achieves the lowest mean perplexity score, indicating a superior predictive capability. However, this advantage is marginal when compared to $\mathbf{K = 40}$. We opted to classify using $\mathbf{K = 40}$ topics. This decision was driven

[6] It's important to note that the training set includes news related to the COVID-19 shock.

Table 2 – 5-Fold Cross Validation: Perplexity Scores for Different Number of Topics

Number of	**Iteration***					**Mean PS****
Topics (K)	1	2	3	4	5	
20	1881.25	1867.72	1869.16	1936.96	1869.60	1884.93
40	1801.16	1825.08	1819.24	1890.85	1822.16	**1831.70**
60	1799.61	1828.84	1798.62	1900.59	1822.35	**1830.00**
80	1879.47	1864.49	1866.37	1927.17	1849.37	1877.37
100	1970.29	1984.40	1997.85	1963.61	1984.38	1980.10

* Each "Iteration" corresponds to the fold used as the test set, e.g., Iteration 1 uses the first fold as the test set, Iteration 2 uses the second fold, and so on.

** "Mean PS" stands for the average perplexity scores for each topic.

by the understanding that having too many topics could lead to the generation of redundant or overly similar topics, which in turn diminishes the distinctiveness and recognizability of individual topics. This observation aligns with the findings of Gan and Qi (2021), who noted potential pitfalls in over-segmenting topics. Therefore, we estimated the LDA model with the following parameters: $K = 40$, $\alpha = 1.25$ and $\eta = 0.055$[7]. We allocated the first 80% of the dataset as the training set and reserved the remaining 20% for the test set.

After determining the number of topics and estimating the model, LDA provides us with the topic-word distribution and the document-topic distribution. However, as we noted earlier, LDA is an unsupervised model, which means it doesn't assign names or labels to the identified topics. The process of assigning labels to the topics is inherently subjective and necessitates our direct involvement. In order to label them accurately, we conduct a manual inspection of the most relevant words linked to each topic. Additionally, we identify the documents that most effectively represent each topic, specifically those that exhibit the highest proportion of topic k within the entire collection. This manual inspection is particularly pertinent when classifying a larger number of topics, such as 40 topics. Through this manual process, we have identified topics that are both relevant and coherent within the Brazilian context. Details of these topics, including their labels, the count of articles most strongly associated with each topic, and the ten most significant words defining each topic, can be found in Appendix B. We observe a range of topics that include global interactions, such as the *United States Economy, Europe,* and *BRICS*; political themes tied to figures and entities like *Lula, Bolsonaro,* and *Brazilian Politics*; macroeconomic subjects including *Monetary Policy, Fiscal Policy,* and *Inflation*; among others.

After utilizing the topic-word distribution to define the labels of the topics, we now leverage the document-topic distribution to construct a time series. Each document is comprised of a blend of the 40 identified topics, represented as proportions that sum to one—indicating a complete representation of the document within these topics.

[7] As defined in equation (3.4), η = 200/V, where V represents the number of unique words following our data preprocessing, as detailed in Chapter 2 and in Appendix A, which includes a detailed example. After this preprocessing, we identified a total of 3,641 unique words.

To capture monthly trends, we aggregate a topic proportion for each month by summing the topic proportions of all documents published within that month. We then normalize these figures by the total sum of topic proportions for the respective month, ensuring that the total remains one[8].

To ensure the stationarity of the series, we apply a logarithmic transformation to the monthly proportions, take the first difference, and then standardize the series. For a detailed visualization of the topic proportions at various time points before and after applying these statistical transformations, refer to Appendix C.

In the forthcoming chapter, we will harness the monthly topic proportions from the LDA to establish a news index. This index will aim to reflect the impact of news topics on asset prices. We will examine Latent Threshold Models (LTM) and Time-Varying Parameter (TVP) regressions to identify the topics that significantly affect the construction of the index. The goal is to develop an index that accurately indicates economic trends.

[8] Documents from July 2011 were included in the LDA estimation model. However, due to the sparse data available for that month (only few days), we have started our topic proportion analysis from the 1st of August, 2011.

4 CONSTRUCTING THE NEWS INDEX

The initial challenge in our study was to transform textual data into a format that could be interpreted and analyzed. This was achieved through the application of LDA, which facilitated the conversion of extensive texts into 40 distinct topic proportion series. However, simply identifying these topics is not sufficient; their true economic value emerges from their ability to capture and forecast economic outcomes. Therefore, we structured our topic proportion series on a monthly basis, aligning them with the need to select a suitable monthly economic indicator for analysis.

In line with Real Business Cycle theory, if agents receive positive news about future economic conditions, optimism increases, leading to higher investment and consumption, and consequently, an economic boom. Following the insights of Beaudry and Portier (2006), we consider stock prices are the type of variable most likely to reflect news. This premise forms the basis of our analysis, where we examine the extent to which the news topics extracted through LDA can forecast movements in the Brazilian stock market, as represented by the IBOVESPA index. We aim to develop an aggregate news index that captures the influence of these topics on Brazilian asset prices, a methodology paralleling that used by Larsen and Thorsrud (2019) in their construction of a news index.

In this section, we will construct two variants of the news index, each subtly different in capturing the temporal relationship between the topics and the IBOVESPA index. The first index, which we refer to as NI TVP, utilizes a Time-Varying Parameter Regression Model approach. The second index, named NI LTM, is based on a Latent Threshold Model.

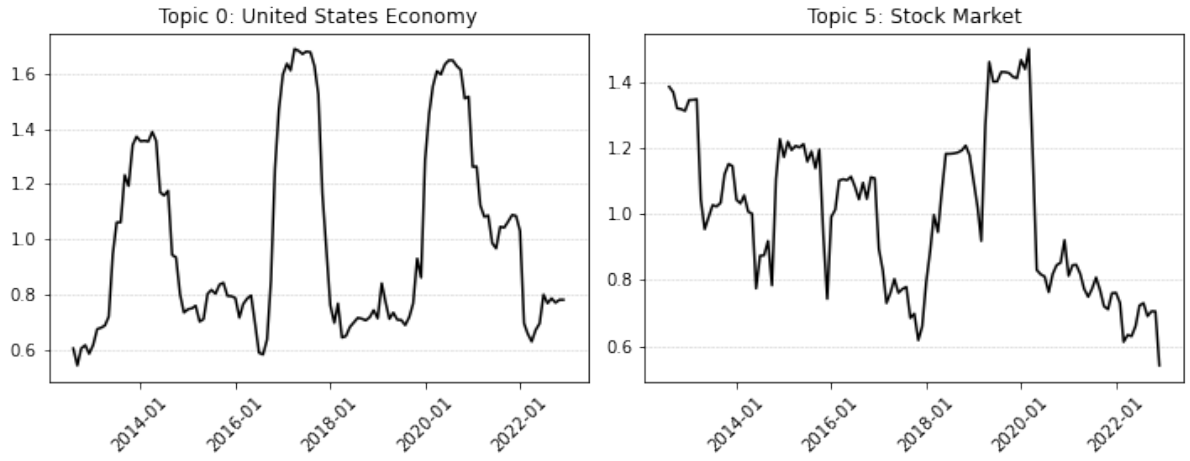
4.1 NI TVP: News Index with Time Varying Parameter

4.1.1 Time Varying Parameter Regression with Stochastic Volatility

We believe that the association between the proportions of various topics and asset prices is not static, but may change over time. To address this, we use Time-Varying Parameter Regression with Stochastic Volatility for each of the 40 topics, with the IBOVESPA index as the dependent variable in each model. This approach requires conducting 40 separate regression models, one for each topic proportion series, to effectively capture the evolving dynamics of their relationships with Brazilian asset prices.

As we investigate the evolution of topic proportions over time, it's clear that these metrics are inherently subject to variations and fluctuations driven by the events reported in the news media. These proportions can be affected by diverse events, which in turn impact the volume and tenor of discussions surrounding these topics. Figure 4 provides a clear illustration of how volatility can shift over time for topics related to the United States and the stock market. Given this variability, we propose that a model with stochastic volatility would be more suitable to capture the nuanced and changing landscape of topic-driven market dynamics.

Figure 4 – Temporal Volatility of Topic Proportions: *US Economy* and *Stock Market*



Note: Standard deviations are calculated using a 12-month rolling window.

The Time Varying Parameter Regression with Stochastic Volatility, introduced by Nakajima (2011), provides a framework for modeling the dynamic relationships in time series data where the influence of predictors on the outcome can change over time, accompanied by changes in volatility. The model can be specified as follows:

$$y_t = z_t' \alpha_t + \epsilon_t \tag{4.1}$$

$$\alpha_{t+1} = \alpha_t + u_t \tag{4.2}$$

$$\sigma_t^2 = \gamma \exp(h_t) \tag{4.3}$$

$$h_{t+1} = \phi h_t + \eta_t \tag{4.4}$$

In the model, ϵ_t is normally distributed with a mean of zero and variance σ_t^2, represented as $\epsilon_t \sim N(0, \sigma_t^2)$. Similarly, u_t is drawn from a normal distribution with mean zero and covariance matrix $\sum$, noted as $u_t \sim N(0, \sum)$, and η_t is also normally distributed with mean zero and variance σ_η^2, denoted by $\eta_t \sim N(0, \sigma_\eta^2)$. Here, t denotes the time index, defined as $t = 1, ..., n$. The term z_t is a $(n \times 1)$ vector corresponding to a specific topic, while α_t is an $(n \times 1)$ vector of time-varying parameters associated with that topic. We assume that the baseline level of α is zero ($\alpha = 0$), the initial state of u is drawn from a normal distribution with zero mean and covariance matrix $\sum_0$ ($u_o \sim N(0, \sum_0)$), γ is positive ($\gamma > 0$), and the initial value of h is zero ($h_0 = 0$). The prior of ϕ is chosen such that it satisfies the condition $|\phi| < 1$ with the initial values set at 0.95.

The model estimation employs a Markov Chain Monte Carlo (MCMC) approach, from which we draw 3000 samples. The burn-in period consists of the first 300 samples, which are discarded to allow the Markov chain to reach its stationary distribution. The priors adopted are consistent with those used by Nakajima (2011)[1]. The dataset for estimation extends from

[1] For a more comprehensive explanation of the model's structure and estimation process, readers are referred to Nakajima (2011).

September 2011 to December 2022. The IBOV index has been processed with the same methods used for the topic series data: applying logarithmic differencing and then standardization.

4.1.2 Dynamic Bayesian Predictive Synthesis

We aim to investigate the predictive capacity of topic time series for the IBOV index. Given the extensive array of topic time series available, it is crucial to evaluate and condense this data set. It is reasonable to presume that certain topics may exert more influence on asset prices than others. One approach to address this issue is to incorporate our topic time series into a Time-Varying Parameter Regression (TVPR) model to predict our target series. The objective is to determine the individual predictive strength of each series concerning the target series.

To this end, we construct 40 distinct forecasting models, each incorporating a single topic time series, to predict asset prices. Drawing from the concept introduced by McAlinn and West (2019), we envision a scenario where the decision-maker seeks to forecast asset prices and receives forecast densities from various sources. In this context, each topic time series represents a distinct source, providing its own forecast to the decision-maker.

In practical terms, we execute a TVPR on individual topic series to project the IBOV index. Our goal is to predict asset prices using the implied posterior $P(y|M)$, which is the predictive probability distribution for a future quantity y based on each model M:

$$p(y|M_i) = \int p(y|\theta^i, M_i)p(\theta^i|M_i)d\theta^i \tag{4.5}$$

In this equation, θ^i represent the parameters of model i, and the integral captures the total predictive distribution by averaging over all possible values of the parameters, weighted by their posterior probability. Specifically, we designate Model M_i as a TVPR that employs topic time series i as the explanatory variable to predict the IBOV index. Each model is constructed to assess the unique contribution of its respective topic to asset price movements. This formulation allows us to assess the predictive power of each topic time series individually.

This process begins with the period from September 2011 to February 2017 as the training phase, during which we implement forward filtering with one-step-ahead predictions. Following this, from March 2017 to March 2019, we perform MCMC-based Bayesian Predictive Synthesis (BPS) analysis, utilizing an 'expanding window' of past data as we move forward in time. This approach extends to the final phase, covering the period from April 2019 to December 2022. In our analysis, we extend the evaluation of each topic series predictive power by comparing their density forecasts within the BPS framework.

