FUNDAÇÃO GETULIO VARGAS
ESCOLA DE ECONOMIA DE SÃO PAULO

João Paulo Zuccoli Tessari

Economic news as data: What's between the lines?

SÃO PAULO
2021

João Paulo Zuccoli Tessari

Economic news as data: What's between the lines?

Dissertação apresentada à Escola de Economia de São Paulo como pré-requisito à obtenção de título de mestre em Economia de Empresas.
Orientador: Marcelo Fernandes

SÃO PAULO
2021

Tessari, João Paulo Zuccoli.

Economic news as data : what's between the lines? / João Paulo Zuccoli Tessari. - 2021.

42 f.

Orientador: Marcelo Fernandes.

Dissertação (mestrado CMEE) – Fundação Getulio Vargas, Escola de Economia de São Paulo.

1. Divulgação de informações. 2. Aprendizado do computador. 3. Volatilidade (Finanças). 4. Macroeconomia. I. Fernandes, Marcelo. II. Dissertação (mestrado CMEE) – Escola de Economia de São Paulo. III. Fundação Getulio Vargas. IV. Título.

CDU 330.101.541

Ficha Catalográfica elaborada por: Isabele Oliveira dos Santos Garcia CRB SP-010191/O
Biblioteca Karl A. Boedecker da Fundação Getulio Vargas - SP

João Paulo Zuccoli Tessari

Economic news as data: What's between the lines?

Dissertação apresentada à Escola de Economia de São Paulo como pré-requisito à obtenção de título de mestre em Economia de Empresas.

Data de aprovação: __/__/____

Banca Examinadora:

Prof. Dr. Marcelo Fernandes
FGV EESP (Orientador)

Prof. Dr. Gustavo Freire
FGV EESP

Prof. Dr. Marcelo Medeiros
PUC-Rio

Agradecimentos

Agradeço ao meu orientador, Marcelo Fernandes, por aceitar conduzir a minha dissertação e pelos importantes conselhos.

Aos membros da banca, Gustavo Freire e Marcelo Medeiros, pela discussão enriquecedora e pelas sugestões.

Aos meus pais e à Julia pelo apoio e paciência.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

Resumo

Agentes econômicos necessitam de informação sobre diferentes aspectos da economia para tomarem decisões de modo apropriado. Entretanto, medidas importantes para acessar o estado atual da economia não se encontram disponíveis instantaneamente. O objetivo do presente trabalho é aplicar a metodologia proposta por Bybee *et al.* (2020) para estimar uma estrutura de tópicos para as notícias econômicas brasileiras. Adicionalmente, medimos a atenção dedicada a cada tópico ao longo do tempo. Com isso, estudamos se as séries de atenção podem ser utilizadas como proxies em tempo real para diferentes indicadores econômicos brasileiros. A base de dados textual abrange o período de Maio de 2000 até Julho de 2020. Os resultados mostram que as séries de tempo extraídas dos artigos econômicos fornecem informações úteis para reconstruir e gerar conhecimento para as séries macroeconômicas no contexto brasileiro ao final do seu período de referência.

Keywords: Análise textual, notícias macroeconômicas, atenção, machine learning, volatilidade

Abstract

Economics agents need information about different aspects of the economy to properly make their decisions. However, important measures to access the economic state are not instantly available. The purpose of this work is to apply the methodology proposed by Bybee *et al.* (2020) to estimate a topic structure for the Brazilian economic news. Also, we measure the attention dedicated to each topic overtime. With that, we study whether the attention time series can be used as a real-time proxy for several Brazilian economic indicators. The textual dataset sample period goes from May 2000 to July 2020. The results show that the time series extract from the texts of the economic articles provides useful information to reconstruct Brazilian macroeconomic variables and generate instantaneous knowledge at the end of the reference period for some of them.

Keywords: Textual analysis, macroeconomic news, attention, machine learning, volatility

List of Figures

Contents

1 Introduction

Gentzkow *et al.* (2019) highlight that new technologies allow a vast amount of text in digital format to be available. Also, they emphasize that the information embedded in texts is a valuable complement to the data in more structured formats and traditionally applied to research. Shapiro *et al.* (2017) point out that information about perceptions on the labor market, prices, interest rates, and other economic factors are easily available in the digital online format. Also, they emphasize that computational advances make it possible to analyze large volumes of text simultaneously.

Many of public and private resources are spent measuring and monitoring information and aspects of the economy (Bybee *et al.*, 2020), since knowledge about the economic situation is important for economic agents to make their decisions properly. With that, methods and tools that allow the process and use of text data as a useful source of information, especially with low or no access cost, are naturally valuable. Accordingly, the number of empirical economic works that use text, in some way, as a source of data is increasing considerably (Gentzkow *et al.*, 2019).

One of the main factors that differentiate text data from other data types commonly used in economics is the very high dimension (Gentzkow *et al.*, 2019). The studies inserted in the growing literature that uses text as a source of data to predict and monitor economic indicators use similar approaches to overcome this fact. Broadly speaking, they use different methods of sentiment analysis, usually counting the appearances of a predefined list of words or using dictionaries like *"The Harvard IV-4"*. They differ from each other mainly by building distinct indicators and using alternative textual data sources. Also, it is worth mentioning that these studies use article texts from the majors US newspapers. Examples of articles that follow this kind of strategy are Fan (2010), Levenberg *et al.* (2014), Tuckett *et al.* (2015), Baker *et al.* (2016) and Shapiro *et al.* (2020).

There are some interesting applications using textual data that seek to tackle this question in the Brazilian context. Nevertheless, compared with the studies just mentioned, they rely on quite distinct textual data sources (e.g, Castro (2016) and Farias *et al.* (2019)). The work that uses the most similar textual data is Azevedo (2017). Still, he uses daily news summaries

instead of directly accessing the news content.

Bybee *et al.* (2020) propose to use a topic model as an alternative approach to overcome the high dimension inherent in text data. Topic models began to be explored recently in the empirical economic research (Bybee *et al.*, 2020). Despite that, there are some compelling applications of them with distinct objectives and contexts from those of the studies discussed here (e.g, Hansen *et al.* (2018) and Hansen *et al.* (2019))

Bybee *et al.* (2020) apply the *Latent Dirichlet Allocation* (LDA) model to reduce the high dimensionality to a scale that is interpretable and rationalizable by humans. They use articles text from *Wall Street Journal* to estimate a topic model and quantify the attention dedicated to each one over time. The attention time series are applied to several applications. Among them, a predictive analysis via vector autoregression and reconstruction of economic and financial indicators.

This work's main contribution is to estimate a topic structure for the Brazilian economic news using unprecedented data. Also, we measure the attention overtime dedicated to each topic to establish quantitative connections between the texts and some numerical variables, widely used to understand the Brazilian economic and financial activity. To this end, we construct models to access contemporary correlation and predictability. Since we developed a web scraper to retrieve online news from a relevant source in the Brazilian context, the article texts we use can be quickly updated. With that, one of our goals is to produce alternative measures to work as a real-time proxy for several traditional Brazilian economic indicators. Such as in Bybee *et al.* (2020), we use the full-textual content of the articles.

Our results present a structure for the economic articles in the Brazilian context. We show that the topic structure that we obtained split the economic articles into broader clusters and then into topics intuitively. We find that the time series extract from the texts of the economic articles provides useful information to reconstruct the macroeconomic variables. These findings reconcile with the results obtained by Bybee *et al.* (2020) for the US economy. Also, we show that the attention series's ability to generate an instantaneous month-end proxy for numerical economic series is, in general, low. However, in some cases, the attention series seems to provide useful instantaneous information at the end of the reference period of the variables.

We also apply the attention series in a attempt to propose a measure of uncertainty based on news texts. For that, we use the implied volatility index for Brazil (Astorino *et al.* (2017)). The results indicate that the measure of uncertainty can only partially address the movements of the implied volatility index. Our historical recovery of the indicator for periods when it is not available can capture some historical events.

The remainder of this paper is organized as follows. Section 2 presents the text preprocessing procedures that we use and the structure of the topic model. Section 3 describes the textual dataset and how we obtained it. In section 4 we discuss the model specification and provide some analyzes of the estimated topics. Section 5 presents the results that we obtain studying the relation between the articles texts and economic data. Finally, section 6 concludes.

2 Text Preprocess and Basic Methodology

The approach of this work largely follows the methodology proposed by Bybee *et al.* (2020). We will start this section by presenting some important definitions for understanding the rest of the article, as well as describing pre-treatments that will be applied to the textual database. The pre-treatments are necessary for representing texts as data and, consequently, making them suitable for statistical approaches, as described by Gentzkow *et al.* (2019).

2.1 Representing Text as Data

To represent text as data, three simplifications are usually performed (Gentzkow *et al.* (2019)): dividing the text into individual documents, reducing the number of linguistic elements and limiting the dependency between the elements within a document.

Each article will be considered as an individual document. Accordingly, the simplification's final result will be the mapping of the news text in numerical vectors that characterize each article. With that, each vector element will represent the count of a language *token* present in that article.

We apply text normalization to reduce the number of linguistic symbols covered. With this procedure, we convert all the text to lower case, and remove elements that are not actually words, such as punctuation and numbers.

In addition, we exclude *"stopwords"* from the analysis. The *"stopwords"* usually include articles and conjunctions and, although they are important for sentence structure, they have little or no meaning in isolation. Also, they contribute little to the distinction of documents (Gentzkow *et al.* (2019)). For that, we use the pre-determined set of *"stopwords"* for the Portuguese language available in the Python library "NLTK".

Finally, we use the *stemming* process, where the word affixes are removed, and their stem replaces the words. For this, the "RSLP Portuguese stemmer" available in the Python library "NLTK" was applied.

The mechanisms described above contribute to reduce the amount of language elements that will be included in the analysis. Therefore, we reduce the dimensionality of the problem and the use of computational capacity. Also, this contributes to obtain a topic model with

more interpretative results (Gentzkow *et al.* (2019)).

To limit the dependence between the elements within the same document, we represent the individual documents through the *"bag-of-words"* method. As a result, word ordering is completely ignored. From the uni-grams originated in this procedure, we constructed bi-grams by pairing each of the *tokens* obtained. Finally, as in Bybee *et al.* (2020), we removed from the analysis linguistic tokens that are present in less than approximately 0.1% of documents.

Consequently, an article is now represented by a vector, the size of which is given by the number of linguistic elements remaining after applying the procedures mentioned above and each entry represents the number of times a given token (uni-gram and bi-gram) is used in that document.

We see that the adoption of the above procedures jointly results in representing the original text through a sparse matrix $\mathbf{w}$. The matrix $\mathbf{w}$ has dimension $D \times V$, where D is the number of news articles and V the number of *tokens* in the vocabulary after reducing the linguistic terms covered. In this perspective, each row of the w_d matrix corresponds to the representation of an article, and each of the individual elements $w_{d,v}$ indicates, as mentioned above, the number of times the term v appears in the document d.

2.2 Latent Dirichlet Allocation

Although the procedures adopted above contribute considerably to reduce the dimensionality, the number of unique elements considered in the vocabulary still makes the dimension of the sparse matrix $\mathbf{w}$ extremely high (Bybee *et al.* (2020))

One of the stages of the present work is to condense each of the news addressed here through a topic model. Accordingly, as in Bybee *et al.* (2020), we apply the Latent Dirichlet Allocation (LDA) model. LDA is an unsupervised learning algorithm, which means, it learns patterns from untagged data. Therefore, there is no prior need for discretionary labels. Each document is summarized by the probabilities of belonging to each of the topics modeled.

To achieve this goal, the LDA model assumes that the vector that counts the quantity of each linguistic expression for a given news item, w_d, is distributed according to a multinomial

distribution:

$$w_d \sim \text{Mult}\left(\Phi'\theta_d, L_d\right) \tag{1}$$

where L_d indicates the scale of the distribution, that is, the number of terms that are present in document d. In addition, we have $\Phi = [\phi_1, \ldots, \phi_K]'$. Hence the kth topic is defined by a vector, ϕ_k, with dimension V (number of unique linguistic terms considered in the vocabulary), with $\phi_{k,v} \geq 0$ for all v and $\sum_v \phi_{k,v} = 1$. With that, we see that a topic consists of a probability distribution over linguistic terms.