4.1.3 Calculating the NI TVP Index

To derive the monthly news index, we integrate the predictive outcomes delineated in the preceding chapters. Our methodology echoes that of Larsen and Thorsrud (2019), with a key variation; we employ a TVPR model instead of a Latent Threshold Model. The formula for constructing the index is as follows:

$$NITVP_t = \sum_{i=1}^{K=40} w_i \alpha_{i,t} n_{i,t} \tag{4.6}$$

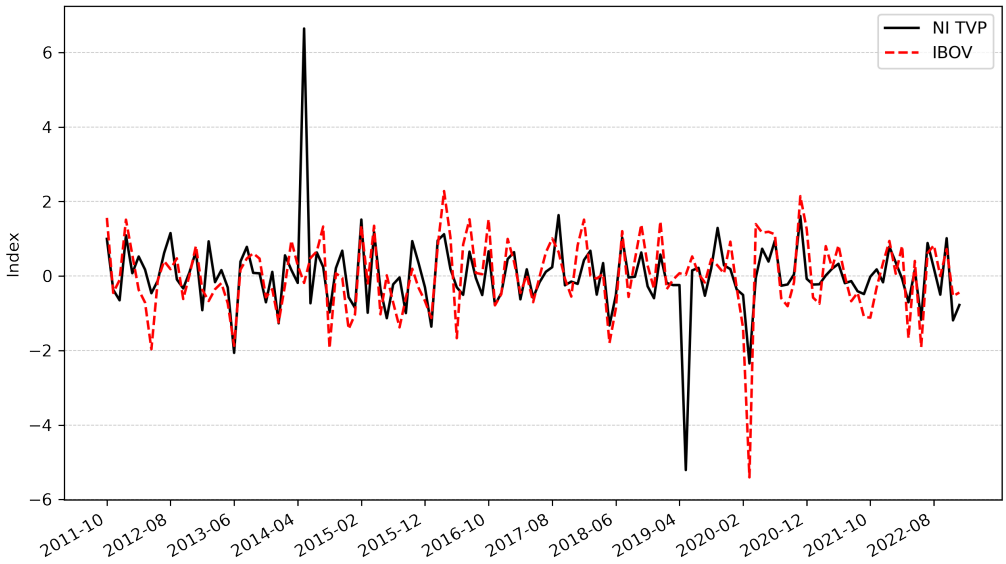
In this equation, $n_{i,t}$ represents the activity of topic i at the simultaneous time t, as identified by the LDA model. The term $\alpha_{i,t}$ denotes the estimated parameter from the TVPR for topic i at time t, and w_i signifies the weight corresponding to the predictive influence of topic i on asset price forecasting.

The weight w_i is calculated using the formula:

$$w_i = \frac{p(y|M_i)}{\sum_{i=1}^{K=40} p(y|M_i)} \tag{4.7}$$

where $p(y|M_i)$ indicates the predictive density of future values y based on the model M_i, while the denominator aggregates these densities across all K topics to normalize the weight of each topic within the overall index. Figure 5 displays the NITVP, as constructed in this section, alongside the IBOV returns from October 2011 to December 2022.

Figure 5 – NI TVP and IBOV Returns



Note: All series are standardized.

4.2 NI LTM: News Index with Latent Threshold Model

The NI LTM index closely resembles the NI TVP index, with a key difference in the treatment of the α term. The BPS process, however, remains consistent with the approach used in the NI TVP index, utilizing the same weighting scheme derived from the initial index.

4.2.1 Latent Threshold Model

Introduced by Nakajima and West (2013), the Latent Threshold Model (LTM) provides a selective mechanism for dynamic variables. The model is articulated as follows:

$$y_t = x_t' b_t + \epsilon_t \tag{4.8}$$

$$b_t = \beta_t s_t \tag{4.9}$$

$$\beta_{t+1} = \mu + \phi(\beta_t - \mu) + \eta_t \tag{4.10}$$

where $s_t = I(|\beta_t| \geq d)$ defines the latent threshold, with d indicating the threshold level and I representing the indicator function. This threshold signifies a temporal selection for the β coefficients; specifically, when $|\beta_t|$ exceeds the threshold, it suggests that the topic is influential in predicting y_t at time t.

Similar to the approach in Section 4.1.1, we execute 40 distinct regression models, one for each topic series. In each model, x_t represents a different topic time series. The unique aspect here is the inclusion of a threshold, which dictates whether a topic at a specific time contributes information to y_t.

In this model, t represents the time index, x_t is a $(n \times 1)$ vector and b_t is a $(n \times 1)$ vector, where n represents the number of observations over time. As per Nakajima and West (2013), the model stipulates that ϵ_t follows a normal distribution with mean zero and variance σ^2 ($\epsilon_t \sim N(0, \sigma^2)$), and η_t also adheres to a normal distribution with mean zero and variance σ_η^2 ($\eta_t \sim N(0, \sigma_\eta^2)$). The prior distributions are set as follows: $\mu \sim N(0, 1)$, $\frac{\phi+1}{2} \sim Beta(20, 1.5)$, $\sigma_\eta^{-2} \sim Gamma(3, 0.03)$, and $\sigma^{-2} \sim Gamma(3, 0.03)$. The priors for ϕ have a mean and standard deviation of (0.86, 0.11); for σ_η^2 and σ^2, they are (0.015,0.015)[2]. The model is estimated using Gibbs Sampling with 3000 iterations, discarding the first 300 as a burn-in period.

4.2.2 Calculating the NI LTM Index

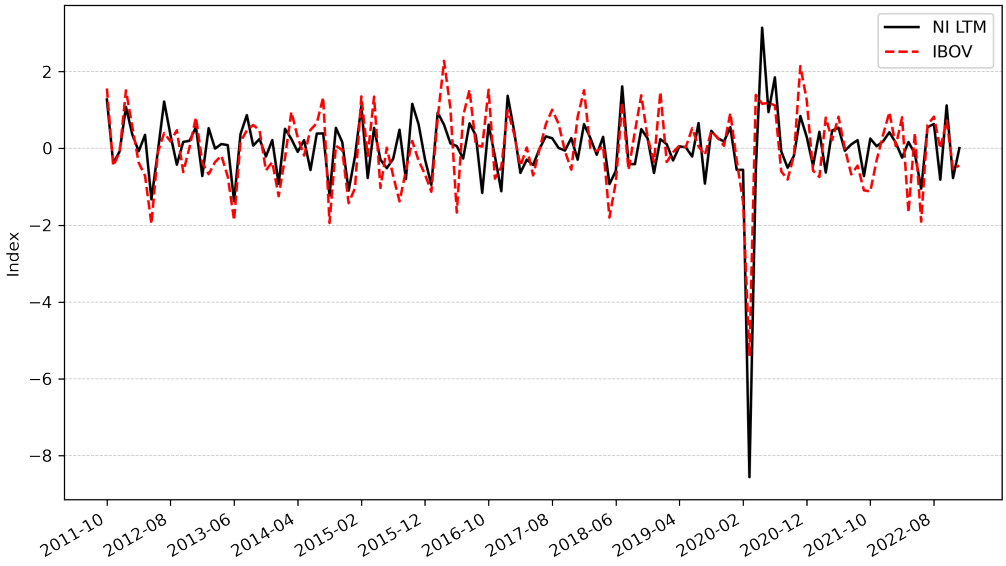
The process of deriving the monthly NI LTM news index closely mirrors the methodology employed for the first index. We utilize the previously defined weights and incorporate topic activities as delineated by the LDA. The formula for constructing the index is as follows:

$$NILTM_t = \sum_{i=1}^{K=40} w_i b_{i,t} n_{i,t} \tag{4.11}$$

In this equation, $n_{i,t}$ represents the activity level of topic i at the concurrent time t, as determined by the LDA model. The term $b_{i,t}$ indicates the parameter estimated by the LTM for topic i at time t, while w_i denotes the weight assigned to the predictive influence of topic i on asset price forecasting, as established in Section 4.1.2.

[2] For a detailed exposition of the model and its assumptions, refer to Nakajima and West (2013).

Figure 6 – NI LTM and IBOV Returns



Note: All series are standardized.

Figure 5 compares the NITVP with IBOV returns, while Figure 6 presents a comparison of NILTM against IBOV returns from October 2011 to December 2022. Both indices exhibit similarities with IBOV returns; however, they do not match exactly. NITVP shows a correlation of 0.49 with IBOV returns, and NILTM demonstrates a stronger correlation, at 0.71.

In the following chapter, we will utilize the two news indices we have developed to explore their effects on macroeconomic variables following news shocks. This analysis will delve into distinguishing between news and noise shocks using Structural VAR models, enhancing our understanding of how these indices influence economic dynamics.

5 RESULTS

In this section, our objective is to quantify the impact of news shocks on the Brazilian business cycle, employing a structural Bayesian VAR (Vector Autoregression) model. Our analysis adopt the identification strategy proposed by Larsen and Thorsrud (2019) for distinguishing between news and noise shocks. This involves incorporating our news index and asset prices into a recursive ordering based on Cholesky identification within the VAR framework. Accordingly, the variables productivity, the news index, and asset prices are sequenced in the VAR system in this specific order.

The logic behind this identification is that the news index is presumed to be independent of immediate productivity changes. A news shock is expected to influence productivity after a lag. The news index serves as a gauge of information that affects the expectations and decisions of economic agents. It offers a refined view of the prevailing economic narratives, focusing on the most critical information that reflects the state of the economy.

In our model, asset prices are positioned following the news index. This ordering is based on the understanding that asset prices represent a wider array of information beyond what the news index captures. Diverging from Larsen and Thorsrud (2019) approach, our news index is designed to be contemporaneous. While their model was based on the predictive power of news topics at time t-1 for asset prices at time t, our index proposes a simultaneous impact, reflecting the rapid processing and immediate effect of news in the market. This index likely mirrors market sentiment, capturing key information at the time of trading. Therefore, asset prices in our VAR ordering are positioned after the news index, acknowledging that they are influenced by the news but also react to a range of other information not encompassed by the news index alone. Consequently, following the approach of Larsen and Thorsrud (2019), we define noise shocks as variations in asset prices that are not explained by the information conveyed in the news.

5.1 Estimation

The time series for our news indices, NITVP and NILTM[1], were constructed from September 2011 to December 2022. However, for our empirical application, we utilize the estimation sample from January 2012 to December 2022, because our productivity series start em January. We incorporate both measures of Total Factor Productivity, TFP and TFPu, both of which are constructed using logarithmic transformations. Asset prices, represented by the IBOV index, are measured as monthly changes, i.e., log(x_t) - log(x_{t-1}).

Given the unavailability of Gross Domestic Product data on a monthly basis, we employ two alternative output measures: *'Monitor do PIB'* (GDP-M) and Central Bank Economic Activity Index (IBC-BR). Additionally, we include the retail index (RI), inflation (π) and the interest rate

[1] The news index series will be utilized as presented in Figures 5 and 6.

(R) in our analysis. The GDP-M, IBC, and RI are measured in log levels, while π is captured in monthly changes, and R is presented in level form (as a percentage).

The model was estimated using the BEAR toolbox program (Dieppe et al. (2016)), applying a Normal Wishart prior. The simulation process encompassed 3000 iterations, which included a burn-in period of 1000 iterations. The model was structured with a configuration of 4 lags[2].

5.2 News vs Noise Shocks

We incorporate a benchmark model similar to that of Larsen and Thorsrud (2019), comprising the following variables in sequence: TFPu, NILTM, IBOV, GDP-M, π (inflation), and R (interest rate). Our model, however, focuses on output indicators rather than consumption.

Upon a news shock, as illustrated in Figure 7, Productivity (TFPu) initially experiences a significant rise, indicating an immediate boost in productivity expectations. This increase is fleeting, as TFPu eventually reverts to its baseline, demonstrating the news' temporary effect on productivity. The IBOV index shows an initial sharp increase, reflecting a positive market response to the shock, but this surge is short-lived. Around the sixth month, the index stabilizes near its original level. GDP-M initially rises in response to the shock but gradually falls back towards the baseline. Inflation (π) initially surges, suggesting a spike in price pressures, but this is followed by a decrease. The interest rate (R) initially dips, indicating a brief period of easing, before rising again, signaling a subsequent tightening.

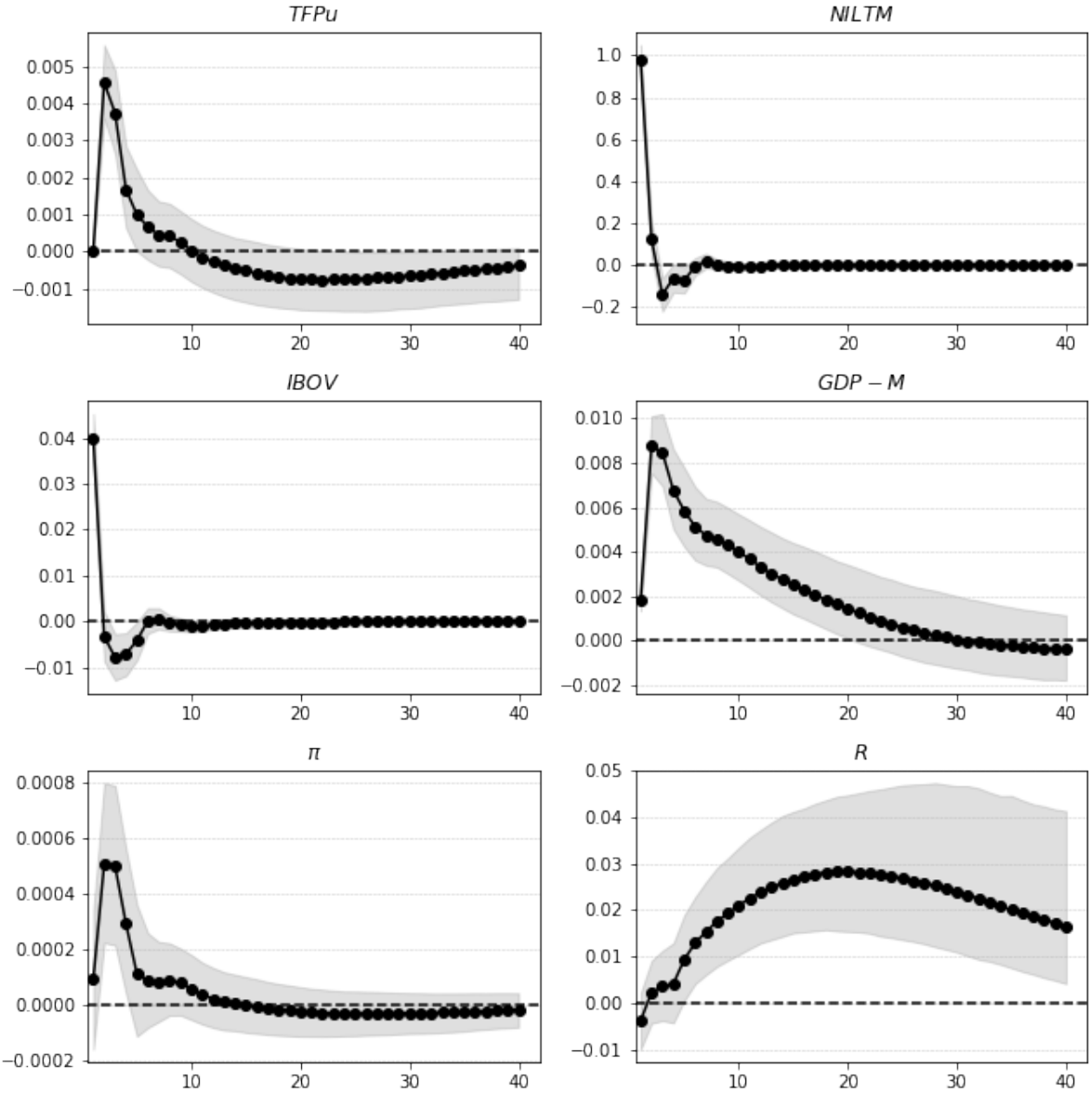
In contrast, as shown in Figure 8, a noise shock leads to an opposite response in Productivity, GDP-M, Inflation, and the Interest Rate compared to a news shock. For instance, TFPu and Inflation exhibit a slight decline before stabilizing. The responses of productivity and the News Index are not significantly different from zero, indicating that our identification methods effectively separate news and noise components, akin to Larsen and Thorsrud (2019). As expected for noise shocks, which are unassociated with economic fundamentals, we observe no significant effects on these variables.

Consistent with certain aspects of Barsky and Sims (2011), we observe a significant rise in stock prices following news shocks. However, this contrasts with their findings, as we see an increase in GDP-M (output) and inflation, whereas Barsky and Sims (2011) report output decline and deflation. Additionally, our model shows a decrease in interest rates initially, in contrast to their observed increase.

While Barsky and Sims (2011) suggest a bust scenario, Barsky and Sims (2012) and Larsen and Thorsrud (2019) indicate a boom period. In our analysis, productivity responses to news shocks seem to be temporary, contrasting with the permanent effects reported in boom period studies. We also find an initial increase in inflation and a decrease in interest rates, which differs from the observed trends of deflation and rising rates in these prior works. Nonetheless, a common finding across our study and the earlier research is the observed rise in asset prices following news shocks.

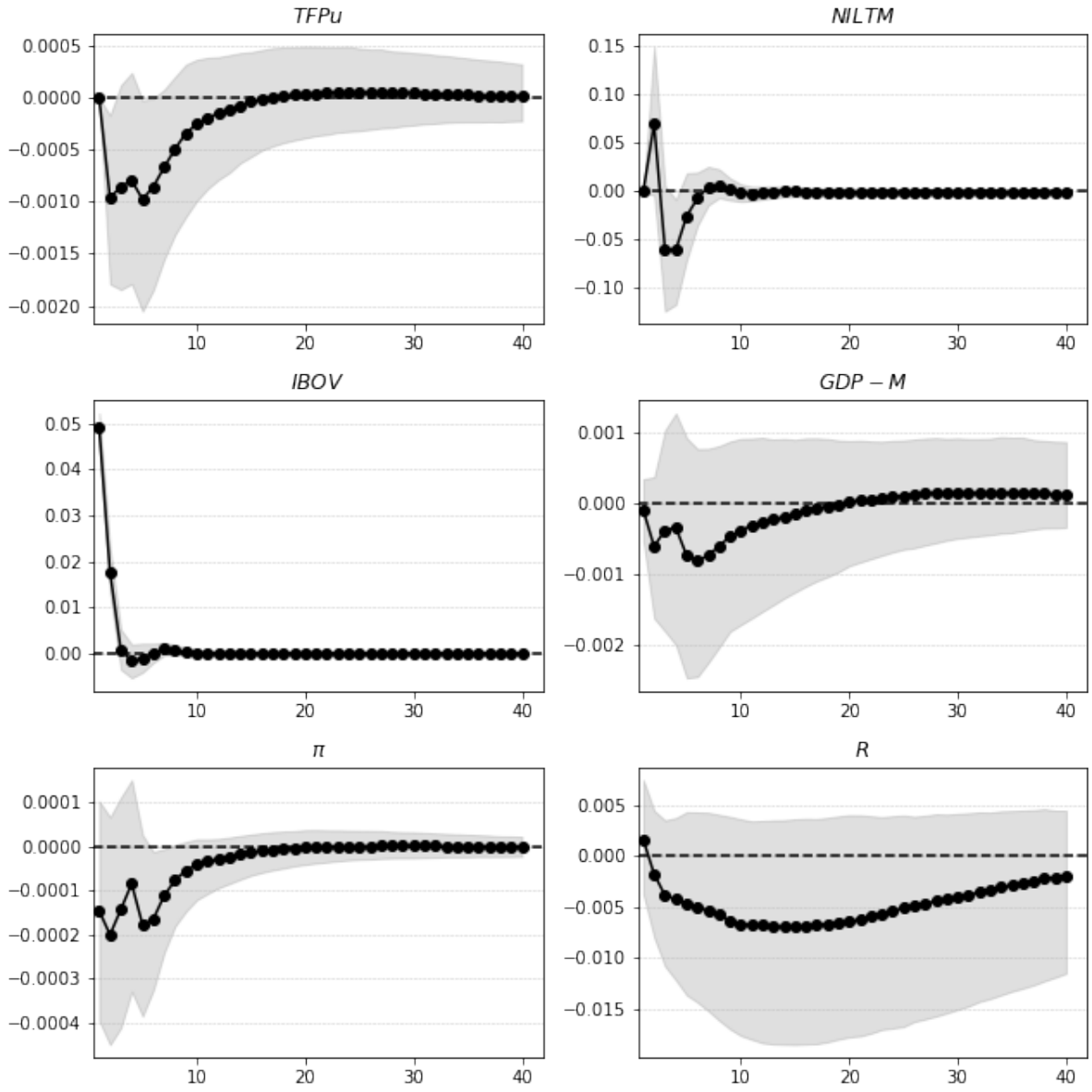
[2] Please refer to Appendix E for detailed information on the VAR model.

Figure 7 – Benchmark Model: News Shock



Each graph reports the percentage response to an initial one-standard-deviation news shock over various response horizons. The solid lines represent the median impulse responses, while the shaded gray areas denote the range within plus or minus one standard deviation. TFP, TFPu, and the variable 'R' are presented in their level form. Inflation (π) and the IBOV index responses are expressed as monthly growth rates. The NILTM index is shown in standard deviations due to its normalization, and GDP-M responses are depicted as percentages.

Figure 8 – Benchmark Model: Noise Shock



Each graph reports the percentage response to an initial one-standard-deviation noise shock over various response horizons. The solid lines represent the median impulse responses, while the shaded gray areas denote the range within plus or minus one standard deviation. TFP, TFPu, and the variable 'R' are presented in their level form. Inflation (π) and the IBOV index responses are expressed as monthly growth rates. The NILTM index is shown in standard deviations due to its normalization, and GDP-M responses are depicted as percentages.

For robustness, we conducted various tests, the details of which are presented in Appendix D. Figure 17 shows the results when substituting GDP-M with IBC-BR, and Figure 18 illustrates the outcomes when replacing TFPu with TFP. In Figure 19, we make both substitutions simultaneously. The results in all these scenarios align closely with our initial findings.

The forecast error variance decomposition (FEVD), depicted in Table 3, is derived from the variables in our benchmark model. The decomposition reveals that news shocks account for almost 14% of the long-run variation in productivity and about 15% over a 4-month horizon. These shocks explain 36% of the variation in asset prices, 30% in the GDP indicator, 7% in inflation, and 17% in the interest rate at a 40-month horizon. The contribution of news shocks remains relatively consistent across the months for these variables, except for the interest rate, which shows a 1.4% contribution at the 4-month horizon. Regarding noise shocks, their impact is generally less pronounced, except in the case of asset prices, where they account for almost 54% of the long-run variance. When compared to the literature, our findings are similar to those of Larsen and Thorsrud (2019), although we observe a more significant impact on TFPu and the interest rate, while their results indicate a larger effect on inflation. Compared to Barsky and Sims (2011), our results show smaller long-run values for all variables except asset prices.

The similarities with Larsen and Thorsrud (2019), as well as the differences from Barsky and Sims (2011), might involve the same arguments put forth by Larsen and Thorsrud (2019) in their comparative analysis. Similarly, these arguments could be applied to the Brazilian context. Firstly, both Brazil and Norway are small, open economies that may be influenced by international business cycle fluctuations. It is possible that *Valor Econômico* primarily focuses on domestic economic developments, which could bias the news towards local events, while international events are less represented in our dataset, thereby reducing the variance attributed to the identified news shocks. Secondly, as we adopt the identification strategy proposed by Larsen and Thorsrud (2019), we also include an unanticipated productivity shock, which accounts for a significant portion of the variability. For instance, in the benchmark model, unanticipated productivity shocks contribute 34.4% to GDP-M, 57.3% to productivity, 4.5% to inflation, and 7.8% to the interest rate in the long run.

In Table 3, 'Total IBOV' and 'Total GDP-M' represent the total variance in IBOV and GDP-M explained by news and noise shocks. We observe that, in the short term, these shocks account for almost all variation in asset prices, and in the long term, they explain about 90% of the variance. This indicates that movements in asset returns are effectively captured by these two types of shocks. Regarding GDP, initially, the shocks are not very significant. However, starting from the 4-month horizon, they account for approximately 36% of the total variance.

We applied forecast error variance decomposition to the same models previously analyzed for impulse responses, with these results also detailed in Appendix D. Table 7 presents the outcomes when GDP-M is replaced with IBC-BR. In both models, the short-run and long-run impacts on IBOV are strikingly similar. However, for IBC-BR, there is a notable difference in the short-run explanation (which is higher) compared to GDP-M. In the long-run, total GDP explains 31.8% of the variance, while IBC-BR accounts for 39.7%. The other variables show

similar results across both models. Figure 8, which illustrates the scenario where TFPu is substituted for TFP, also reveals very similar results for all variables. Finally, in Figure 9, where both TFP and IBC-BR replace TFPu and GDP-M, the only significant deviation from the benchmark GDP-M model is observed in the 1-month horizon and 4-month horizon, with other variables exhibiting comparable outcomes.

Table 3 – Forecast Error Variance Decomposition: Benchmark Model

	h=1	**h=4**	**h=8**	**h=16**	**h=24**	**h=40**
News Shock						
TFPu	0.000	0.148	0.125	0.116	0.125	0.137
NILTM	0.994	0.918	0.905	0.900	0.897	0.892
IBOV	0.399	0.371	0.370	0.368	0.366	0.365
GDP-M	0.019	0.353	0.355	0.341	0.321	0.299
Inflation	0.004	0.053	0.059	0.063	0.066	0.071
Interest Rate	0.005	0.014	0.037	0.095	0.135	0.174
Noise Shock						
	h=1	**h=4**	**h=8**	**h=16**	**h=24**	**h=40**
TFPu	0.000	0.012	0.020	0.022	0.023	0.024
NILTM	0.000	0.019	0.022	0.022	0.022	0.022
IBOV	0.593	0.568	0.554	0.548	0.544	0.538
GDP-M	0.001	0.007	0.012	0.015	0.017	0.019
Inflation	0.005	0.017	0.025	0.026	0.027	0.027
Interest Rate	0.003	0.012	0.014	0.017	0.019	0.021
Total IBOV	0.992	0.939	0.924	0.916	0.910	0.903
Total GDP-M	0.020	0.360	0.367	0.356	0.337	0.318