Additionally, we have $\theta_d = (\theta_{d,1}, \ldots, \theta_{d,K})'$ with $\theta_{d,k} \geq 0$ for all k and $\sum_k \theta_k \theta_{d,k} = 1$. Then, each document in the model is represented as a probability distribution over the topics. Consequently, it is possible to determine the attention level of each article to each of the topics.

We use the *gensim* library, available for the *Python* programming language to implement the topic model. By doing so, we estimate the model using the procedure proposed by Hoffman *et al.* (2010).

3 Textual Dataset

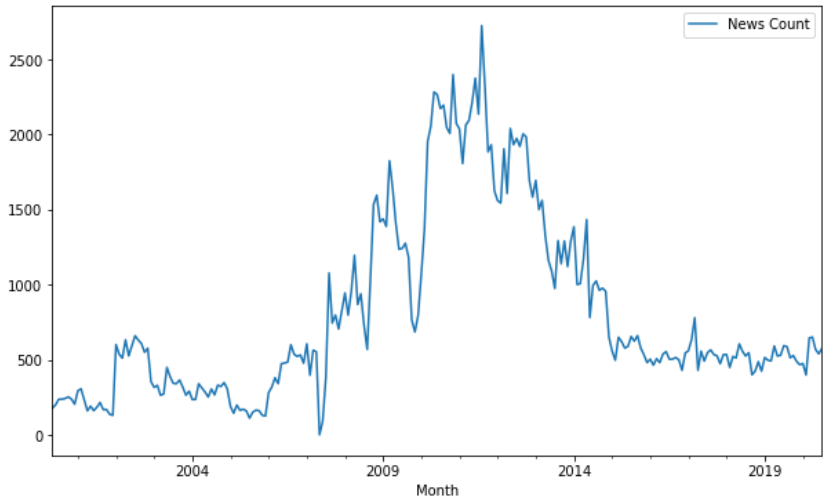
We built the textual dataset applied in this work from scratch. For this purpose, we developed a web crawler and scraper to collect news from the "O Estado de São Paulo" newspaper website. "O Estado de São Paulo" is a major Brazilian newspaper since 1875. The idea was to use textual data with characteristics similar to those present in the news text used by Bybee *et al.* (2020), containing the full content of the news.

To this end, we access the "O Estado de São Paulo" online news search website. The website's content shows a limited amount of news headlines containing an implicit link to access the full text and summary by page sorted chronologically. To extract all the news text for the larger sample period as possible, the developed web crawler will navigate through the pages as if we were clicking on a "load more" button at the bottom of the page. This process was repeated until the last available news was reached.

We collected all the links to access the complete article content. Following this, we accessed each link and scraped the textual content for each page loaded during this procedure . With this, 194,466 news were collected covering the period from May 2000 to July 2020. The search was restricted to the "Economia" section tag.

Figure 1 shows the news count per month for the sample period. We see that for only one of the analyzed months (May 2007), the news count equals zero since there is no article available on the online search website for that month. It is necessary to keep in mind that we may suffer from the absence of news availability for some periods. Notice that the number of articles per month changes over time, either due to availability or changes in news production. However, the topic model addresses the changes in news production by modeling news term proportions rather than term count levels (Bybee *et al.* (2020)).

Figure 1: News count per month



Note: The graph plots the article count each month

Using the collection of news texts obtained and described above, we apply the pre-processing presented in previous sections. With that, we ended up with a dictionary containing 7,855 unique linguistic terms, determining the dimension of the count vector, w_d, that represents each news item. Therefore, the dimension of the matrix $\mathbf{w}$ is given by 194,466 x 7,855.

We developed the crawler and scraper, as well as the procedures for pre-processing the text data, in a way that it is possible to apply both to retrieve new articles and make them suitable for statistical applications.

4 Model Specification and Estimation

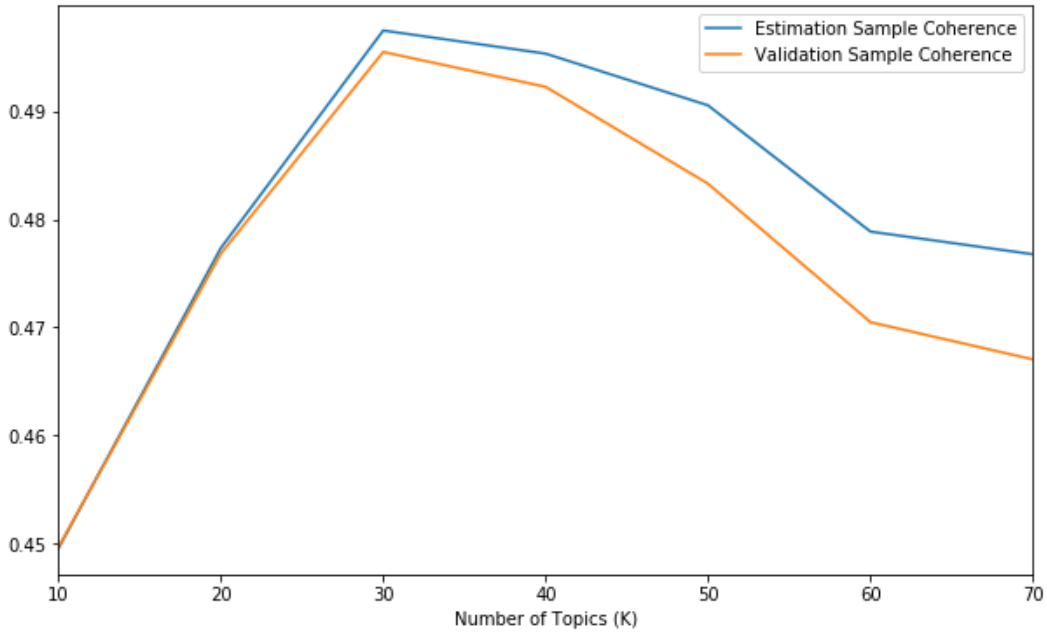
4.1 Determining the Number of Topics

Usually, the held-out documents' predictive likelihood is used to determine the number of LDA models' topics. However, as Chang *et al.* (2009) pointed out, the predictive likelihood of unseen data correlates negatively with human interpretability. We maximize a measure of coherence across words in a given topic to ensure that they somehow relate to each other. Coherence of a set of words measures the hanging and fitting together of single words or subsets of them (Röder *et al.* (2015)). A coherence measure can be seen as a composition of parts. The first part is the segmentation that is used to split the word set into smaller pieces. Then, these pieces are compared against each other using a confirmation measure based on word probabilities. Finally, an aggregation method is applied to the scalar values. In particular, we employ Röder *et al.* (2015) coherence measure Cv given their evidence that it yields the highest correlation with human rating. Intuitively, topic coherence measures the degree of semantic similarity between words in the topic.

We do not want to establish specific topics tailored to the sample of texts used in this work. In particular, the topics generated here should not have extremely specific content regarding certain periods, that is, we do not want to distort the choice of the number of topics by a particular subsample. The structure of topics and words should handle well new articles' texts. Ideally, the methodology proposed here should be easily implemented, applied and updated for new data and different time horizons. We estimate the model considering different numbers of topics, repeating this exercise for aleatory subsamples, always holding 20% of the texts as a validation sample.

Figure 2 plots the mean coherence value across the aleatory subsamples for different model specifications. The analyses suggest that a specification with 30 topics yields the highest coherence. Also, validation samples and text samples used to estimate the models shows similar behavior on average. For most subsamples, individually, the highest level of coherence was also achieved with 30 topics.

Figure 2: Number of topics selection



Note: the graph plots the mean coherence values for all the aleatories subsamples

4.2 Keywords and Attention Over Time

The LDA model determines the topics in an unsupervised manner. With that, we need to identify the subject of each topic and label it accordingly. To this end, we sorted the *tokens* by their relative frequency within the topic compared to the term's general frequency. By doing this, we bring down the importance of terms that are too common among the topics. Using the terms ordered in this way, we are more likely to correctly identify each topic's semantics (Bybee *et al.* (2020)).

Also, we want to measure the attention devoted to each of the topics by the news in a given month. We average the vectors that represent each of the documents, w_d, within that month. Therefore, we define the monthly attention to a topic by its corresponding entry in the average vector. That is, the monthly attention to one topic will be given by the average probability of each document belonging to that topic in that month [1].

[1]Since we have no articles available for May 2007, we use the topics' attention from the month immediately preceding it

4.3 Do these topics make sense?

In this subsection we present some exercises using the results obtained by specifying the LDA model adopting the number of topics equal to 30. We follow some of the applications and visualizations performed by Bybee *et al.* (2020). These exercises are performed to understand better the topics obtained and check if the label assigned to each one makes sense from an economic perspective.

First, we visually inspect the time series for some of the generated topics, to observe if they present behavior consistently, at least, with the Brazilian economy's main facts. We focus on series of attention that we believe to have movements and historical factors that are relatively easy to identify.

Next, we apply the hierarchical clustering method proposed by Ward Jr (1963) to the representative vectors[2] of the topics to verify which topics are closer to each other from the semantic point of view. With that, we can check and understand if the structure obtained can be economically rationalized.

Finally, we use the t-SNE visualization method proposed by Maaten & Hinton (2008) to visualize the articles' disposition in a two-dimensional plane. It is possible, for example, to visualize how the articles distribution in the plane relates to the main topic of each article.

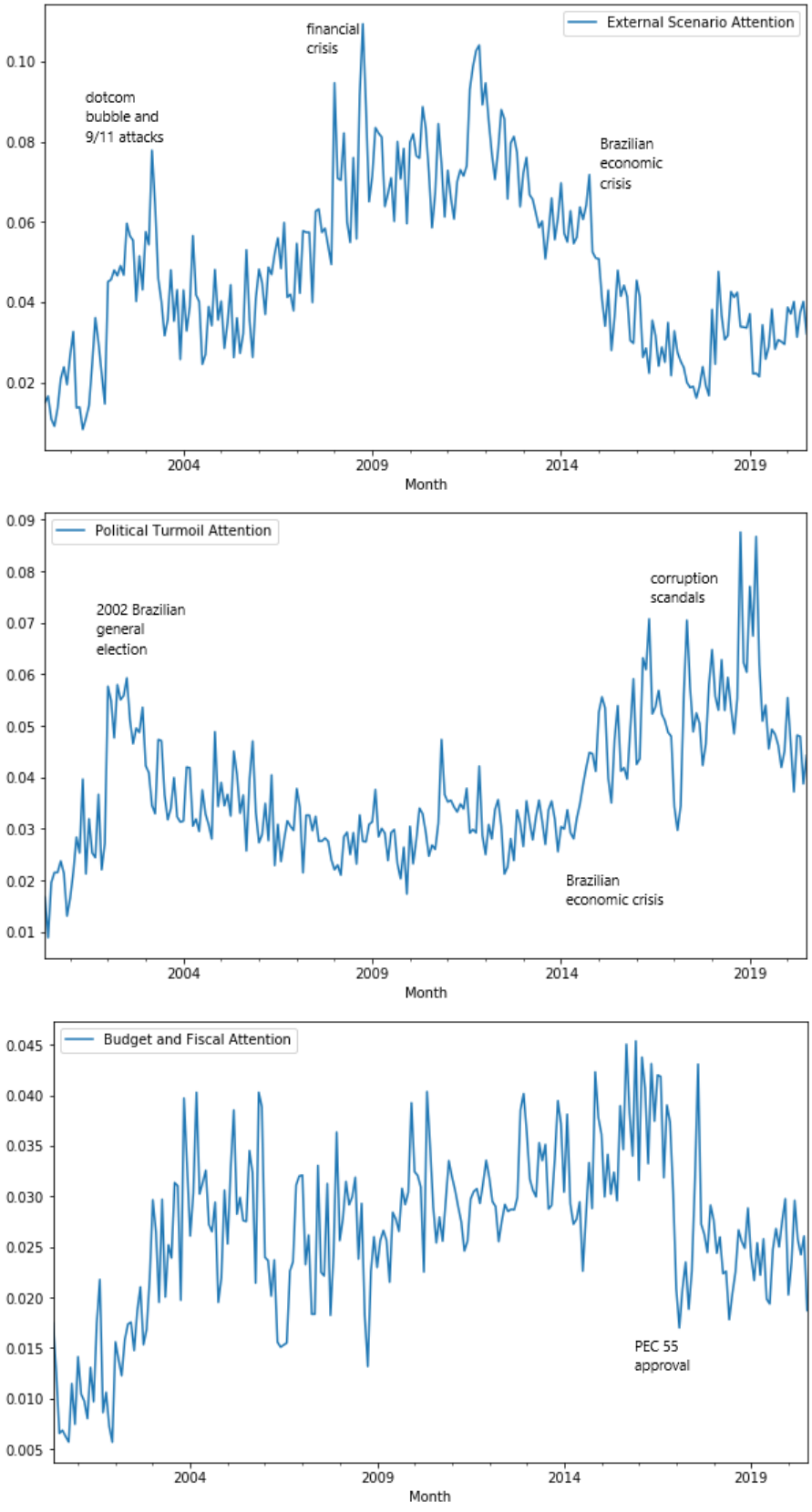
4.3.1 Inspecting the Attention Series

Figure 3 shows the attention time series for the topics "external scenario", "political turmoil" and "budget and fiscal", respectively. The "external scenario" topic presents a upward trend until 2012. During this period, hikes occur around 2001 and 2002, possibly related to the dotcom bubble and the 9/11 attacks. Also, the topic attention spikes in 2008 with the financial crisis and remains high henceforth. With the Brazilian economic crisis in 2014 the "external scenario" topic attention declines. The Brazilian crisis was followed and intensified by a political crisis with a series of corruption scandals. With that, the "political turmoil" topic attention spikes and remains high until the 2019. Furthermore, there is an increase around the general election of 2002. Finally, the "budget and fiscal" attention series present an upward trend until

[2]For the hierarchical clustering and t-SNE visualization we previously applied the "StandardScaler" and "Normalize" functions of the "scikit-learn" library available for the Python language.

2016, when the "PEC 55" constitutional amendment was approved.

Figure 3: Topics attention time series



4.3.2 Topics Structure

Figure 4: Resulting dendrogram from hierarchical clustering

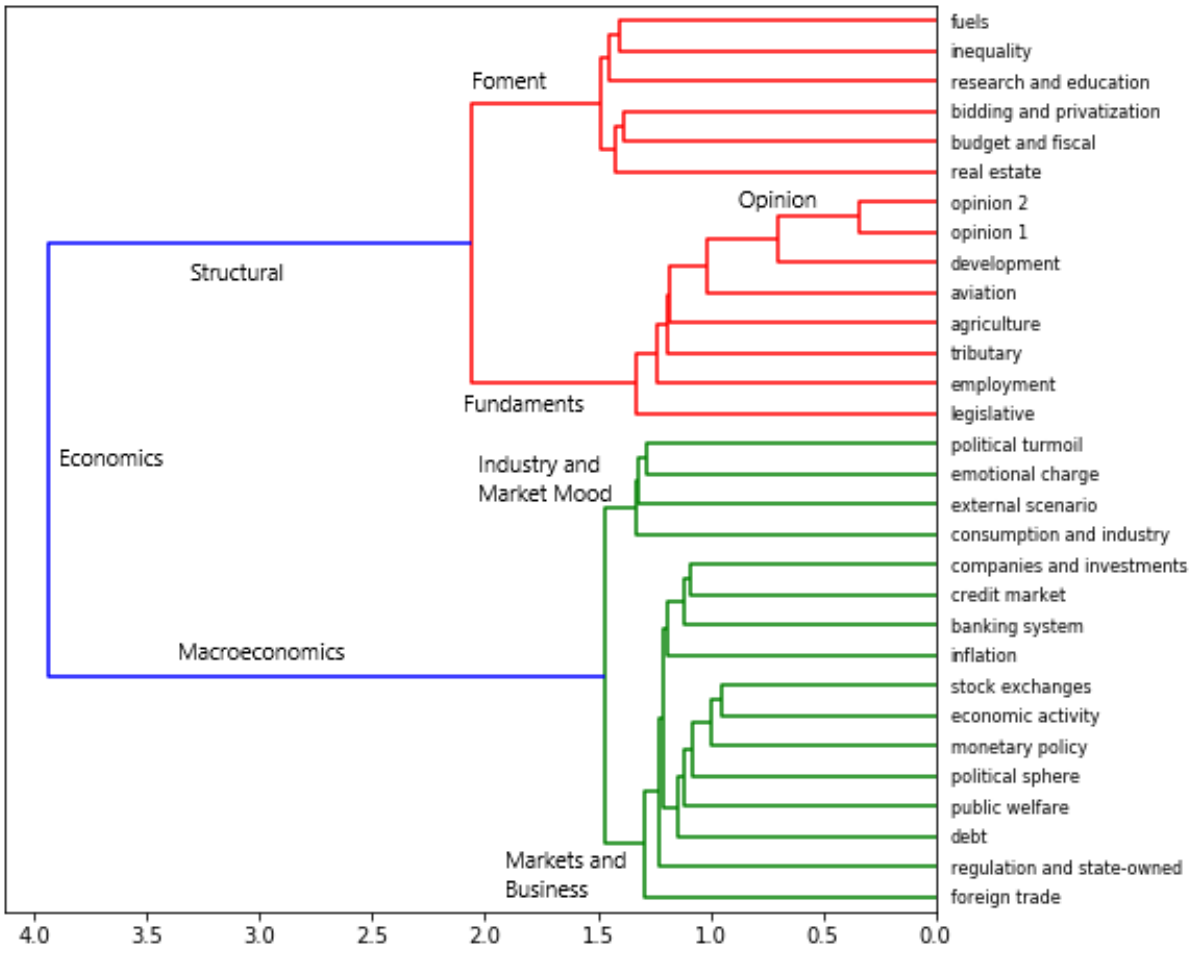


Figure 4 presents the dendrogram representing the results obtained by applying the hierarchical clustering method (Ward Jr (1963)) to the topics representative vectors. The structure of the dendrogram is completely data-driven. We provide the labels to each topic representative vector based on the analysis of the keywords. Also, we identify the broadest level as "Economics" since we only cover articles from the economy section. The topics are split into two branches that we tag as "Structural" and "Macroeconomics".

The "Structural" branch contains the topic cluster "Foment" that includes the "real estate", "bidding and privatization" and "research and education" topics and "Fundaments" cluster that includes topics such as "legislative", "tributary" and "development". The "Macroeconomics" branch also contains two topic clusters and includes topics such as "external scenario", "consumption and industry", "economic activity" and "monetary policy". These

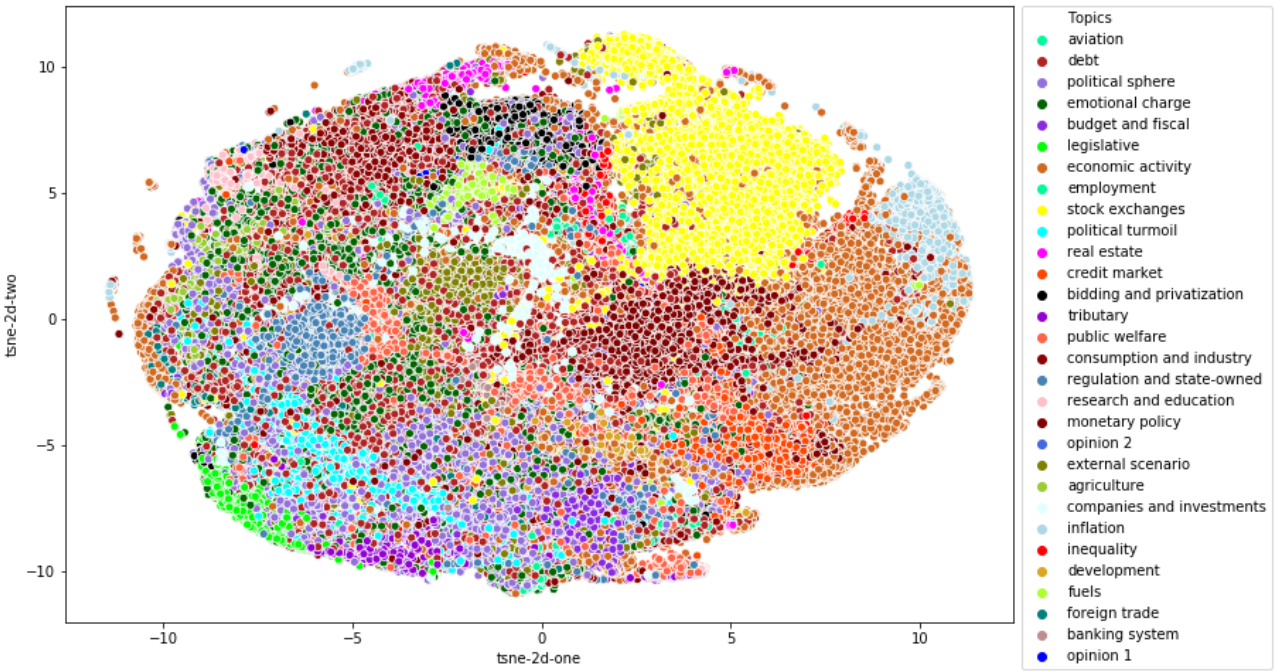
coherent connections and aggregations between the topics provide evidence of the semantic validity of the topic model.

4.3.3 News Level Visualization

The t-distributed stochastic nearest neighbor embedding (t-SNE) technique developed by Maaten & Hinton (2008) allows data visualization at the news level. It is worth remembering that each news text is represented by a vector, w_d, which can be interpreted as a probability distribution over the estimated topics. The t-SNE technique summarize the information contained in the vectors of 30 dimensions representing each of the news in a two-dimensional space. This is done so that articles that have similar high-dimensional vectors have nearby representative points in the plane.

The method is capable of capture local structures of large dimension data and, at the same time, reveal structures of a more global character, such as the presence of clusters (Maaten & Hinton (2008)). We can then analyze how articles that have the same dominant topic are distributed in the two-dimensional space. It is expected that articles that share the same dominant topic tend to co occur spatially in the plane. Therefore, we may observe different clusters being modeled. Also, news with dominant topics which are related with many other topics should appear more disperse, while those that can relate with fewer topics should be more concentrated (Bybee *et al.* (2020)).

Figure 5: t-SNE two-dimensional visualization



Note: Each dot represents a article and is colored according to their dominant topic

We colored each point according to the dominant topic of the article it represents. Observing Figure 5 it is possible to notice a non-random behavior. Dots of the same color tend to occupy nearby spaces. Therefore, in general, articles that share the same dominant topic tend to be represented by similar coordinates. As expected, we can see clusters formed by dots of the same color.

Figure 6 shows the same plane shown in Figure 5, but highlights the color of only one of the topics while the other points are filled in gray. For the individual graphs in Figure 6, we selected topics that we consider to be easy to judge if they relate to many of other topics or only a few.

The first two graphs in Figure 6 show articles which the main topic are “stock exchanges” and “monetary policy”, respectively. Note that both topics can relate to a large number of other topics. With that, it is possible to observe that the colored dots in both graphs are relatively scattered, allowing a relationship with several other topics.