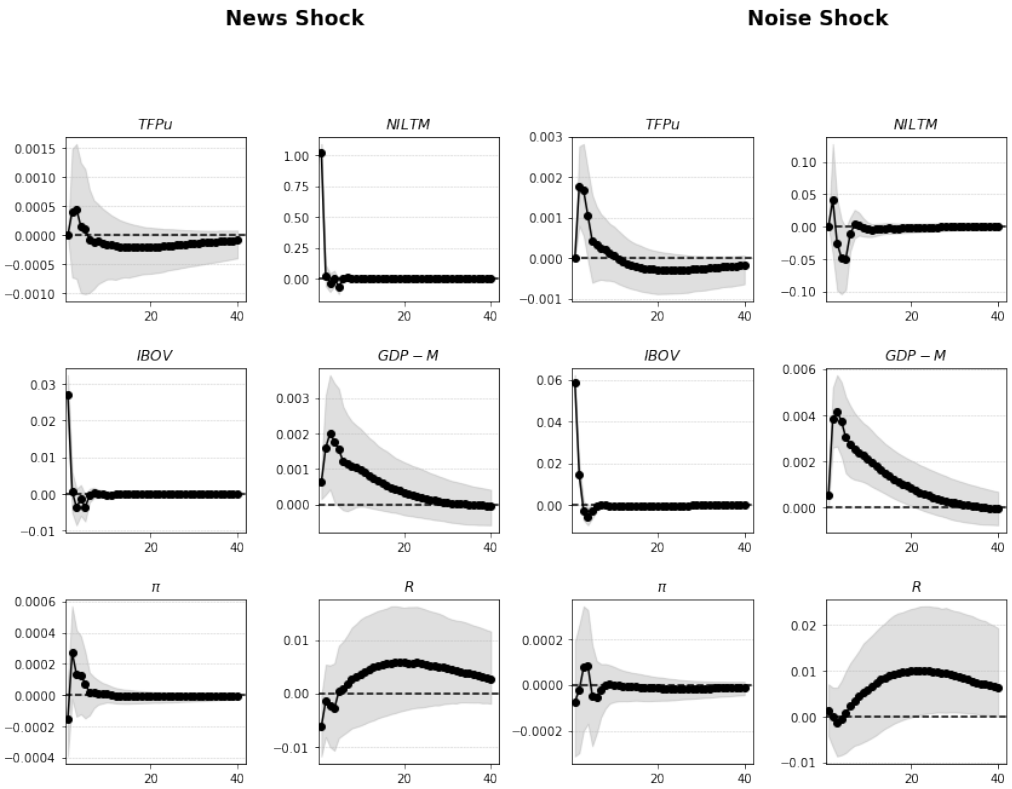
The letter 'h' refers to the forecast horizon. The numbers denote the proportion of the forecast error variance for each variable that is attributable to our identified news shocks at various forecast horizons. 'Total IBOV' represents the total variance of the IBOV index explained by both news and noise shocks combined. 'Total GDP-M' indicates the total variance of the GDP proxy explained by both news and noise shocks combined.

5.2.1 NITVP

Upon analyzing the NITVP news index, distinct response patterns emerged following news and noise shocks, differing from the reactions seen with the NILTM index. As depicted in Figure 9, which showcases the benchmark model with NITVP replacing NILTM, the responses to news shocks were akin to those observed with NILTM, except in the case of inflation, where deflation was noted. The impact on other variables was less marked compared to NILTM and lacked significance, particularly in the case of TFPu.

Conversely, noise shocks elicited a more substantial increase in both TFP and GDP compared to news shocks. This contrasts with the NILTM scenario, where different trajectories were observed for the two types of shocks, and some variables showed significant changes. In the NITVP case, however, the trajectories for news and noise shocks appeared more similar, with several variables demonstrating significant changes. Consequently, distinguishing between news and noise shocks becomes more challenging with the NITVP index.

Figure 9 – NITVP: News and Noise Shocks



On the left side are the responses to news shocks, while on the right side are the responses to noise shocks. Each graph reports the percentage response to an initial one-standard-deviation shock over various response horizons. The solid lines represent the median impulse responses, while the shaded gray areas denote the range within plus or minus one standard deviation. TFP, TFPu, and the variable 'R' are presented in their level form. Inflation (π) and the IBOV index responses are expressed as monthly growth rates. The NITVP index is shown in standard deviations due to its normalization, and GDP-M responses are depicted as percentages.

Table 4 – Forecast Error Variance Decomposition with NITVP index

	News Shock					
	h=1	**h=4**	**h=8**	**h=16**	**h=24**	**h=40**
TFPu	0.000	0.009	0.0129	0.015	0.0169	0.018
NITVP	0.9996	0.933	0.914	0.905	0.899	0.892
IBOV	0.172	0.166	0.168	0.167	0.166	0.164
GDP-M	0.001	0.021	0.025	0.026	0.026	0.026
Inflation	0.004	0.022	0.026	0.026	0.027	0.027
Interest Rate	0.008	0.014	0.014	0.016	0.017	0.019
	Noise Shock					
	h=1	**h=4**	**h=8**	**h=16**	**h=24**	**h=40**
TFPu	0.000	0.028	0.028	0.031	0.033	0.036
NITVP	0.000	0.013	0.018	0.019	0.019	0.020
IBOV	0.804	0.756	0.740	0.733	0.729	0.723
GDP-M	0.001	0.085	0.095	0.098	0.094	0.088
Inflation	0.004	0.016	0.020	0.022	0.022	0.024
Interest Rate	0.003	0.010	0.012	0.019	0.025	0.034

The letter 'h' refers to the forecast horizon. The numbers denote the proportion of the forecast error variance for each variable that is attributable to our identified news shocks at various forecast horizons.

The forecast error variance decomposition presented in Table 4, featuring the NITVP index, indicates that both news and noise shocks contribute similarly and modestly. The most significant contribution from news shocks is to the IBOV, at around 16%, which is less than half compared to the benchmark model. Conversely, noise shocks account for 72% of the variation in asset prices, a figure that exceeds the contribution observed in the benchmark model.

The FEVD unveils marked differences between the NITVP and NILTM indices. Notably, the long-run impacts of news and noise shocks are consistently lower for the NITVP index. Within the realm of the IBOV variable, the disparity is pronounced: news shocks explain 36% of the variance with the NILTM index, as opposed to just 16% with NITVP. Conversely, noise shocks have a more significant effect at 72% with NITVP, compared to 54% with NILTM.

The underlying reasons for these variations may be rooted in the fundamental construction

of the news indices. The NILTM index incorporates a threshold that filters out less impactful topics in explaining asset returns. When a topic does not contribute meaningfully to asset price fluctuations, it is systematically excluded from the index's calculation. The NITVP index, conversely, maintains all topics, which means it reflects the cumulative impact of a broader array of news items, including those with minimal individual explanatory power.

The NITVP index, by encompassing a broad range of topics, might unintentionally capture noise from news items of marginal relevance, potentially exaggerating the influence of noise shocks in the variance decomposition. This inclusive approach may reflect short-lived fluctuations in the dataset that do not correspond with core economic developments but are instead related to transient, non-systemic variations. In contrast, the NILTM index implements a selective threshold, concentrating on news that meets a specified relevance bar. This methodological decision could make the NILTM index more resonant with structural changes in economic conditions, thus offering a more precise depiction of the economic narrative affecting the variables.

5.3 Three-Variable System

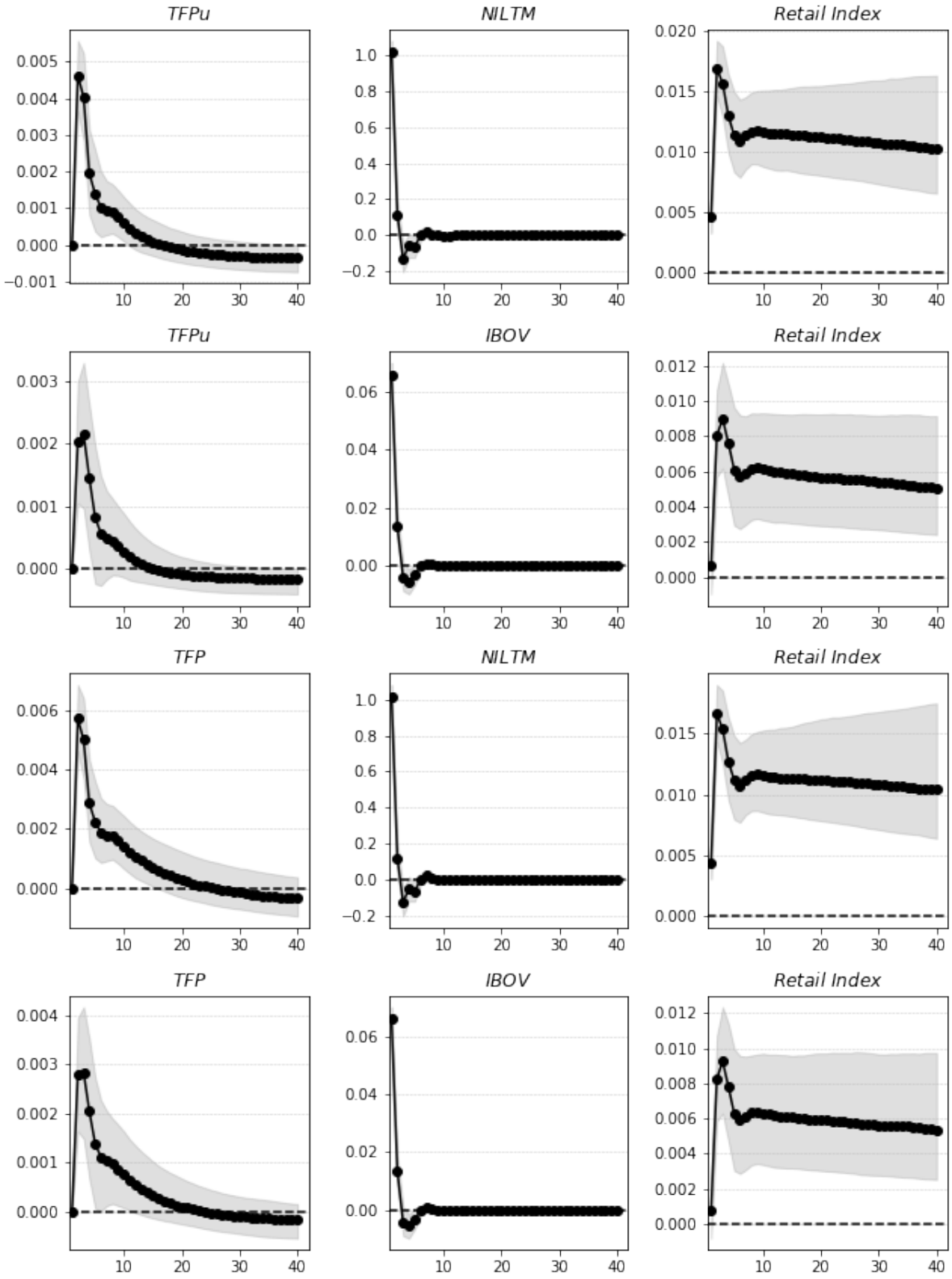
In this section, we show that the identification strategy proposed by Larsen and Thorsrud (2019), which utilizes a news index, yields consistent results. Drawing inspiration from Beaudry and Portier (2006), we consider a three-variable system comprising Productivity, Stock Prices, and Consumption, arranged in that order. For our analysis, we use the Retail Index (RI) as a proxy for consumption. We identify news shocks based on unexpected innovations in asset prices and the news index.

Figure 10 presents the impulse responses to news shocks for each model. The first two models utilize the news index and asset prices to identify news shocks with the TFPu productivity measure. The latter two models employ the TFP productivity measure. In all the cases as shown in Figure 10, there is a distinct and enduring increase in the retail index following news shocks, consistent with the patterns observed by Beaudry and Portier (2006). Turning our attention to the productivity measures, both TFP and TFPu initially show an increase in response to news shocks. This is in line with our previous findings but contrasts with Beaudry and Portier (2006), where TFP and TFPu are suggested to have a permanent increase in the long-run. In our analysis, however, we observe that the rise in TFP and TFPu is followed by a gradual decline, indicating that the effects of news shocks on these productivity measures are not enduring. Moreover, the impulse responses indicate more significant effects when the news index is utilized.

As shown in Table 5, when news shocks are identified using the news index, they explain 13.6% and 14.6% of the variance in TFPu and TFP, respectively. In contrast, when identified using asset prices, the explanation rate for both measures is around 4.5%. A significant difference is also observed in the Retail Index, with approximately 37% explained using the news index compared to 10% with asset prices.

These findings support the identification approach of Larsen and Thorsrud (2019) using the news index. They suggest that using the news index, as opposed to asset prices, leads to more robust results. Consistent with their argument, relying solely on asset prices might blur the distinction between news and noise, as they can be a composite of both. The news index, on the other hand, provides a more accurate reflection of true information, distinguishing more clearly between the effects of news and noise shocks.

Figure 10 – Three-Variable System: Impulse Responses to News Shocks



In the first row, the impulse responses are identified using the NILTM index to capture news shocks. The second row presents impulse responses identified using asset prices to discern news shocks. Both of these rows employ TFPu as the measure of productivity. In the third row, news shocks are identified once again through the NILTM index, while the fourth row utilizes asset prices for this purpose. However, in these cases, the measure of productivity used TFP. Each graph reports the percentage response to an initial one-standard-deviation shock over various response horizons.