Meanwhile, the graphs placed at the bottom of Figure 6 highlight the articles in which the main topics are “agriculture” and “tributary”, respectively. In this case, the articles seem to occupy a marginal position in the two-dimensional plane and also present a less disperse

behavior, this is probably related to the fact that those topics are narrower, in the sense that few other topics can act as a non-primary subject for these articles

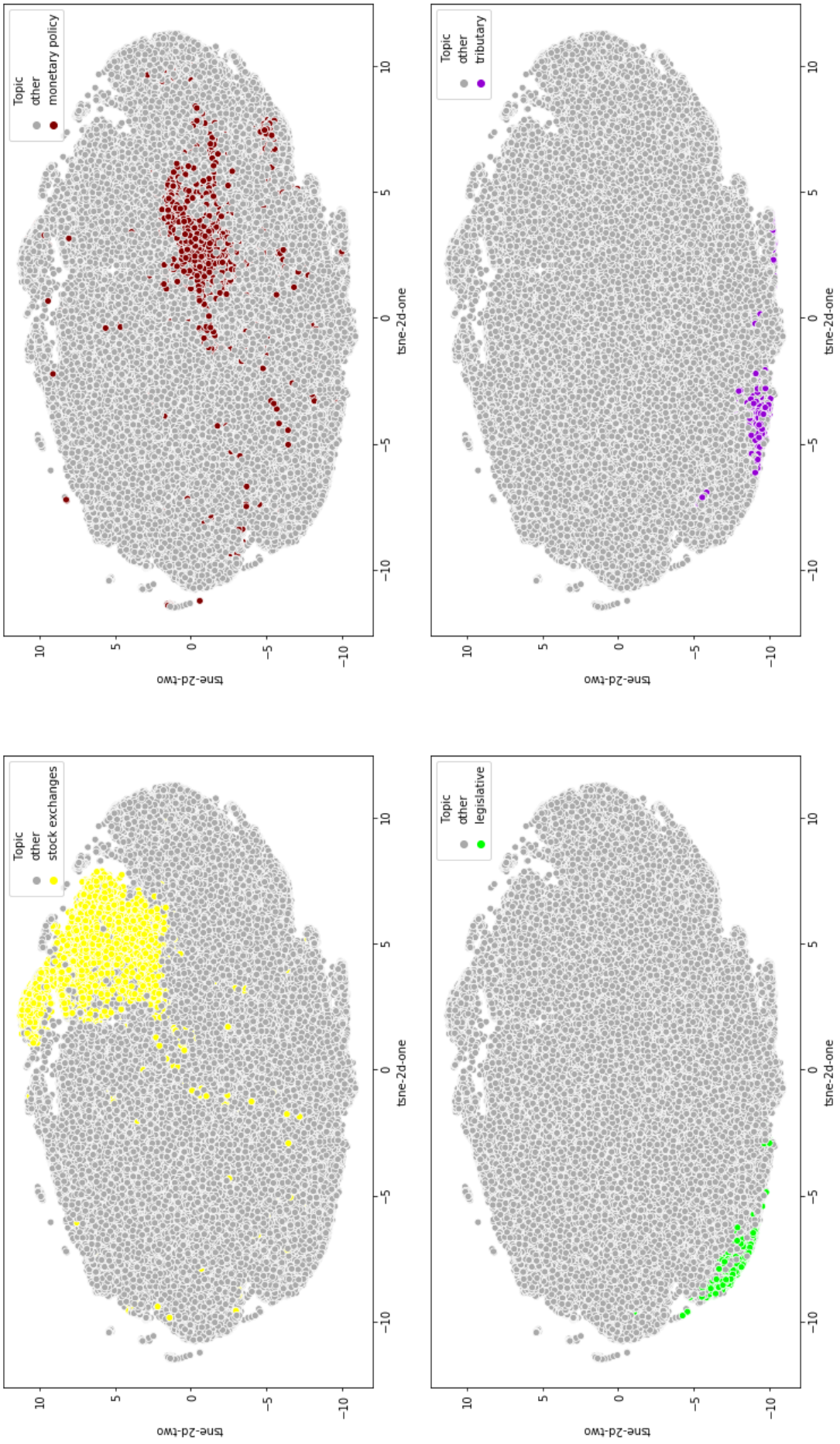


Figure 6: t-SNE visualization by individual topics

5 Results and Applications

5.1 Macroeconomic Data Series

We inspect the extent to which the monthly changes in the topics attention series are related to some Brazilian macroeconomic data series. In particular, we study the CAGED net formal job creation series and the growth of the Monthly Survey of Industry - Physical Production (PIM-PF), Index of Economic Activity of the Central Bank (IBC-Br) and Fundação Getulio Vargas' Consumer Confidence Index (ICC) series [3]. Also, we check the contemporaneous predictive power of the attention series to assess the feasibility of using the attention series to generate an instantaneous month-end proxy for these variables. Even though 30 attention series is not a large amount, our text data starts in May 2000. Since we work with monthly frequency data, using a method that does not penalize the number of variables would be subject to an overfit.

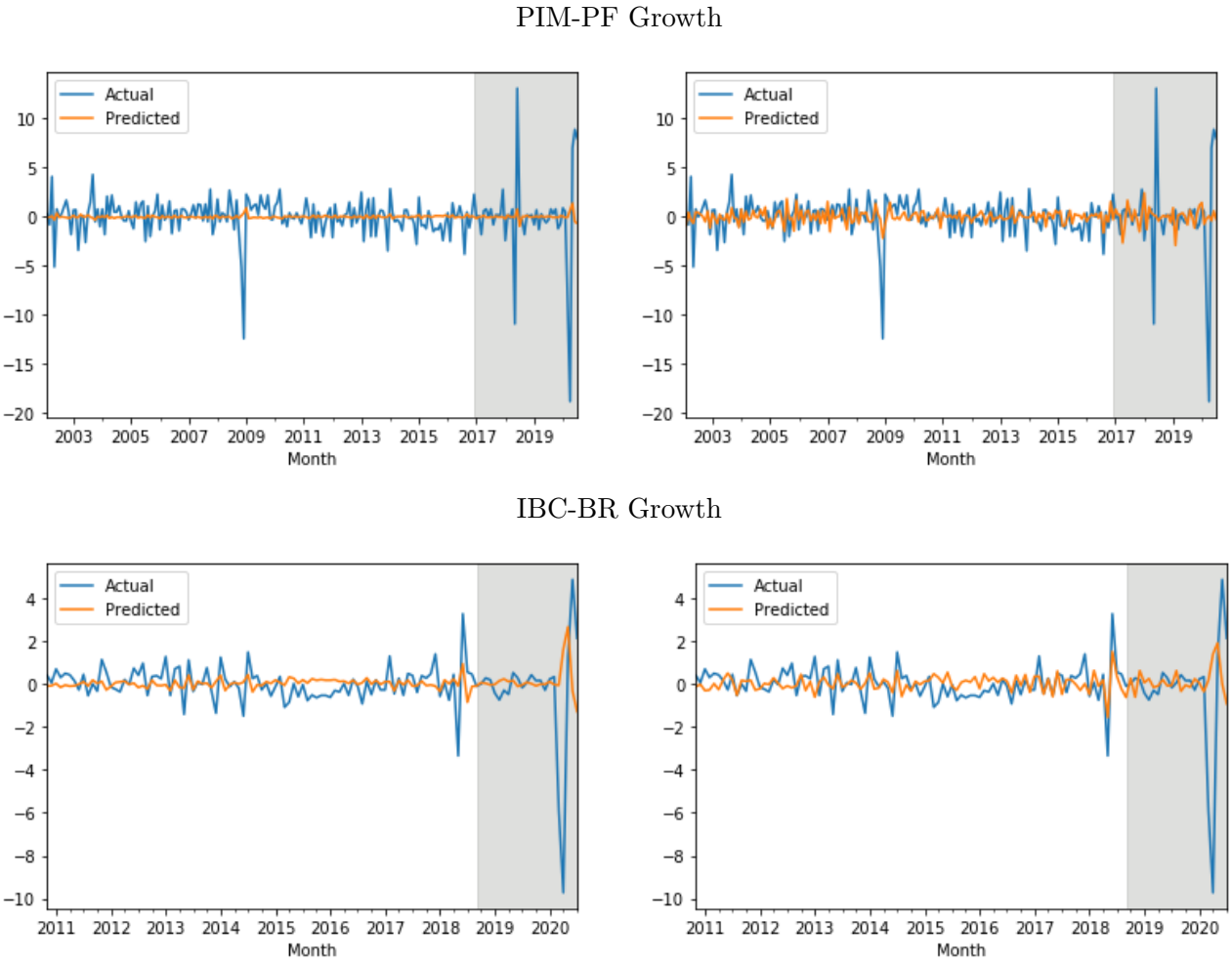
We apply lasso penalization setting the penalization parameter to choose exactly five explanatory variables to address this problem. Besides the attention series, we add an autoregressive term among the variables to be selected. After the five-variable selection, we apply the selected variables in ordinary least squares regressions to reconstruct and predict the macroeconomic series, following Belloni *et al.* (2013). The five-variable approach provides uniformity and comparability across the different macroeconomic data series that we explore (Bybee *et al.* (2020)). As a comparative exercise, we estimate a model considering only the autoregressive component.

The graphs on the left of Figures 7 and 8 show the models' results that use only the autoregressive component as an explanatory variable. The results of the five-variable approach are presented on the graphs placed on the right of the figures. The shaded area in the graphs corresponds to the out-of-sample period [4].

[3] We retrieve the macroeconomic data series from Bloomberg as they were released. With that, we avoid using future information for series that may have been revised. Also, since the ICC series is calculated using the information of the first three weeks of the month, in this case, we use attention series mid-month changes. We seasonally adjusted the CAGED net formal job creation series. We collected the other series in their seasonally adjusted month-over-month percentage changes

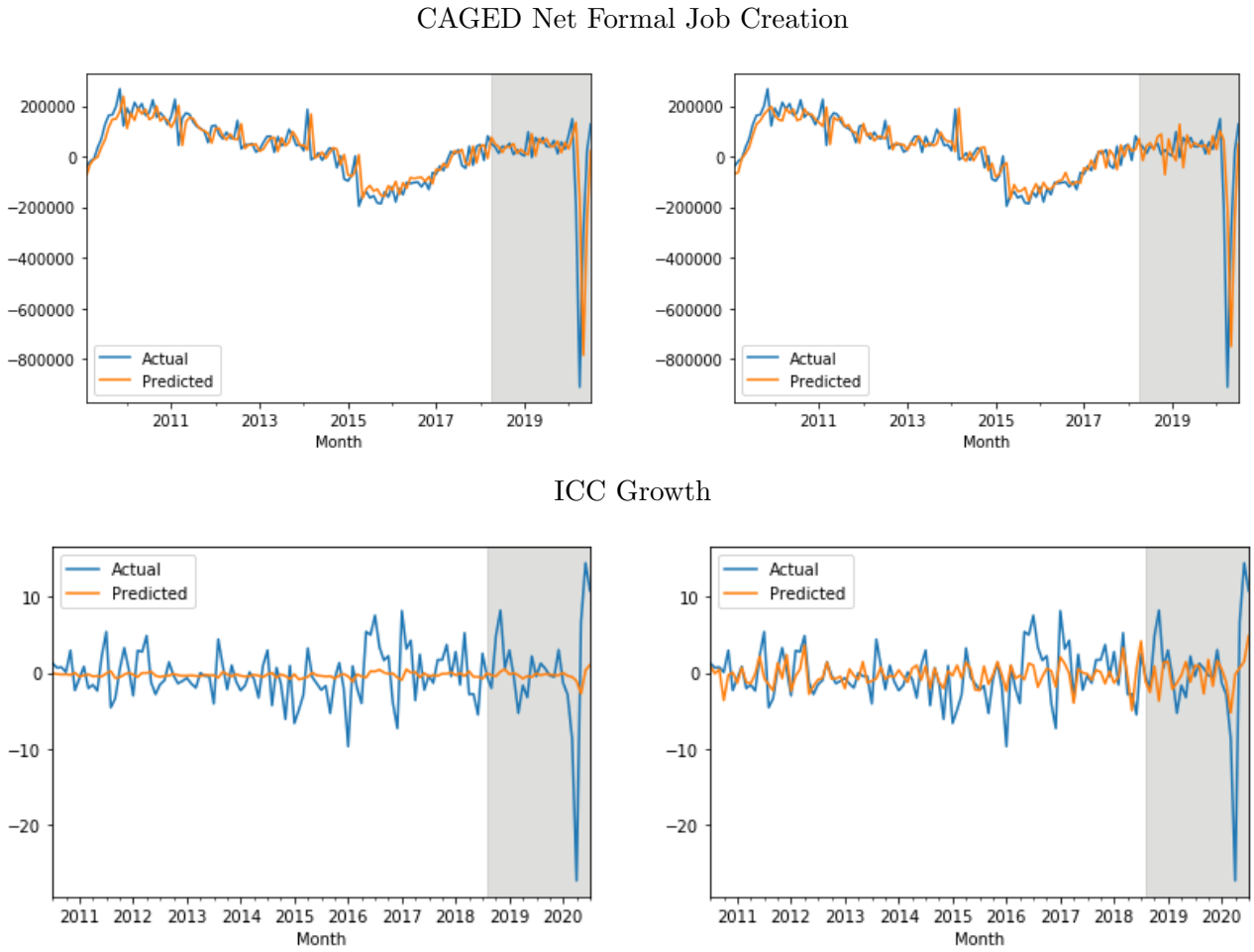
[4] We select 20% of the data for the test samples

Figure 7: Macroeconomic data series results (PIM-PF and IBC-BR)



Note: The panels placed on the left side of the figure show the results of the models that uses only the autoregressive component as an explanatory variable. The panels placed on the right side plot the results of the five-variable approach.

Figure 8: Macroeconomic data series results (CAGED Net Formal Job Creation and ICC)



Note: The panels placed on the left side of the figure show the results of the models that uses only the autoregressive component as an explanatory variable. The panels placed on the right side plot the results of the five-variable approach.