Table 5 – Three-Variable System: Forecast Error Variance Decomposition

News Shock Using the News Index

	h=1	**h=16**	**h=40**		**h=1**	**h=16**	**h=40**
TFPu	0.000	0.130	0.136	TFP	0.000	0.145	0.146
NILTM	0.995	0.969	0.968	NILTM	0.996	0.972	0.970
RI	0.063	0.390	0.378	RI	0.057	0.383	0.366

News Shock Using the Asset Prices

	h=1	**h=16**	**h=40**		**h=1**	**h=16**	**h=40**
TFPu	0.000	0.038	0.042	TFP	0.000	0.046	0.047
IBOV	0.989	0.954	0.951	IBOV	0.989	0.954	0.950
RI	0.002	0.109	0.105	RI	0.003	0.113	0.108

The letter 'h' refers to the forecast horizon. The numbers denote the proportion of the forecast error variance for each variable that is attributable to our identified news shocks at various forecast horizons. In the upper section, news shocks are discerned based on unexpected innovations in the news index, considered separately for two scenarios: with TFP and TFPu. In the lower section, the identification of news shocks is based on the fluctuations in asset prices, again examined for both TFP and TFPu cases.

6 CONCLUSION

This study presented a comprehensive analysis of the influence of news on the Brazilian economic cycle, employing a combination of econometric models and textual data analysis. By utilizing the Latent Dirichlet Allocation (LDA) model on a vast corpus of journalistic articles from *"Valor Econômico*, we were able to transform these texts into a series of monthly news indices (NITVP and NILTM), featuring easily interpretable topics. This approach offered new insights into the dynamics of financial markets and economic activity.

Our analyses revealed that news and noise shocks, as captured by these news indices, have a substantial impact on both asset prices and key macroeconomic variables. Notably, these shocks explained a significant proportion of the variation in asset prices in both the short and long term, underscoring the importance of news information in market fluctuations. This aspect of our analysis bears a resemblance to the findings of Larsen and Thorsrud, further validating the significance of news content in economic analysis.

Furthermore, the results highlighted notable differences between the NITVP and NILTM indices in terms of response to news and noise shocks, with significant implications for interpreting economic data. Comparing our findings with previous studies, such as Beaudry and Portier (2006) and Larsen and Thorsrud (2019), reinforced the validity of our approach and showed that using news indices can provide more accurate insights than analyses based solely on asset prices. This alignment with Larsen and Thorsrud's methodology not only lends credibility to the application of news indices in economic analysis but also allows for a complementary understanding of the economic dynamics, enriching the broader insights established by Beaudry and Portier (2006).

We conclude that integrating textual data into econometric models offers a powerful tool for better understanding the nuances of the economic cycle. This approach enriches our understanding of the underlying forces shaping the Brazilian economy. In summary, this work paves new pathways for economic research, demonstrating the invaluable role of textual data in economic analysis.

REFERENCES

APEL, Mikael; GRIMALDI, Marianna Blix. The information content of central bank minutes. *Working Paper Series*, Sveriges Riksbank, n. 261, 2012.

ASUNCION, Arthur; WELLING, Max; SMYTH, Padhraic; TEH, Yee Whye. On smoothing and inference for topic models. *Uncertainty in Artificial Intelligence*, AUAI Press, p. 27–34, 2009.

BAKER, Scott R.; BLOOM, Nicholas; DAVIS, Steven J. Measuring economic policy uncertainty. *The Quarterly Journal of Economics*, Oxford University Press, v. 131, n. 4, p. 1593–1636, 2016.

BARSKY, Robert B; SIMS, Eric R. News shocks and business cycle. *Journal of Monetary Economics*, Elsevier, v. 58, p. 273–289, 2011.

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.

BEAUDRY, Paul; PORTIER, Frank. Stock prices, news, and economic fluctuations. *The American Economic Review*, American Economic Association, v. 96, n. 4, p. 1293–1307, 2006.

BLEI, David M. Probabilistic topic models. *Communications of the ACM*, v. 55, n. 4, p. 77–84, 2012.

BLEI, David M; NG, Andrew Y; JORDAN, Michael I. Latent dirichlet allocation. *Journal of Machine Learning Research*, v. 3, p. 993–1022, 2003.

DIEPPE, Alistair; LEGRAND, Romain; VAN, Roye Bjorn. The bayesian estimation, analysis and regression (bear) toolbox. *Working Paper Series*, European Central Bank, 2016.

ELLINGSEN, Jon; LARSEN, Vegard H; THORSRUD, Leif Anders. News media versus fred-md for macroeconomic forecasting. *Journal of Applied Econometrics*, Wiley, v. 37, n. 1, p. 63–81, 2022.

FERREIRA, Leonardo N. Forecasting with var-text and dfm-text models: exploring the predictive power of central bank communication. *Working Papers Series*, Central Bank of Brazil, n. 559, 2021.

GAN, Jingxian; QI, Yong. Selection of the optimal number of topic for lda topic model - taking patent policy analysis as an example. *entropy*, MDPI, v. 23, n. 10, 2021.

GARCÍA, Diego. Sentiment during recessions. *The Journal of Finance*, The American Finance Association, v. 68, n. 3, p. 1267–1300, 2013.

GENTZKOW, Matthews; KELLY, Bryan; TADDY, Matt. Text as data. *Journal of Economic Literature*, American Economic Association, v. 57, n. 3, p. 535–574, 2019.

GOMES, Victor; PESSÔA, Samuel de Abreu; VELOSO, Fernando A. Evolução da produtividade total dos fatores na economia brasileira: uma análise comparativa. *Pesquisa de Planejamento Econômico*, v. 33, n. 3, p. 389–434, 2003.

GRIFFITHS, Thomas L; STEYVERS, Mark. Finding scientific topics. *Proceedings of the National Academy of Sciences*, v. 101, p. 5228–5235, 2004.

HANSEN, Stephen; MCMAHON, Michael. Shocking language: Understanding the macroeconomic effects of central bank communication. *Journal of International Economics*, Elsevier, v. 99, p. S114–S133, 2016.

HANSEN, Stephen; MCMAHON, Michael; PRAT, Andrea. Transparency and deliberation within the fomc: A computational linguistic approach. *The Quarterly Journal of Economics*, Oxford University Press, v. 133, n. 2, p. 801–870, 2018.

HOFFMAN, Matthew D; BLEI, David M; BACH, Francis. Online learning for latent dirichlet allocation. *Neural Information Processing Systems*, 2010.

JÚNIOR, José Ronaldo de Castro Souza; CORNELIO, Felipe M. Estoque de capital fixo no brasil: Séries desagregadas anuais, trimestrais e mensais. Instituto de Pesquisa Econômica Aplicada, n. 2580, 2020.

KALAMARA, Eleni; TURRELL, Arthur; REDL, Chris; KAPETANIOS, George; KAPADIA, Sujit. Making text count: Economic forecasting using newspaper text. *Journal of Applied Econometrics*, Wiley, v. 37, n. 5, p. 896–919, 2022.

LARSEN, Vegard H; THORSRUD, Leif A. The value of news for economic developments. *Journal of Econometrics*, Elsevier, v. 210, p. 203–218, 2019.

LARSEN, Vegard H; THORSRUD, Leif Anders; ZHULANOVA, Julia. News-driven inflation expectations and information rigidities. *Journal of Monetary Economics*, Elsevier, v. 117, p. 507–520, 2021.

MARDONES-SEGOVIA, Constanza; WHEELER, Jordan M; CHOI, Hye-Jeong; WANG, Shiyu; COHEN, Allan S. Model selection for latent dirichlet allocation in assessment data. *Psychological Test and Assessment Modeling*, v. 65, n. 1, p. 3–35, 2023.

MCALLIN, Kenichiro; WEST, Mike. Dynamic bayesian predictive synthesis in time series forecasting. *Journal of Econometrics*, Elsevier, v. 210, p. 155–169, 2019.

NAKAJIMA, Jouchi. Time-varying parameter var model with stochastic volatility: An overview of methodology and empirical applications. *Institute for Monetary and Economic Studies*, Bank of Japan, v. 29, p. 107–142, 2011.

NAKAJIMA, Jouchi; WEST, Mike. Bayesian analysis of latent threshold dynamic models. *Journal of Business and Economic Statistics*, American Statistical Association, v. 31, n. 2, p. 151–164, 2013.

TETLOCK, Pau C; SAAR-TSECHANSKY, Maytal; MACSKASSY, Sofus. More than words: Quantifying language to measure firms'fundamentals. *The Journal of Finance*, The American Finance Association, v. 63, n. 4, p. 1437–1467, 2008.

THORSRUD, Leif Anders. Words are the new numbers: A newsy coincident index of the business cycle. *Journal of Business Economic Statistics*, American Statistical Association, v. 38, n. 2, 2018.

APPENDIX A – DETAILED EXAMPLE OF DATA PREPROCESSING

In Figure 11, an example of a news article prior to the preprocessing detailed in Chapter 2 is presented. This example includes the headline in bold followed by the complete text of the article.

Figure 11 – News Article: Example

Focus: mercado aponta estabilidade em inflação para 2011
SÃO PAULO - O mercado manteve a projeção de inflação para este ano, medida pelo Índice de Preços ao Consumidor Amplo (IPCA), prevendo avanço de 6,31%, segundo estimativa do boletim Focus, do Banco Central (BC), divulgado hoje.
A atual projeção manteve a expectativa de inflação para este no mesmo patamar da leitura anterior do relatório do BC, mas ampliou a projeção de inflação para 2012, que ficou em 5,28%, contra 5,20% na semana passada. Este aumento, entretanto, resultou de avanço nas projeções de apenas 11 analistas, dos cerca de 100 consultados pelo BC para a elaboração do boletim. Do total, oito reduziram suas projeções de inflação para o próximo ano, e os demais mantiveram a projeção inalterada, informa fonte econômica do governo.
Para os próximos 12 meses o mercado ampliou a expectativa de inflação para 5,40%, mantendo pela sexta semana consecutiva a perspectiva de alta. No levantamento da semana passada os analistas consultados pelo BC projetavam inflação de 5,37% para o mesmo período.
Já para o Índice Geral de Preços – Mercado (IGPM), a mediana das projeções para este ano aponta para 5,31%, o que representa um avanço das expectativas, que indicavam 5,22% há uma semana.
Para o Índice de Preços ao Consumidor (IPC-Fipe), a projeção para 2011 é de 5,15%, avançando pela sexta semana consecutiva. No boletim anterior a expectativa para o indicador era de crescimento de 5,13% neste ano.
A mediana dos analistas consultados pelo BC ampliou a previsão para crescimento do Índice Geral de Preços – Disponibilidade Interna (IGP-DI) para 5,39% para este ano, ligeiro avanço na comparação com o crescimento de 5,34% projetado para o indicador na semana passada. (Bruno De Vizia | Valor)

* Note: This news article is available at <https://valor.globo.com/brasil/noticia/2011/07/25/focus-mercado-aponta-estabilidade-em-inflacao-para-2011.ghtml>. Last accessed on November 3, 2023. Originally published on July 25, 2011.

As previously mentioned, based on the assumption that the first paragraph conveys the main point of the news article to the reader and in order to achieve computational efficiency, we have condensed the articles to include only the title and the first paragraph. This reduction is illustrated in Figure 12.

Figure 12 – News Article: Title and First Paragraph

Focus: mercado aponta estabilidade em inflação para 2011
SÃO PAULO - O mercado manteve a projeção de inflação para este ano, medida pelo Índice de Preços ao Consumidor Amplo (IPCA), prevendo avanço de 6,31%, segundo estimativa do boletim Focus, do Banco Central (BC), divulgado hoje.

After this initial reduction, we begin preprocessing the text data. Firstly, we eliminate punctuation and special characters, which are highlighted in green. Secondly, we remove numbers and the names of publication cities, shown in blue. Thirdly, we discard stop-words, marked in red.

Figure 13 – News Article: Cleaning Process

Focus: mercado aponta estabilidade em inflação para 2011
SÃO PAULO - O mercado manteve a projeção de inflação para este ano, medida pelo Índice de Preços ao Consumidor Amplo (IPCA), prevendo avanço de 6,31%, segundo estimativa do boletim Focus, do Banco Central (BC), divulgado hoje.

After completing the text cleaning process in Figure 13, we convert all words to lowercase and then tokenize them, yielding the final result as shown in Figure 14.

Figure 14 – News Article: After Cleaning Process

'*focus*'; '*mercado*'; '*aponta*'; '*estabilidade*'; '*inflação*'; '*mercado*'; '*manteve*'; '*projeção*'; '*inflação*'; '*medida*'; '*índice*'; '*preços*'; '*consumidor*'; '*ipca*'; '*prevendo*'; '*avanço*'; '*estimativa*'; '*boletim*'; '*focus*'; '*banco*'; '*central*'; '*bc*'; '*divulgado*';

APPENDIX B – LDA: TOPICS AND THEIR LABELS

The **'Topic'** column lists the various topics identified from the output of LDA. The **'Label'** column provides a descriptive name for each topic, derived from our subjective interpretation of the ten most significant words associated with that topic. These labels, which are not an output of the LDA process, reflect our understanding and analysis of the content of each topic. **'Number of Documents'** quantifies the articles predominantly characterized by a specific topic, indicating the topic's representativeness within the document corpus. Additionally, the key terms are presented in Portuguese alongside their English translations.

Table 6 – Thematic Mapping: Topics and Their Labels

Topic	**Label**	**Number of Documents**	**Important Words**
Topic 0	United States Economy	18515	eua, estados unidos, fed, americano, mercados, espera, americana, foco, globais, riscos
			USA, United States, Federal Reserve, American, markets, wait, American (feminine), focus, global, risks
Topic 1	Knowledge	14130	vacinação, social, estudo, digital, aumentar, população, educação, precisa, especialistas, problemas
			vaccination, social, study, digital, increase, population, education, needs, specialists, problems
Topic 2	Fiscal Policy	12793	fiscal, recursos, pagamento, estados, medidas, reduzir, orçamento, dívida, pública, auxílio
			tax, resources, payment, states, measures, reduce, budget, debt, public, aid
Topic 3	Fear	13029	estado, guerra, segurança, direito, decidiu, autoridades, icms, valores, especial, Minas Gerais
			state, war, security, right, decided, authorities, tax, values, special, Minas Gerais

Continued on next page

Table 6 – Continued from previous page

Topic	Label	Number of Documents	Important Words
Topic 4	Communication	11956	acordo, serviços, rede, uso, negociações, acesso, via, negociaçãoo, serviço, internet
			agreement, services, network, use, negotiations, access, route, negotiation, service, internet
Topic 5	Stock Market	39380	queda, alta, dólar, forte, bolsa, ibovespa, sessão, movimento, exterior, york
			drop, rise, dollar, strong, stock market, ibovespa, session, movement, abroad, New York
Topic 6	Financial Transactions	16588	mercado, compra, ativos, venda, investidores, financeiro, aquisição, papéis, comprar, vender
			market, purchase, assets, sale, investors, financial, acquisition, securities, buy, sell
Topic 7	Inflation	18391	alta, preços, índice, março, dezembro, outubro, novembro, abril, maio, fevereiro
			rise, prices, index, March, December, October, November, April, May, February
Topic 8	Labor Unions	10500	nacional, união, São Paulo, entidade, cidades, paulista, democracia, trabalhadores, reajuste, concessão
			national, union, São Paulo, entity, cities, from São Paulo, democracy, workers, adjustment, concession
Topic 9	Oil	23090	produção, petróleo, preço, recorde, commodities, combustíveis, volume, sul, conforme, toneladas

Continued on next page

Table 6 – Continued from previous page

Topic	Label	Number of Documents	Important Words
			production, oil, price, record, commodities, fuels, volume, south, as per, tons
Topic 10	Rules	9096	dias, medida, prazo, mudança, mudanças, tema, regras, Brasília, transição, podera
			days, measure, term, change, changes, theme, rules, Brasília, transition, may
Topic 11	Macroeconomics	15119	setor, aumento, crescimento, investimento, prevê, pib, expansão, desempenho, anual, ritmo
			sector, increase, growth, investment, predicts, GDP, expansion, performance, annual, pace
Topic 12	Corporate Announcements	12643	empresa, companhia, informou, anunciou, ceo, administração, funcionários, comunicado, realizada, anuncia
			company, corporation, informed, announced, CEO, administration, employees, statement, held, announces
Topic 13	Trade Balance	20534	bilhões, us, bilhão, bi, total, ação, positivo, comercial, euros, negativo
			billions, US, billion, billion (abbreviation), total, share/action, positive, trade, euros, negative
Topic 14	Corporate Credit	17217	empresas, crédito, operações, maiores, bancos, clientes, caixa, dinheiro, companhias, custo
			companies, credit, operations, largest, banks, clients, cash, money, companies, cost
Topic 15	Rating	7251	risco, sp, relatório, afirma, nota, segue, ve, avaliação, avalia, alto
			risk, SP, report, states, note, follows, see, evaluation, assesses, high

Continued on next page

Table 6 – Continued from previous page

Topic	Label	Number of Documents	Important Words
Topic 16	Research	8311	pesquisa, principais, impacto, europa, mostra, pressão, semanas, econômico, instituto, divulgada
			research, main, impact, Europe, shows, pressure, weeks, economic, institute, released
Topic 17	Monetary Policy	15554	inflação, Banco Central, BC, acima, IBGE, expectativa, abaixo, divulgado, meta, média
			inflation, Central Bank, CB, above, IBGE (Brazilian Institute of Geography and Statistics), expectation, below, disclosed, target, average
Topic 18	Lula	36018	ex, Lula, pt, Lula Silva, governador, eleições, eleitoral, campanha, TSE, partido
			former, Lula, PT (Workers' Party), Lula Silva, governor, elections, electoral, campaign, TSE (Superior Electoral Court), party
Topic 19	Brazilian Politics	26888	projeto, senado, lei, câmara, congresso, proposta, análise, casa, aprovou, aprovação
			project, senate, law, chamber, congress, proposal, analysis, house, approved, approval
Topic 20	Brazilian Presidents	19336	presidente, executivo, atual, vice, domingo, eleito, cargo, comando, chefe, entrevista
			president, executive, current, vice, Sunday, elected, position, command, chief, interview
Topic 21	Uncertainty	17723	economia, global, política, brasileira, crise, cenário, recuperação, econômica, internacional, ambiente

Continued on next page

Table 6 – Continued from previous page

Topic	Label	Number of Documents	Important Words
			economy, global, politics, Brazilian, crisis, scenario, recovery, economic, international, environment
Topic 22	Justice	38797	federal, STF, decisão, processo, justiça, Supremo Tribunal Federal, público, pedido, ministros, defesa
			federal, Supreme Federal Court, decision, process, justice, Supreme Federal Court , public, request, ministers, defense
Topic 23	Global Data	10257	dados, mundo, divulgação, demanda, quase, principal, manter, devido, continua, Japão
			data, world, disclosure, demand, almost, main, maintain, due, continues, Japan
Topic 24	Negotiations	13676	início, operação, negócios, volta, frente, busca, destaques, vista, estratégia, passou
			beginning, operation, business, return, front, search, highlights, view, strategy, passed
Topic 25	Europe	8235	Ucrânia, ficar, série, começa, locais, evitar, Reino Unido, vida, começou, qualquer
			Ukraine, stay, series, begins, places, avoid, United Kingdom, life, started, any
Topic 26	Copom Meetings	12986	juros, taxa, redução, reunião, corte, taxas, longo, semana passada, mantém, espaço
			interest, rate, reduction, meeting, cut, rates, long, last week, keeps, space
Topic 27	Results	28826	milhões, trimestre, período, lucro, resultado, receita, anterior, registrou, lucro líquido, resultados

Continued on next page

Table 6 – Continued from previous page

Topic	Label	Number of Documents	Important Words
			millions, quarter, period, profit, result, revenue, previous, recorded, net profit, results
Topic 28	Rio de Janeiro	12130	rio, gestão, região, cidade, Rio de Janeiro, centro, áreas, modelo, construção, receber
			river, management, region, city, Rio de Janeiro, center, areas, model, construction, receive
Topic 29	Industry Production	19621	passado, janeiro, setembro, vendas, queda, julho, agosto, junho, indústria, comparação
			past, January, September, sales, drop, July, August, June, industry, comparison
Topic 30	BRICS	29868	Brasil, país, China, países, brasileiro, comércio, mundial, Argentina, Índia, UE
			Brazil, country, China, countries, Brazilian, trade, global, Argentina, India, EU
Topic 31	Russia	12716	Rússia, pessoas, brasileiros, tempo, evento, possibilidade, causa, sábado, curtas, encontro
			Russia, people, Brazilians, time, event, possibility, cause, Saturday, shorts, meeting
Topic 32	Bolsonaro	13157	Bolsonaro, Petrobras, conta, Twitter, estatal, família, redes sociais, fala, alvo, fica
			Bolsonaro, Petrobras, account, Twitter, state-owned, family, social media, speaks, target, stays
Topic 33	Government Policy	17473	governo, ministro, afirmou, ministério, secretário, secretaria, fazenda, defende, Guedes, deixar

Continued on next page

Table 6 – Continued from previous page

Topic	Label	Number of Documents	Important Words
			government, minister, stated, ministry, secretary, secretariat, treasury, defends, Guedes, leave
Topic 34	Energy	14200	energia, diretor, agência, leilão, contratos, contrato, gás, geração, responsável, consumidores
			energy, director, agency, auction, contracts, contract, gas, generation, responsible, consumers
Topic 35	Institutional Investing	12884	capital, investimentos, fundo, participação, vale, plataforma, controle, negócio, restrições, acionistas
			capital, investments, fund, participation, worth, platform, control, business, restrictions, shareholders
Topic 36	Arthur Lira	10950	feira, terça, semana, futuro, agenda, manhã, melhor, voltou, Lira, força
			fair, Tuesday, week, future, schedule, morning, better, returned, Lira, strength
Topic 37	Health	16381	ações, saúde, casos, oferta, tecnologia, vacinas, vacina, parceria, informações, planos
			actions, health, cases, offer, technology, vaccines, vaccine, partnership, information, plans
Topic 38	Banks and Development	11400	banco, plano, programa, projetos, desenvolvimento, objetivo, infraestrutura, instituição, financiamento, atuação
			bank, plan, program, projects, development, goal, infrastructure, institution, financing, performance
Topic 39	Brand Expansion	10997	cerca, base, produtos, marca, chegou, custos, capacidade, linha, chega, milhão

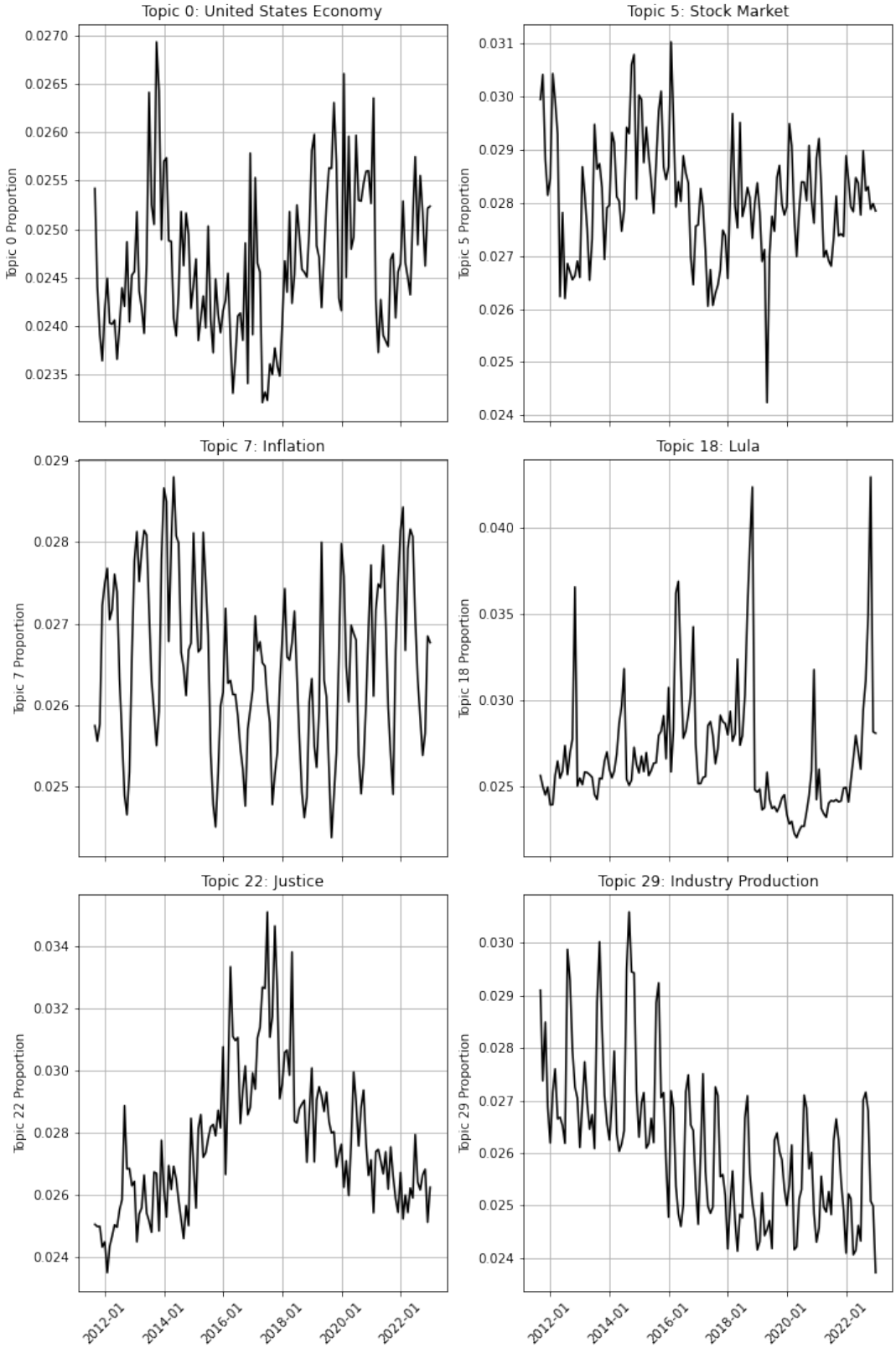
Continued on next page

Table 6 – Continued from previous page

Topic	Label	Number of Documents	Important Words
			about, base, products, brand, arrived, costs, capacity, line, arrives, million