We find that the PIM-PF growth is more closely related with the attention series to "monetary policy","economic activity", "employment", "political sphere", and "legislative". Regarding the in-sample results, the autoregressive component alone explains 0.5% of the 'PIM-PF' growth variability. The five selected topics can explain 13.11% of the variable variance. We find that both models provide less explanatory power in the out-of-sample results than simply fitting a horizontal line.

In the IBC-BR growth case we selected the topics "regulation and state-owned", "fuels", "companies and investments", "opinion 2", and the autoregressive term. The regression containing only the autoregressive term explains 7.31% of the series variability. The five variables model provides an extra explanatory power accounting for 22.88% of the series' variance. Again, in the out-of-sample case, both models are worse than fitting a horizontal line.

The autoregressive component explains a large fraction of the CAGED net formal job creation series variability (76.04%). In this case, we select the attention series to the topics "fuels","political turmoil","inflation", "public welfare", and the autoregressive term. Although the autoregressive component explains a significant portion of the in-sample variability, the topics provide further variability explanation (78.38%). For the out-of-sample period, the autoregressive term explains a very small portion of the series variability (1.51%). We find that the five selected topics contribute significantly to improve the model performance out-of-sample, fitting 10.86% of the series' variability.

For the ICC growth series, we selected the topics "monetary policy", "credit market", "fuels", "companies and investments", and "agriculture". Our results show that the five topics improve the model's ability to track the series for both the in-sample and out-of-sample periods. The model considering only the autoregressive component explains 0.83% of the in-sample variability and 4.02% of the out-of-sample variability. We find that the model with the five selected topics can explain 21.64% and 10.04% of the series variability in-sample and out-of-sample respectively.

The in-sample results indicate that the attention series contain useful information to better reconstructs the macroeconomic series. Combining the attention series with the autoregressive term raised the explained variability for all the macroeconomic series, even in the CAGED

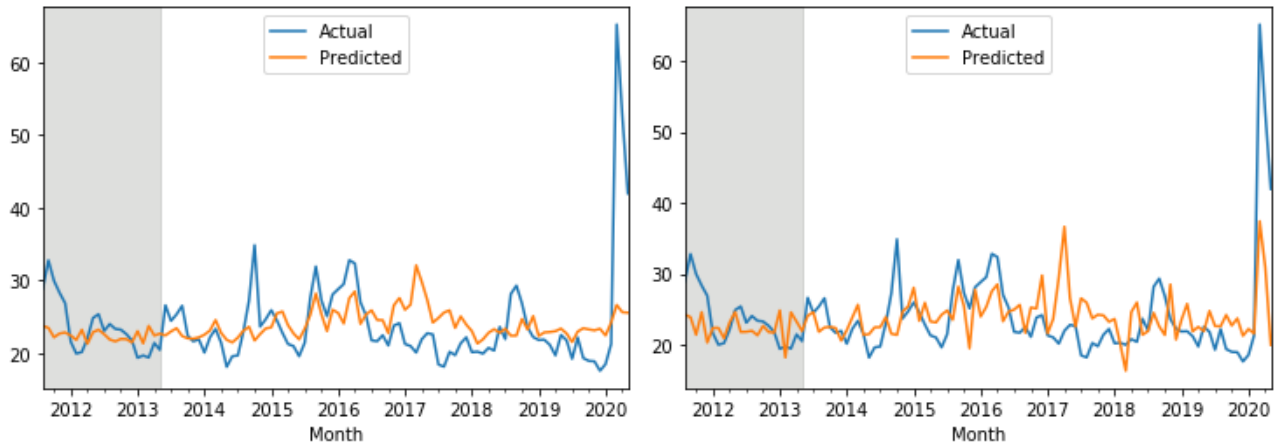
net formal job creation case in which the autoregressive term largely explains the variability alone. With that, our findings reconcile with the results obtained by Bybee *et al.* (2020) for the American economy.

The out-of-sample results show that the ability of the attention series to generate an instantaneous month-end proxy for these variables is, in general, low. Although, for the ICC and CAGED net formal job creation series, we find that the attention series seems to provide useful instantaneous information about these variables. In both cases, the topics' attention series contributed to increase the portion of explained variability. In the PIM-PF and IBC-BR cases, the models perform poorly, especially during the drop in the series associated with the coronavirus outbreak.

5.2 News Implied Volatility

We also apply the five-variable approach to construct a text-based measure of uncertainty, similarly to Manela & Moreira (2017). We use the implied volatility index for Brazil (IVol-BR) proposed by Astorino *et al.* (2017). The IVol-BR data series starts in July 2011. With that, we use the period that the IVol-BR is available to estimate and check the relation between it and the topic attention series. Since our textual data starts in May 2000, we use the estimated dependency to predict the implied volatility in periods that the IVol-BR data is not available. Besides, as the IVol-BR series is daily we use averaged it within a month. Baker *et al.* (2016) show that the Economic Policy Uncertainty (EPU) index for the American economy is highly correlated with VIX, an index of 30-day option implied volatility in the S&P500 index. With that, we select the five variables considering the attention series and the EPU index among the possible explanatory variables. This procedure is similar to that applied for the macroeconomic variables.

Figure 9: IVol-BR results

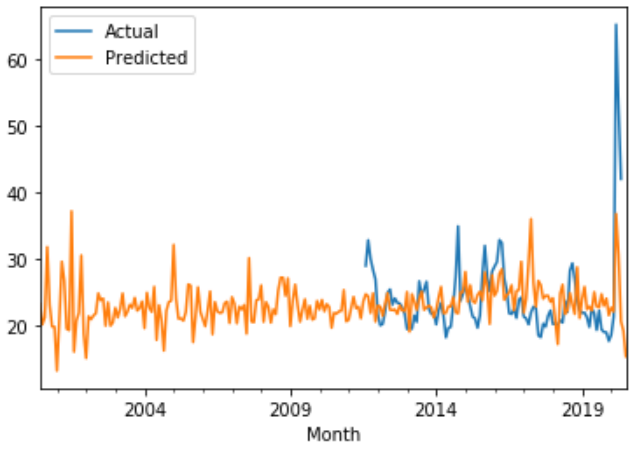


Note: The panels placed on the left side of the figure show the result of the model that uses only the EPU index as an explanatory variable. The panels placed on the right side plot the results of the five-variable approach.

We selected the topics "debt", "companies and investments", "agriculture", "public welfare", and the "EPU" index. The graph on the left side of Figure 9 shows the Ivol-BR and the fitted values using only the EPU index as an explanatory variable. The graph on the right of Figure 9 shows the news implied volatility series generated by us using the selected topics combined with the EPU index. The shaded area represents the out-of-sample period.

The model that explores only the contemporary relationship between the EPU index and the IVol-BR explains a small portion of the series' variability, 8.30% for the in-sample period. The five variables that we select help to explain a larger portion of the IVol-BR variability (21.36%). Some hikes in the IVol-BR series seem to be captured by the model, such as the hike associated with the coronavirus outbreak. However, for example, the volatility at the end of 2015 and beginning of 2016 is not addressed. The out-of-sample results for both models are poor since they perform worse than a horizontal line and do not capture the volatility associated with the euro area debt crisis (Astorino *et al.* (2017)).

Figure 10: Predicted implied volatility



On the other hand, the predicted implied volatility for the period that the IVol-BR is not available (Figure 10) suggests a high level in the series until 2002, in 2004, and between 2007 and 2008. Looking at the historical facts, the period from 2000 to 2002 was marked by the dotcom crisis, the 9/11 attacks, and the Brazilian presidential elections. Meanwhile, in 2004 the Brazilian stock market was affected by the rise in American interest rates, high oil prices, and geopolitical tensions. Also, in the years 2007 and 2008, the global financial crisis happened.

6 Conclusion

This work assesses whether economic articles contain useful information to reconstruct and generate month-end proxies for macroeconomic variables in the Brazilian context. For that, we built a textual data set scrapping online articles from a major Brazilian newspaper. This data set includes the entire textual content for the collected articles. We apply the Latent Dirichlet Allocation model to extract a topic structure implied in the news content. Furthermore, we define and measure the monthly attention that the articles dedicated to each topic. With that, we convert the textual content into data suitable for statistical procedures.

After providing a label for each topic based on its word distribution, we find that the topic structure obtained seems reasonable from the economic point of view. Also, we show that topics' attention time series can recover at least some stylized facts about the Brazilian economy, reinforcing the economic consistency of the topics. We find that the time series extract from the texts of the economic articles provides useful information to reconstruct and generate period-end instantaneous knowledge for the macroeconomic variables. These findings reconcile with the results obtained by Bybee *et al.* (2020) for the American economy.

The attention time series are relatively general and probably can be applied in several other contexts. Accordingly, a natural extension of this work is to investigate whether this series can provide useful information about variables not addressed here. Our tool for collect the texts and the approach for process the articles' content can be used to retrieve and treat new data. With that, one of the main contributions of this work is to establish a topic structure for the Brazilian economic articles since the attention time series associated with it can be applied in future research in several contexts and applications.

References

Astorino, Eduardo Sanchez, Chague, Fernando, Giovannetti, Bruno, & Silva, Marcos Eugênio da. 2017. Variance premium and implied volatility in a low-liquidity option market. *Revista Brasileira de Economia*, **71**(1), 3–28.

Azevedo, Luis Fernando Pereira. 2017. *Impactos econômicos e financeiros de notícias*. Ph.D. thesis.

Baker, Scott R, Bloom, Nicholas, & Davis, Steven J. 2016. Measuring economic policy uncertainty. *The quarterly journal of economics*, **131**(4), 1593–1636.

Belloni, Alexandre, Chernozhukov, Victor, *et al.* 2013. Least squares after model selection in high-dimensional sparse models. *Bernoulli*, **19**(2), 521–547.

Bybee, Leland, Kelly, Bryan T, Manela, Asaf, & Xiu, Dacheng. 2020. *The structure of economic news*. Tech. rept. National Bureau of Economic Research.

Castro, Alexandre Samy de. 2016. Judicial indicators and business cycles in Brazil. *Available at SSRN 2821897*.

Chang, Jonathan, Gerrish, Sean, Wang, Chong, Boyd-Graber, Jordan, & Blei, David. 2009. Reading tea leaves: How humans interpret topic models. *Advances in neural information processing systems*, **22**, 288–296.

Fan, DP. 2010. Predicting the index of consumer sentiment when it isnt measured. *JSM Proceedings, AAPOR,(Alexandria, VA)*.

Farias, Renato, Campelo Junior, Aloísio, & Bittencourt, Viviane Seda. 2019. Social media index as proxy of the Brazilian consumer confidence index.

Gentzkow, Matthew, Kelly, Bryan, & Taddy, Matt. 2019. Text as data. *Journal of Economic Literature*, **57**(3), 535–74.

Hansen, Stephen, McMahon, Michael, & Prat, Andrea. 2018. Transparency and deliberation

within the FOMC: a computational linguistics approach. *The Quarterly Journal of Economics*, **133**(2), 801–870.

Hansen, Stephen, McMahon, Michael, & Tong, Matthew. 2019. The long-run information effect of central bank communication. *Journal of Monetary Economics*, **108**, 185–202.

Hoffman, Matthew, Bach, Francis R, & Blei, David M. 2010. Online learning for latent dirichlet allocation. *Pages 856–864 of: advances in neural information processing systems*.

Levenberg, Abby, Pulman, Stephen, Moilanen, Karo, Simpson, Edwin, & Roberts, Stephen. 2014. Predicting economic indicators from web text using sentiment composition. *International Journal of Computer and Communication Engineering*, **3**(2), 109–115.

Maaten, Laurens van der, & Hinton, Geoffrey. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, **9**(Nov), 2579–2605.

Manela, Asaf, & Moreira, Alan. 2017. News implied volatility and disaster concerns. *Journal of Financial Economics*, **123**(1), 137–162.

Röder, Michael, Both, Andreas, & Hinneburg, Alexander. 2015. Exploring the space of topic coherence measures. *Pages 399–408 of: Proceedings of the eighth ACM international conference on Web search and data mining*.

Shapiro, Adam Hale, Wilson, Daniel J, *et al.* 2017. What's in the News? A New Economic Indicator. *FRBSF Economic Letter*, **2017**, 10.