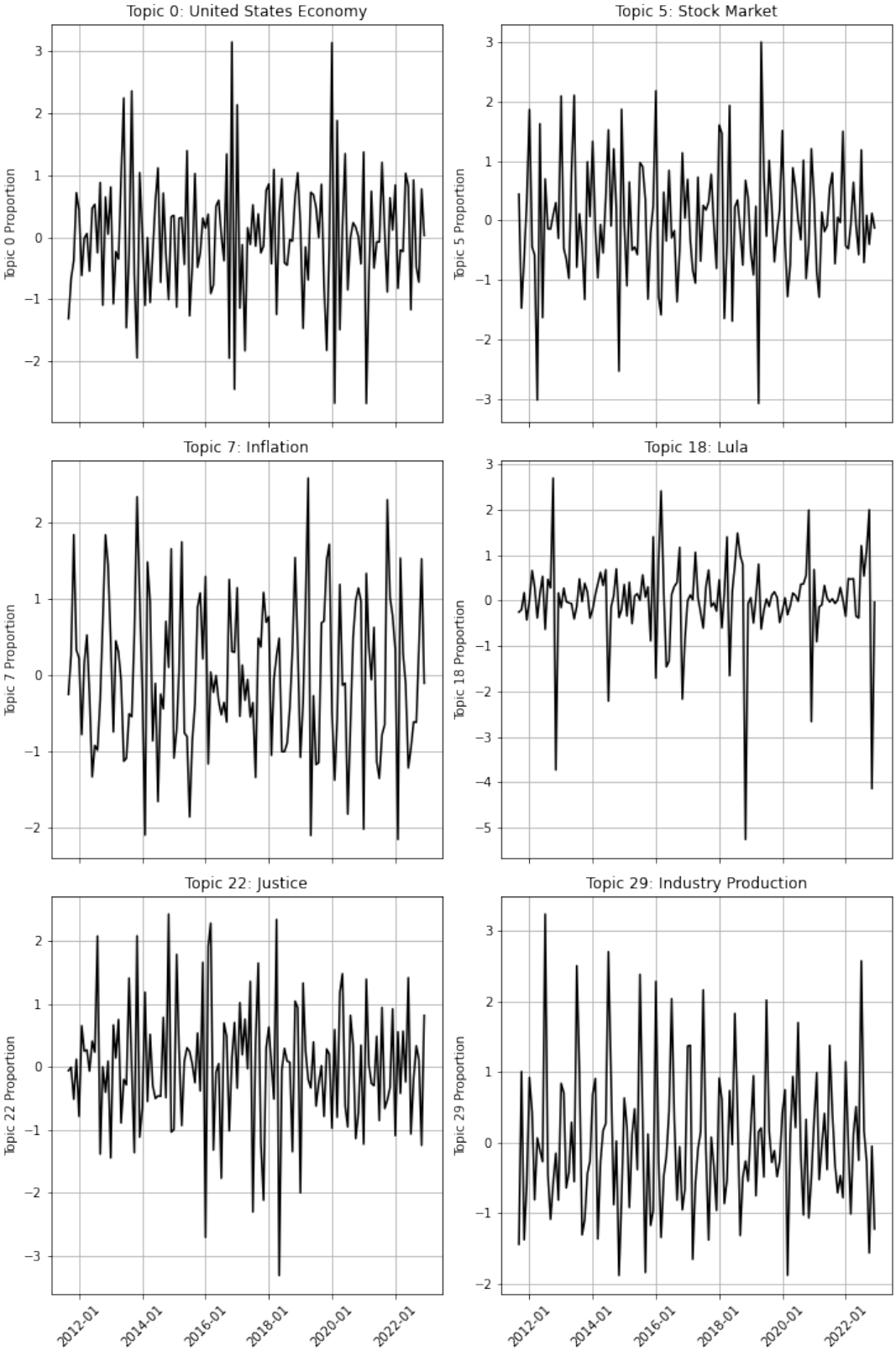
APPENDIX C – LDA: RESULTS

Figure 15 – LDA Topic Distribution: Before Transformation



Note: Topic proportion for selected 6 topics over time by month before any transformation.

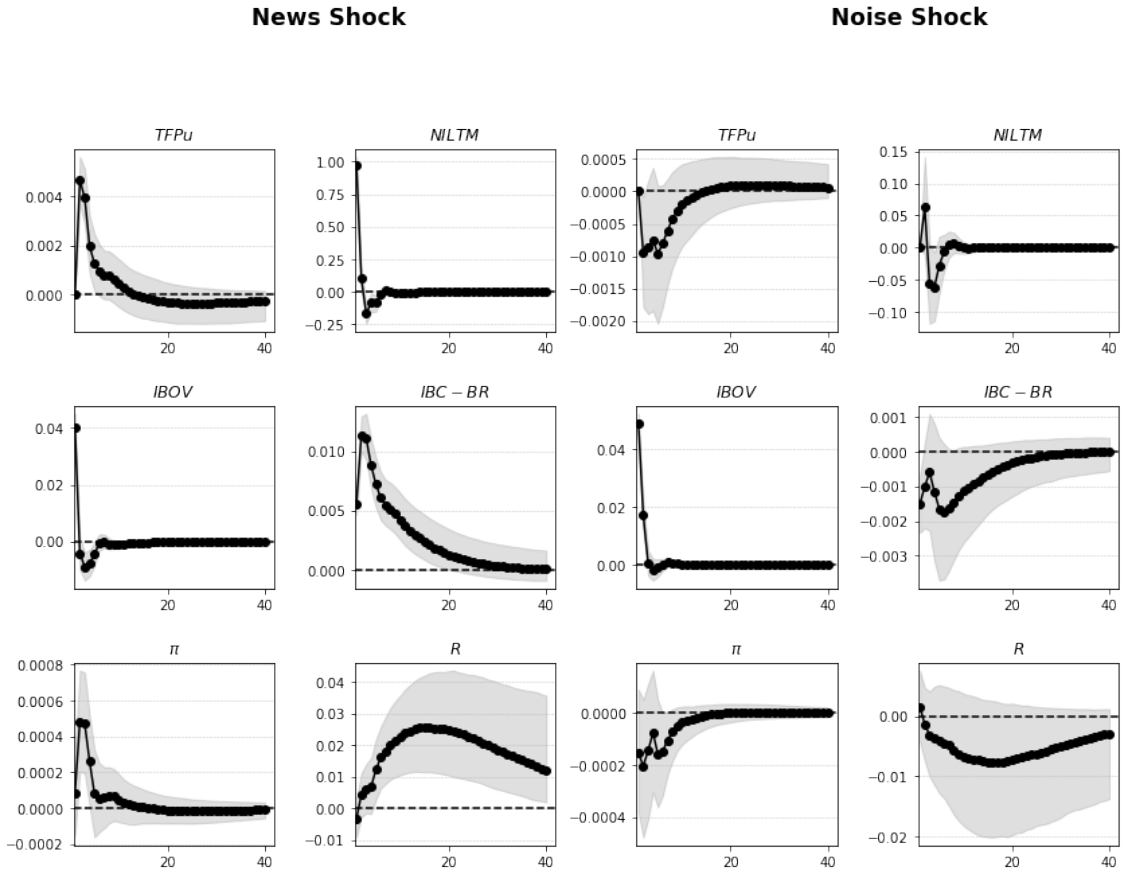
Figure 16 – LDA Topic Distribution: After Transformation



Note: The graph illustrates the proportion of six selected topics over time by month, following a logarithmic transformation, first differencing, and standardization of the data.

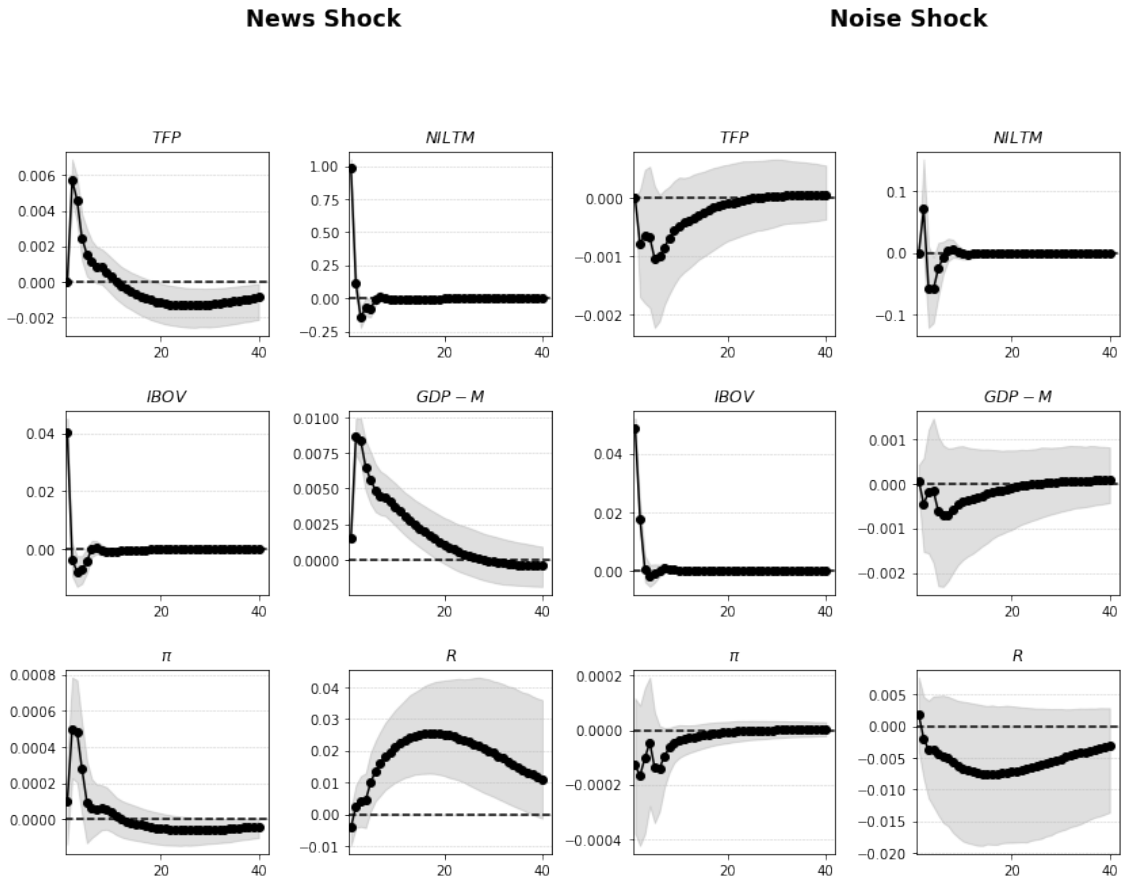
APPENDIX D – ADDITIONAL RESULTS

Figure 17 – Alternative Model 1



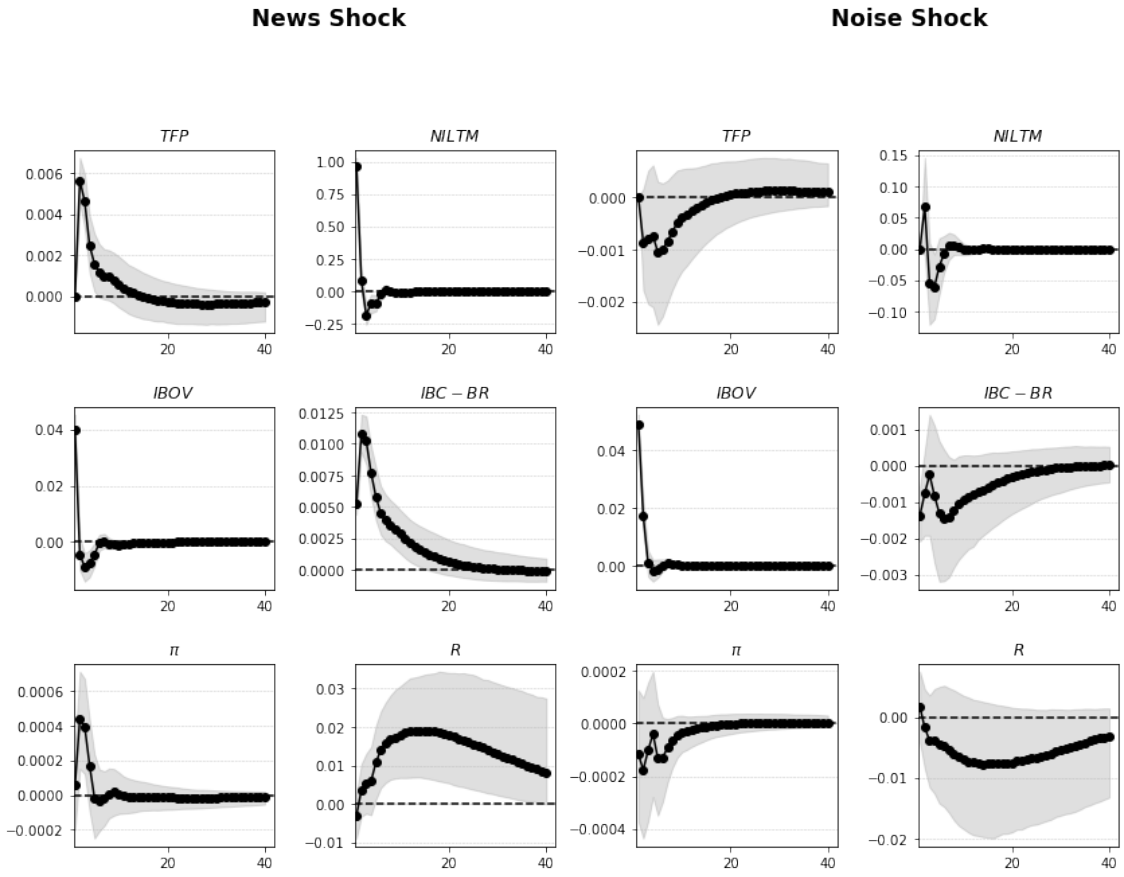
On the left side are the responses to news shocks, while on the right side are the responses to noise shocks. Each graph reports the percentage response to an initial one-standard-deviation shock over various response horizons. The solid lines represent the median impulse responses, while the shaded gray areas denote the range within plus or minus one standard deviation. TFPu, and the variable 'R' are presented in their level form. Inflation (π) and the IBOV index responses are expressed as monthly growth rates. The NILTM index is shown in standard deviations due to its normalization, and IBC-BR responses are depicted as percentages.