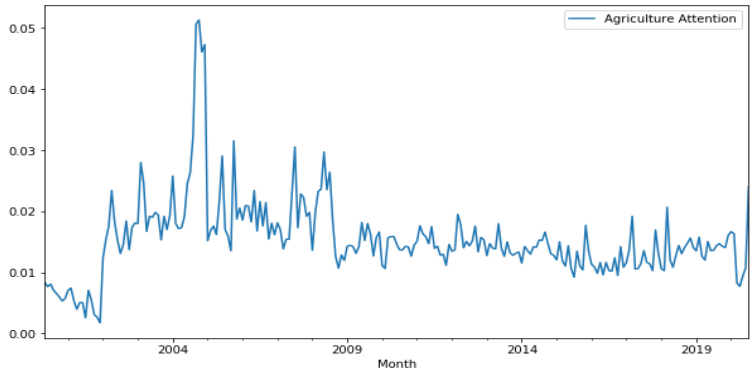
Shapiro, Adam Hale, Sudhof, Moritz, & Wilson, Daniel. 2020. Measuring news sentiment. Federal Reserve Bank of San Francisco.

Tuckett, David, Ormerod, Paul, Smith, Robert Elliott, & Nyman, Rickard. 2015. *Improving economic prediction: A new method for measuring economic confidence and its impact on the evolution of the us economy*. Tech. rept. Working Paper. https://www. researchgate. net/publication/275333686.

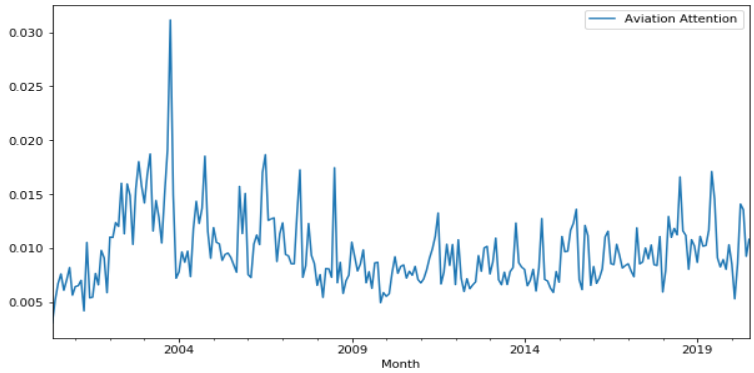
Ward Jr, Joe H. 1963. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, **58**(301), 236–244.

7 Appendix

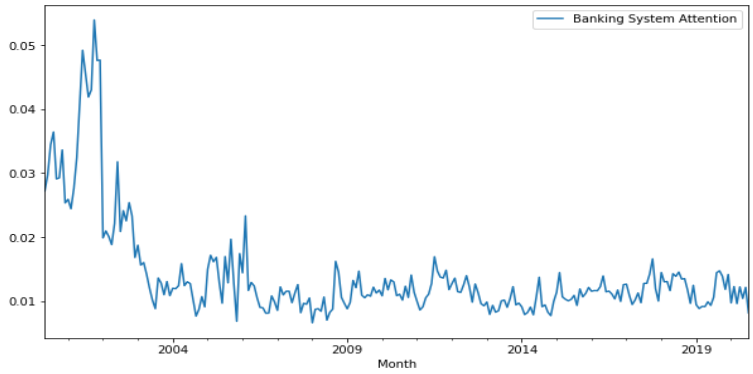
Figure 11: Topics attention time series and keywords



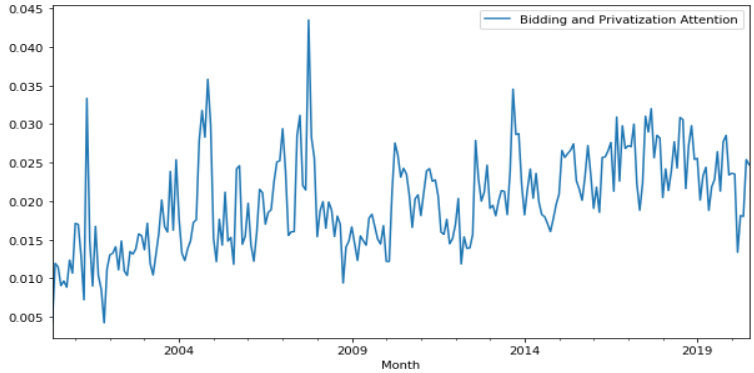
Keywords: “agrícola”, “agricult”, “bid”, “banc interameric”, “grã”, “desenvolv bid”, “ocd”, “cristin”, “algod”, “agronegócio”, “agricul”, “minist agricult”, “amazôn”, “mei ambi”, “organiz cooper”, “agropecu”, “isenç impost”, “mour”, “planti”, “boicot”, “agroindústr”, “sall”, “pecuár”, “brit”, “satisfatór”, “boi”, “embrap”, “hect”, “carbono”, and “agropecuár”



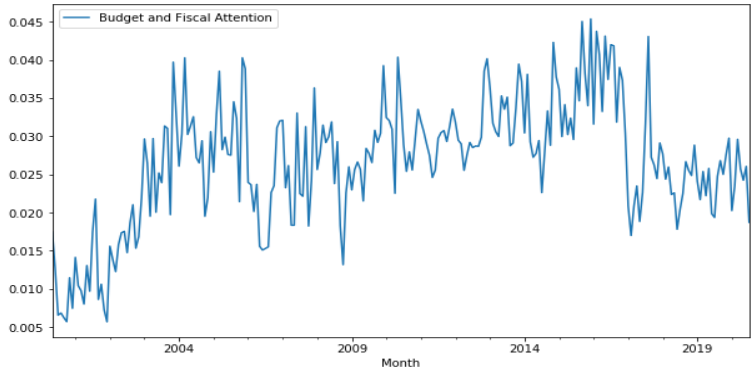
Keywords: “sindicat”, “aére”, “avi”, “passag”, “hug”, “san”, “teleconfer”, “assent”, “gol”, “avi civil”, “aeron”, “lat”, “rotat”, “expir”, “anac”, “azul”, “tripul”, “tarif”, “decol”, “pilot”, “rot”, “bloc”, “aeroporto”, “voo”, “inviabiliz”, “negoci”, “imp”, “categor”, “assim” and “viag”



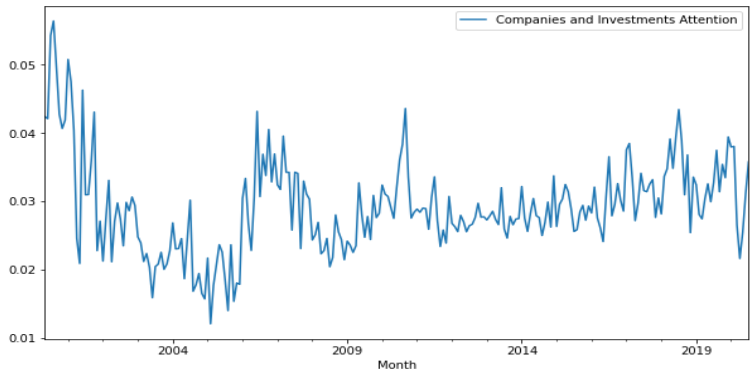
Keywords: “títul”, “bônus”, “resgat”, “bancar”, “solic”, “meirell”, “patrimon”, “centav”, “sacrific”, “itaipu”, “societ”, “intermediár”, “acend”, “indevid”, “fach”, “gall”, “travess”, “domicili”, “emit”, “magistr”, “término”, “emiss”, “perfil”, “tesour”, “atrel”, “aplic”, “fund”, and “cart”



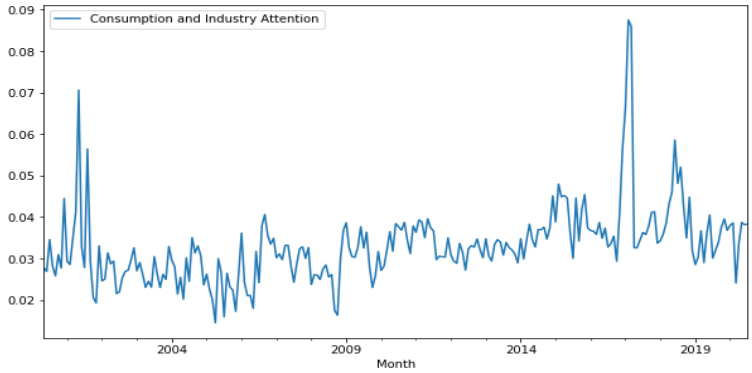
Keywords: “fund pens”, “obr”, “licit”, “sud”, “port alegr”, “min geral”, “br”, “arrend”, “mat gross”, “inici priv”, “bah”, “goi”, “dutr”, “desest”, “pesc”, “rodov”, “sane básic”, “pedágio”, “gasoduto”, “assol”, “águ esgot”, “km”, “revit”, “ferrov”, “condu”, “paran”, “esco”, “marc regulatóri”, “ppp” and “angr”



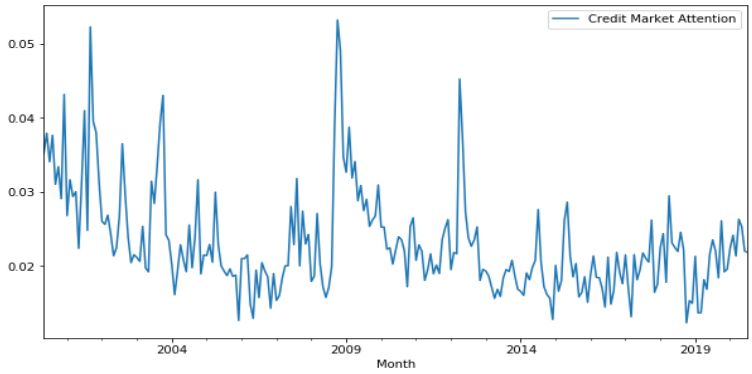
Keywords: “pib”, “déficit”, “produt intern”, “brut pib”, “ajust fiscal”, “orçament”, “lisbo”, “esbarr”, “novel”, “reform previdenciár”, “janel oportun”, “contenç gast”, “tabu”, “direç contrár”, “lup”, “orçamentar”, “desp obrigatór”, “desp discricionár”, “solv”, “superávit”, “burac”, “est municípi”, “romb”, “gast”, “fiscal”, “cresç”, “met”, “reform estrut”, “públic”, and “desp”.



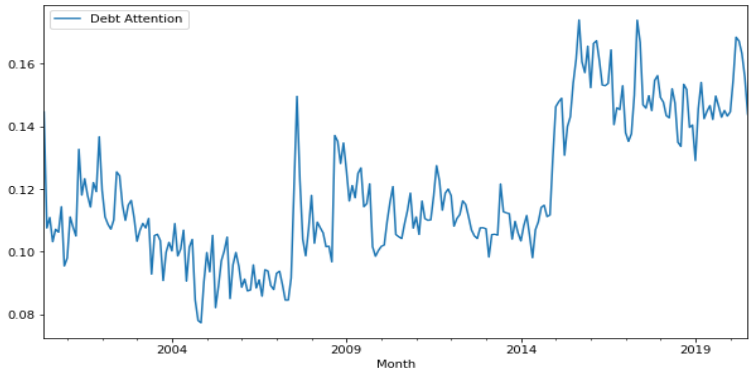
Keywords: “corre”, “ges”, “pn”, “diversific”, “ambev”, “performance”, “verifique”, “desfaz”, “portfóli”, “carrefour”, “unit”, “secov”, “klabin”, “inc”, “pratel”, “suzan”, “invist”, “sinôn”, “equity”, “estre”, “alvar”, “enxut”, “cvc”, “sucumb”, “csn on”, “quadruplic”, “guid”, “olímp”, “fund privat”, and “lucrat”.