Figure 18 – Alternative Model 2



On the left side are the responses to news shocks, while on the right side are the responses to noise shocks. Each graph reports the percentage response to an initial one-standard-deviation shock over various response horizons. The solid lines represent the median impulse responses, while the shaded gray areas denote the range within plus or minus one standard deviation. TFP, and the variable 'R' are presented in their level form. Inflation (π) and the IBOV index responses are expressed as monthly growth rates. The NILTM index is shown in standard deviations due to its normalization, and GDP-M responses are depicted as percentages.

Figure 19 – Alternative Model 3



On the left side are the responses to news shocks, while on the right side are the responses to noise shocks. Each graph reports the percentage response to an initial one-standard-deviation shock over various response horizons. The solid lines represent the median impulse responses, while the shaded gray areas denote the range within plus or minus one standard deviation. TFP, and the variable 'R' are presented in their level form. Inflation (π) and the IBOV index responses are expressed as monthly growth rates. The NILTM index is shown in standard deviations due to its normalization, and IBC-BR responses are depicted as percentages.

Table 7 – Forecast Error Variance Decomposition: Alternative Model 1

	h=1	h=4	h=8	h=16	h=24	h=40
News Shock						
TFPu	0.000	0.152	0.131	0.121	0.125	0.131
NILTM	0.996	0.906	0.890	0.885	0.882	0.879
IBOV	0.398	0.374	0.374	0.373	0.372	0.370
IBC-BR	0.198	0.468	0.438	0.406	0.388	0.373
Inflation	0.004	0.048	0.055	0.059	0.062	0.065
Interest Rate	0.005	0.018	0.050	0.102	0.129	0.148
Noise Shock						
	h=1	**h=4**	**h=8**	**h=16**	**h=24**	**h=40**
TFPu	0.000	0.012	0.018	0.019	0.020	0.022
NILTM	0.000	0.017	0.021	0.021	0.021	0.021
IBOV	0.594	0.562	0.548	0.541	0.538	0.533
IBC-BR	0.015	0.012	0.020	0.023	0.024	0.024
Inflation	0.005	0.017	0.024	0.025	0.026	0.026
Interest Rate	0.003	0.011	0.013	0.018	0.020	0.022
Total IBOV	0.992	0.936	0.922	0.914	0.910	0.903
Total IBC-BR	0.213	0.479	0.458	0.429	0.412	0.397

The letter 'h' refers to the forecast horizon. The numbers denote the proportion of the forecast error variance for each variable that is attributable to our identified news shocks at various forecast horizons. 'Total IBOV' represents the total variance of the IBOV index explained by both news and noise shocks combined. 'Total IBC-BR' indicates the total variance of the IBC-BR indicator explained by both news and noise shocks combined.

Table 8 – Forecast Error Variance Decomposition: Alternative Model 2

	h=1	**h=4**	**h=8**	**h=16**	**h=24**	**h=40**
			News Shock			
TFP	0.000	0.173	0.140	0.121	0.132	0.150
NILTM	0.996	0.922	0.909	0.905	0.902	0.897
IBOV	0.398	0.372	0.371	0.368	0.367	0.365
GDP-M	0.013	0.330	0.322	0.301	0.283	0.266
Inflation	0.004	0.048	0.055	0.059	0.062	0.065
Interest Rate	0.005	0.018	0.050	0.102	0.129	0.148
			Noise Shock			
	h=1	**h=4**	**h=8**	**h=16**	**h=24**	**h=40**
TFP	0.000	0.012	0.018	0.019	0.020	0.022
NILTM	0.000	0.017	0.021	0.021	0.021	0.021
IBOV	0.594	0.562	0.548	0.541	0.538	0.533
GDP-M	0.015	0.012	0.020	0.023	0.024	0.024
Inflation	0.005	0.017	0.024	0.025	0.026	0.026
Interest Rate	0.003	0.011	0.013	0.018	0.020	0.022
Total IBOV	0.992	0.934	0.919	0.909	0.905	0.898
Total GDP-M	0.028	0.342	0.342	0.324	0.307	0.290

The letter 'h' refers to the forecast horizon. The numbers denote the proportion of the forecast error variance for each variable that is attributable to our identified news shocks at various forecast horizons. 'Total IBOV' represents the total variance of the IBOV index explained by both news and noise shocks combined. 'Total GDP-M' indicates the total variance of the GDP proxy explained by both news and noise shocks combined.

Table 9 – Forecast Error Variance Decomposition: Alternative Model 3

	h=1	h=4	h=8	h=16	h=24	h=40
News Shock						
TFP	0.000	0.154	0.120	0.103	0.103	0.104
NILTM	0.996	0.901	0.885	0.880	0.878	0.874
IBOV	0.395	0.374	0.375	0.372	0.369	0.367
IBC-BR	0.181	0.438	0.377	0.312	0.289	0.273
Inflation	0.004	0.039	0.045	0.048	0.050	0.051
Interest Rate	0.004	0.016	0.039	0.066	0.080	0.086
Noise Shock						
	h=1	**h=4**	**h=8**	**h=16**	**h=24**	**h=40**
TFP	0.000	0.009	0.016	0.018	0.019	0.021
NILTM	0.000	0.017	0.021	0.021	0.021	0.021
IBOV	0.596	0.564	0.552	0.544	0.539	0.535
IBC-BR	0.012	0.010	0.018	0.021	0.022	0.023
Inflation	0.004	0.016	0.023	0.024	0.024	0.025
Interest Rate	0.003	0.012	0.014	0.018	0.021	0.022
Total IBOV	0.991	0.938	0.927	0.916	0.908	0.902
Total IBC-BR	0.193	0.448	0.395	0.333	0.311	0.296

The letter 'h' refers to the forecast horizon. The numbers denote the proportion of the forecast error variance for each variable that is attributable to our identified news shocks at various forecast horizons. 'Total IBOV' represents the total variance of the IBOV index explained by both news and noise shocks combined. 'Total IBC-BR' indicates the total variance of the IBC-BR indicator explained by both news and noise shocks combined.

APPENDIX E – BAYESIAN VAR

Following the methodology established by Dieppe et al. (2016)[1], the Bayesian VAR model employed in this study is implemented using the BEAR toolbox, which can be represented as follows:

$$y_t = \alpha + \phi_1 y_{t-1} + ... + \phi_p y_{t-p} + \epsilon_t \tag{E.1}$$

for $t = (1, ..., T)$. Here, $y_t = (y_{1,t}, ..., y_{n,t})$ is a $(n \times 1)$ vector of endogenous variables, $\phi_1, \phi_2, ..., \phi_p$ are $(n \times n)$ parameter matrices, α is $(n \times 1)$ vector of constants, and $\epsilon_t = (\epsilon_{1,t}, ..., \epsilon_{n,t})'$ is a vector of residuals ($\epsilon_t \sim \mathcal{N}(0, \Sigma)$) with $E(\epsilon_t \epsilon_t') = \Sigma$ and $E(\epsilon_t \epsilon_s') = 0$ for $t \neq s$.

This can be expressed in more compact notation as follows:

$$Y = XB + \epsilon \tag{E.2}$$

with

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_T \end{pmatrix}, \quad X = \begin{pmatrix} y_0' & y_{-1}' & \cdots & y_{1-p}' & 1 \\ y_1' & y_0' & \cdots & y_{2-p}' & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ y_{T-1}' & y_{T-2}' & \cdots & y_{T-p}' & 1 \end{pmatrix}, \quad B = \begin{pmatrix} A_1' \\ A_2' \\ \vdots \\ A_p' \\ \alpha' \end{pmatrix}, \quad \text{and} \quad \mathcal{E} = \begin{pmatrix} \epsilon_1' \\ \epsilon_2' \\ \vdots \\ \epsilon_T' \end{pmatrix} \tag{E.3}$$

We can rewrite the model:

$$y = \bar{X}\beta + \epsilon \tag{E.4}$$

with:

$$y = \text{vec}(Y), \quad \bar{X} = I_n \otimes X, \quad \beta = \text{vec}(B), \quad \varepsilon = \text{vec}(\mathcal{E}) \tag{E.5}$$

In this model, ε follow a multivariate normal distribution $\varepsilon \sim \mathcal{N}(0, \Sigma)$, with $\Sigma = \Sigma \otimes I_T$.

E.1 The normal-Wishart prior

In normal-Whishart prior distribution, we assume that both β and Σ are unknown. For β, a multivariate normal distribution is assumed for the prior:

$$\beta \sim \mathcal{N}(\beta_0, \Sigma \otimes \Phi_0) \tag{E.6}$$

where β_0 is $(q \times 1)$ vector, Φ_0 is a $(k \times k)$ diagonal matrix, and Σ is the usual VAR residual variance-covariance matrix, with $k = np + 1$ and $q = nk$.

[1] Please refer to Dieppe et al.(2016) for detailed information.

For β_0, we set values to be around 1 for the first lag coefficients of each variable and 0 for the coefficients of other variables. Φ_0 represents the variance of the parameters for a single equation in the VAR.

Define the variance as:

$$\sigma_{a_{ij}}^2 = \left(\frac{1}{\sigma_j^2}\right)\left(\frac{\lambda_1}{l^{\lambda_3}}\right)^2 \tag{E.7}$$

where σ_j^2 represents the estimated residual variance for variable j within the BVAR framework, which is inferred from separate autoregressive (AR) regressions for each variable, while l denoting the lag considered by the coefficient. λ_1 is the overall tightness parameters, and λ_3 is the scaling coefficient that controls the decay of influence for higher lags.

For constant variable, the variance is defined as

$$\sigma_c^2 = (\lambda_1\lambda_4)^2 \tag{E.8}$$

where λ_4 adjusts the tightness for exogenous variables. We set $\lambda_1 = 0.2$, $\lambda_3 = 1$ and $\lambda_4 = 300$.

For the prior distribution of Σ, an inverse Wishart distribution is considered:

$$\Sigma \sim \mathcal{IW}(S_0, \alpha_0) \tag{E.9}$$

where

$$S_0 = (\alpha_0 - n - 1)\begin{pmatrix} \sigma_1^2 & 0 & 0 & 0 \\ 0 & \sigma_2^2 & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & \sigma_n^2 \end{pmatrix} \tag{E.10}$$

and

$$\alpha_0 = n + 2 \tag{E.11}$$

With these priors in place, the conditional posterior of β is given by:

$$\pi(\beta \mid y) \sim \mathcal{N}(\bar{\beta}, \Sigma \otimes \Phi) \tag{E.12}$$

where

$$\bar{\Phi} = \left[\Phi_0^{-1} + X'X\right]^{-1} \tag{E.13}$$

$$\bar{\beta} = \text{vec}(\hat{B}), \quad \bar{B} = \bar{\Phi}\left[\Phi_0^{-1}B_0 + X'Y\right] \tag{E.14}$$

The conditional posterior of Σ is then characterized as:

$$\pi(\Sigma \mid y) \sim \mathcal{IW}(\bar{\alpha}, \bar{S}) \tag{E.15}$$

where

$$\hat{\alpha} = T + \alpha_0 \tag{E.16}$$

$$\bar{S} = Y'Y + S_0 + B_0'\Phi_0^{-1}B_0 - \bar{B}'\bar{\Phi}^{-1}\bar{B} \tag{E.17}$$

E.2 Cholesky Identification

Consider a reduced-form VAR model where $\varepsilon_t \sim \mathcal{N}(0, \Sigma)$.

$$y_t = \alpha + \phi_1 y_{t-1} + ... + \phi_p y_{t-p} + \epsilon_t \tag{E.18}$$

This model can be expressed as a structural VAR model:

$$D_0 y_t = F + D_1\phi_1 y_{t-1} + ... + D_p\phi_p y_{t-p} + \eta_t \tag{E.19}$$

where $\eta_t \sim \mathcal{N}(0, \Gamma_t)$ and represents a vector of structural innovations.

Define:

$$D = D_0^{-1} \tag{E.20}$$

By premultiplying both sides of the equation by D (interpreted as the structural matrix), we have the following relationship:

$$\Sigma = E(\varepsilon_t\varepsilon_t') = E(D\eta_t\eta_t'D') = DE(\eta_t\eta_t')D' = D\Gamma D' \tag{E.21}$$

In the Cholesky identification approach, we assume that $\Gamma = I$, that is, an identity matrix.

$$\Sigma = DD' \tag{E.22}$$

The objective is to find a lower triangular matrix D that satisfies the equation E.22.