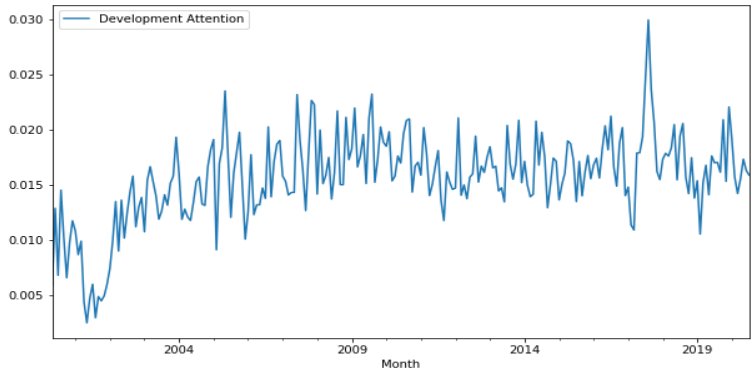
Keywords: “loj”, “veícul”, “fabric”, “eletrôn”, “aparelh”, “supermerc”, “eletroeletrôn”, “marketing”, “fábr”, “roup”, “eletric”, “móvel”, “alme”, “farmác”, “calç”, “automo”, “plás”, “mil tonel”, “passei”, “informa”, “digit”, “térm”, “volkswagen”, “vidr”, “comercializ”, “frot”, “pad”, “alumíni”, “cour”, and “kg”.



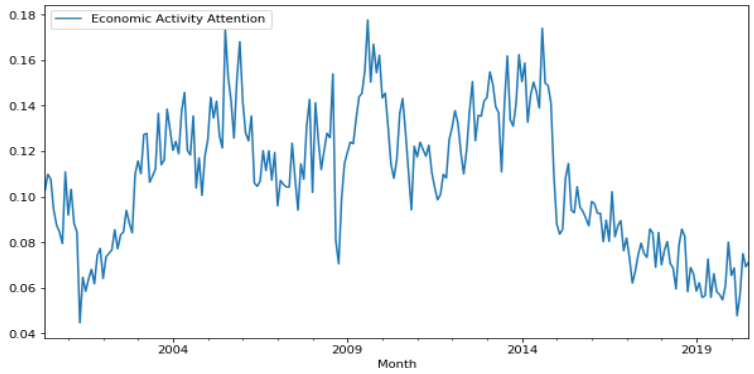
Keywords: “emprést”, “inadimpl”, “saqu”, “cart crédit”, “brad”, “itaú”, “calot”, “bb”, “emprest”, “spread”, “santand”, “rota”, “provision”, “microempr”, “itaú unibanc”, “crédit”, “alien”, “modal”, “contábel”, “banc”, “crédit imobili”, “pesso físic”, “linh”, “cliente”, “institu”, “cobr”, “capt recurs”, “financi”, “brad santand”, and “financ”.



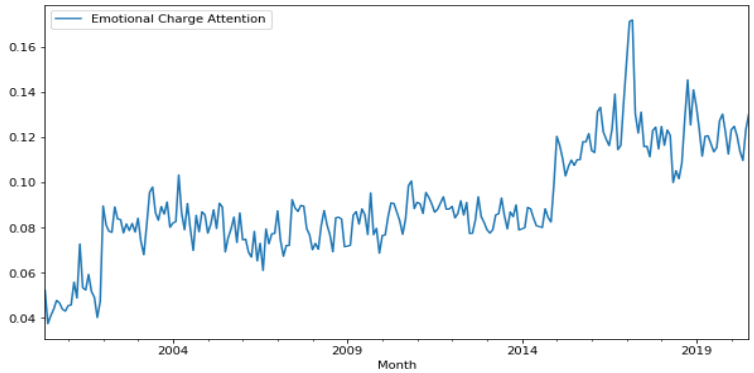
Keywords: “ach”, “cois”, “moratór”, “demor”, “rol”, “aí”, “difícil”, “peg”, “gente”, “insolv”, “acostum”, “saind”, “flávi”, “empurr”, “dig”, “sei”, “imprevis”, “torc”, “fiad”, “planilh”, “perc”, “lá cá”, “aplique”, “mesm proporç”, “surpres”, “aquel époc”, “metrô”, “lastr”, “explod”, and “vim”.



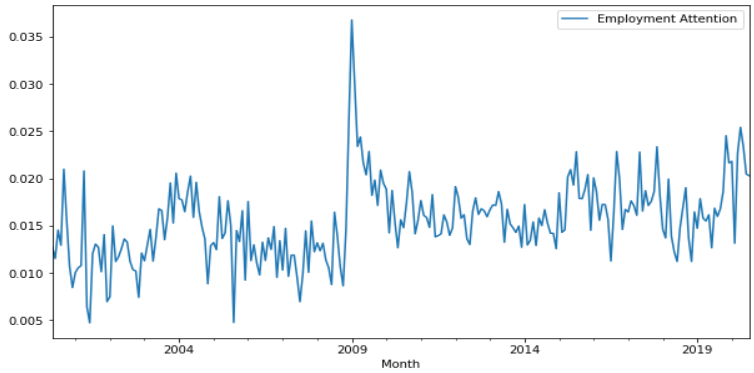
Keywords: “bilhã”, “bnd”, “nacion desenvolv”, “soc bnd”, “csn”, “lucr líqu”, “tjlp”, “patrimôni líqu”, “petrol”, “desinvest”, “bilionár”, “capitane”, “psi”, “desembols”, “provis”, “bi”, “obtev”, “contável”, “totaliz”, “milhã”, “dilu”, “fomente”, “patrimônio”, “capt”, “lucr”, “cifr”, “estatal”, “aloc”, “líqu”, and “gir torn”



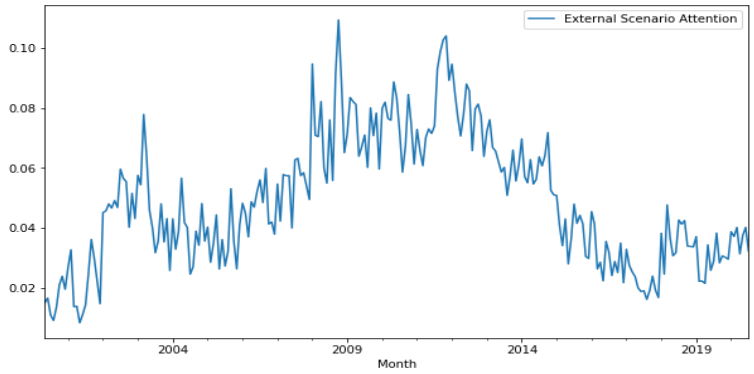
Keywords: “trimestr”, “chequ espec”, “retr”, “dia útel”, “produç industr”, “terc trimestr”, “confianç consum”, “quart trimestr”, “igual períod”, “centr oest”, “sald posi”, “prim quinzen”, “vend varej”, “quinzen”, “sazon”, “mater construç”, “sond”, “carnav”, “sald nega”, “indústr transform”, “bimestre”, “fund poç”, “º trimestr”, “atíp”, “séri histór”, “compar anual”, “ajust sazon”, “comparativ”, “bald”, and “fever”.



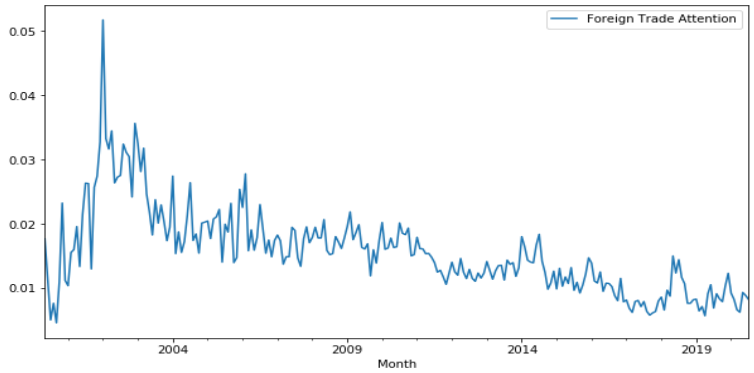
Keywords: “racion”, “lul”, “remédi”, “medic”, “capaz”, “aspect”, “etc”, “rupt”, “naçã”, “neutr”, “corrupç”, “disp”, “impl”, “vál”, “vulner”, “dimens”, “multiplic”, “drama”, “pen”, “examin”, “requ”, “paci”, “engan”, “desaparec”, “insustent”, “borg”, “mau”, “desesper”, “transmit”, and “má”.



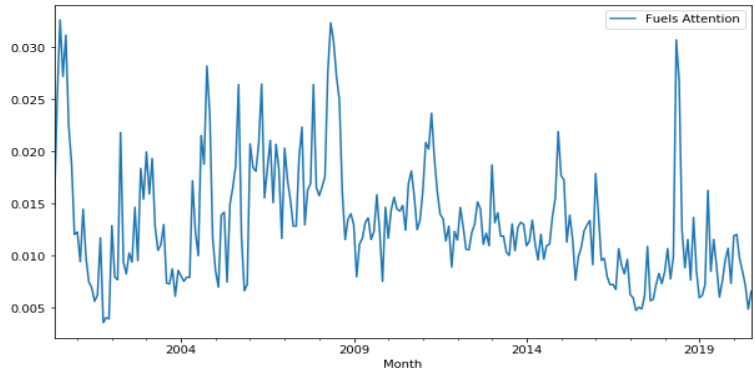
Keywords: “empreg”, “vag”, “post trabalh”, “repôs”, “cart assim”, “desempreg”, “desocup”, “autôn”, “admiss”, “mil vag”, “patrã”, “popul economic”, “salar”, “empreg formal”, “informal”, “férias cole”, “ativ pea”, “bic”, “cadastr geral”, “desempreg caged”, “demiss voluntár”, “incomum”, “caged”, “aloj”, “ral”, “demit”, “segur desempreg”, “nacion amostr”, and “cim azered”.



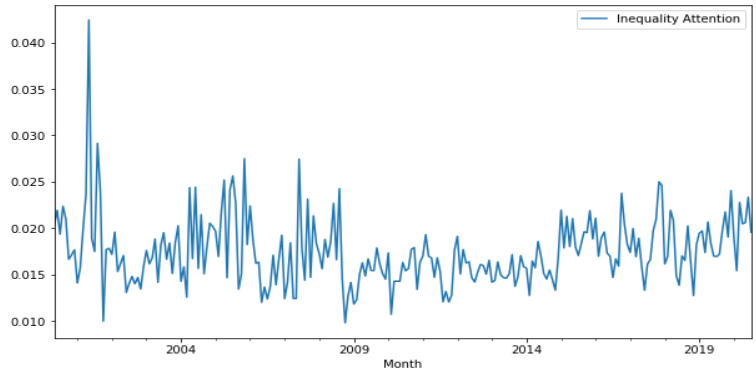
Keywords: “fmi”, “fund monet”, “internac fmi”, “méxic”, “washington”, “chin”, “canad”, “orient médi”, “ásia”, “alemanh”, “espanh”, “rúss”, “inalter”, “bank”, “agênc classific”, “dire gerent”, “israel”, “moody”, “mexic”, “be”, “turqu”, “britân”, “chilen”, “retali”, “sober”, “hong kong”, “paril”, “asiá”, “indonés”, and “cas branc”.



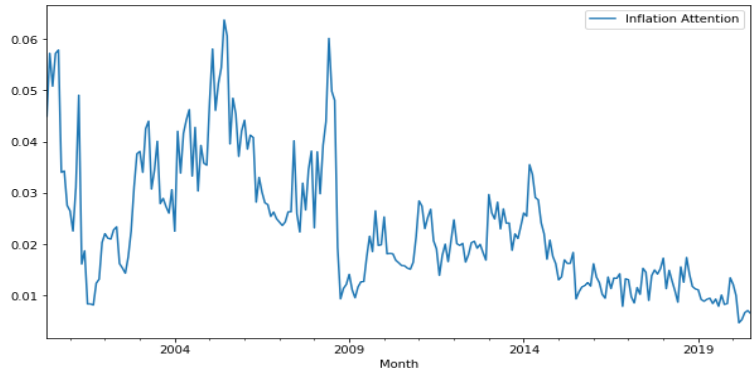
Keywords: “argentin”, “us bilhã”, “cont corr”, “comérci exteri”, “embra”, “companh aére”, “milhã tonel”, “embarque”, “celulos”, “chef gabinet”, “malás”, “reabert”, “camex”, “aeronave”, “jat”, “dad prelimin”, “odebrecht”, “gener motor”, “airlin”, “tap”, “joint ventur”, “delt”, “malh”, “recomeç”, “vo”, “barret”, “prejuíz líqu”, “emir árab”, “tráfeg aére”, and “airbu”.



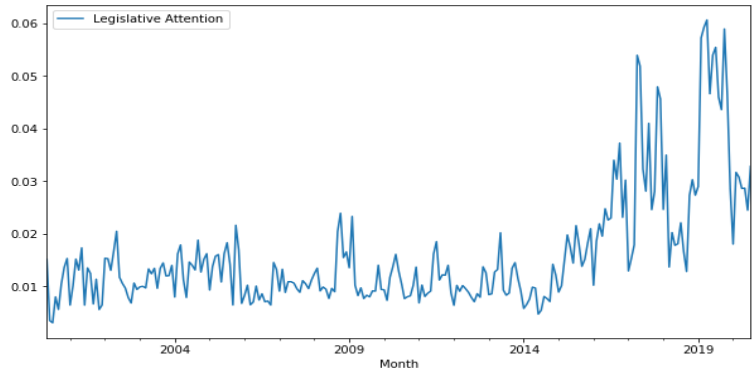
Keywords: “combust”, “gás”, “refin”, “óle”, “grev”, “figueired”, “gás natur”, “cni”, “anp”, “gás co”, “glp”, “cru”, “confeder”, “tav”, “pin”, “ench”, “propíci”, “incógnit”, “óle diesel”, “querosen avi”, “petrolíf”, “leã”, “petróle gás”, “can açúc”, “liquefeit”, “etanol”, “silenci”, “caminhon”, “biodiesel”, and “lun”.



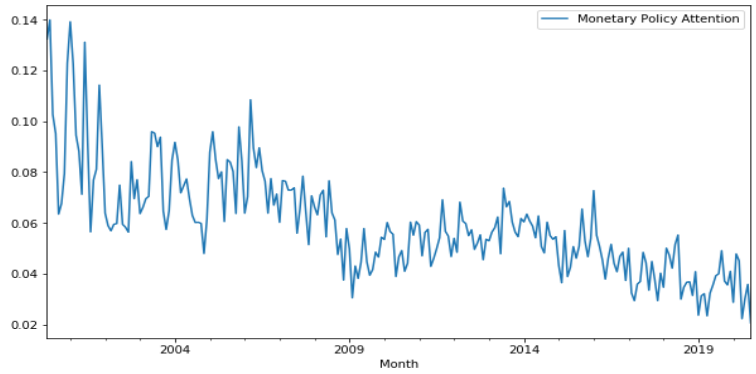
Keywords: “sac”, “salári mín”, “pobr”, “fgv”, “ipe”, “desigualdade”, “furt”, “assist soc”, “rend per”, “violent”, “carente”, “per capit”, “desig”, “regress”, “escalon”, “usual”, “descend”, “ibr fgv”, “lucr divid”, “veloz”, “vivenci”, “envelhec popul”, “depósit”, “famíl”, “rend”, “majorit”, “faix”, “benefíci previdenci”, “poup”, and “indivídu”.



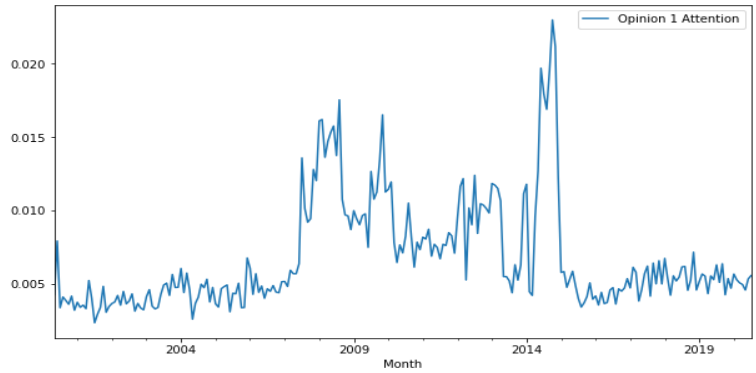
Keywords: “ipc”, “fip”, “consum ampl”, “defl”, “lanterna”, “esti”, “surpreend positiv”, “quantit”, “tempor balanç”, “prév”, “institut pesquis”, “varg fgv” “fund getúli”, “perd fôleg”, “alimente”, “it”, “anal ouv”, “infl”, “índic”, “reajust”, “clima”, “estatís ibg”, “projeç”, “brasil geograf”, “marfrig”, “vari”, “domicíli”, “transport”, “preç”, and “rep”.



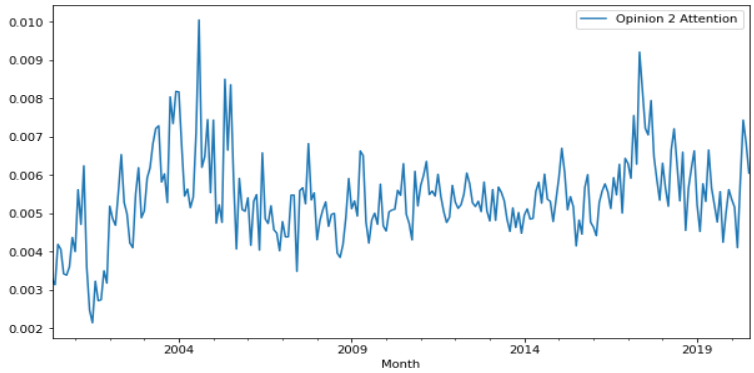
Keywords: “sem”, “congress”, “vot”, “deput”, “parlament”, “projet lei”, “câm deput”, “emend”, “lei complement”, “psb”, “emend constituc”, “rogéri”, “coaliz”, “tecnic”, “pl”, “propost emend”, “incen fiscal”, “tramit congress”, “bezerr”, “pdt”, “plen câm”, “câm”, “pb”, “flexibiliz”, “carimb”, “materi”, “pec”, “desvincul”, “unific”, and “mp”.



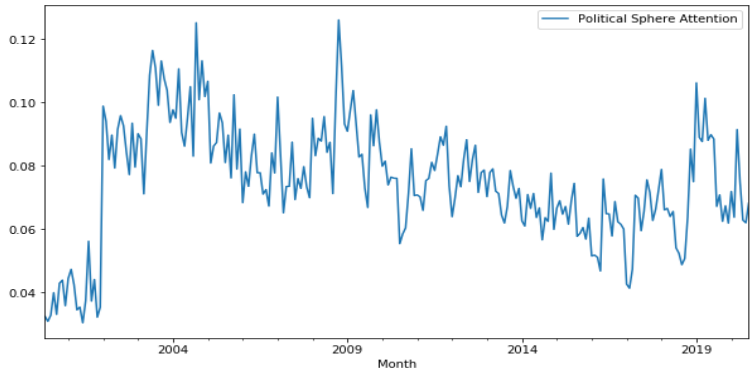
Keywords: “selic”, “copom”, “tax básic”, “polít monetár”, “reun copom”, “luiz fern”, “contrat swap”, “classific risc”, “desenrol”, “atrat”, “déficit nom”, “cmn”, “nacion cmn”, “surj”, “aliv”, “ânc”, “ancor”, “baliz”, “colchã”, “afet negativ”, “critério”, “preponder”, “gatilh”, “spread diferenç”, “indefinid”, “previsibil”, “remun”, “antev”, “timid”, and “pipoc”.



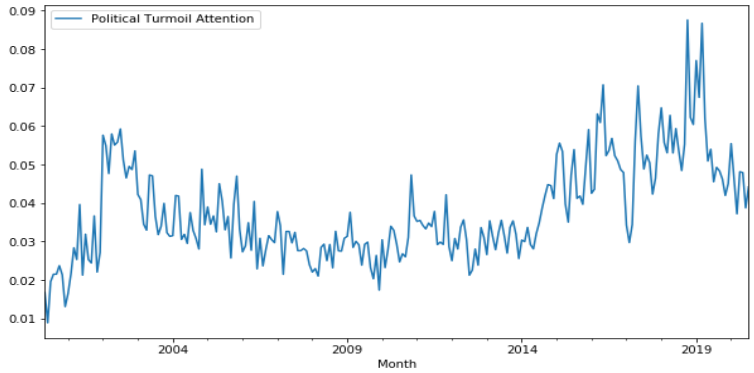
Keywords: “celul”, “vié baix”, “report”, “magalh”, “guim”, “marin”, “roch”, “cent”, “febraban”, “comparec”, “conservador”, “buen”, “chec”, “holding”, “maur”, “reav”, “turm”, “autarqu”, “estági”, “gabriel”, “person”, “dispers”, “adjunt”, “consul”, “fernand”, “membr conselh”, “credenc”, “esport”, “penúlt”, and “ª pos”.



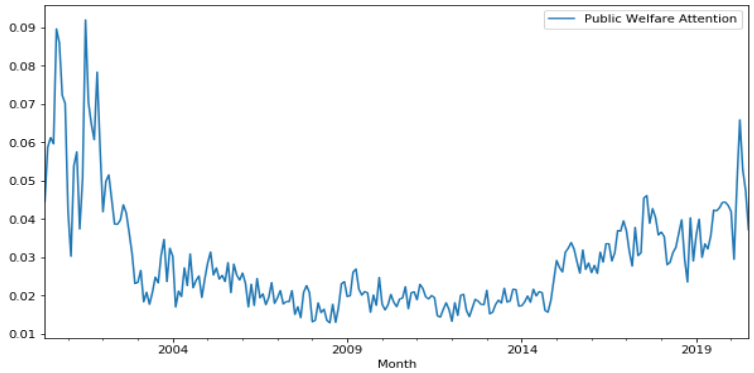
Keywords: “oliv”, “revist”, “fábi”, “lim”, “net”, “antoni”, “mari”, “albert”, “fiat”, “josé robert”, “vá”, “paul fiesp”, “portugu”, “gilbert”, “guilherm”, “andrad”, “propagand”, “august”, “alemã”, “pernambuc”, “mend”, “walt”, “pinh”, “lacerd”, “helen”, “ultr”, “sergi”, “mendonç barr”, “mart”, and “mesquit”.



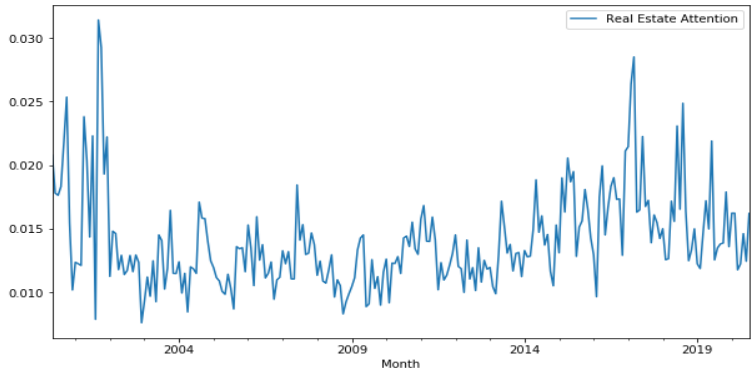
Keywords: “reiter”, “freit”, “repórt”, “cid”, “entrev cole”, “leal”, “pagu”, “redig”, “paláci alvor”, “gued”, “michel tem”, “tecl”, “reformul”, “moder”, “suced”, “destrav”, “diss”, “almoç”, “secret tesour”, “sér”, “envi congress”, “neg”, “ministr”, “maranh”, “legít”, “afirm”, “extraordinár”, “nest terç”, “nest quint”, and “cerimôn”.



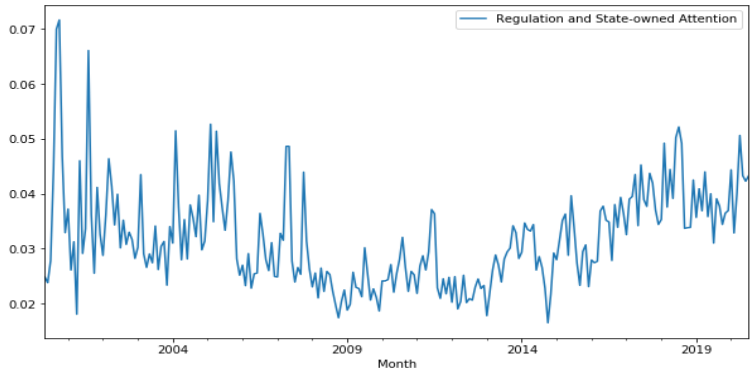
Keywords: “eleitor”, “opôs”, “pt”, “investig”, “armíni frag”, “coll”, “embaix”, “presid repúbl”, “açõ ordinár”, “convoc”, “minorit”, “sindic”, “tribun”, “queix”, “chancel”, “repúbl”, “eleit”, “cartilha”, “minist justiç”, “itamaraty”, “inquérito”, “comand”, “ileg”, “pov”, “renunci”, “supost”, “albuquerqu”, “araúj”, “cas civil”, and “secretár”.



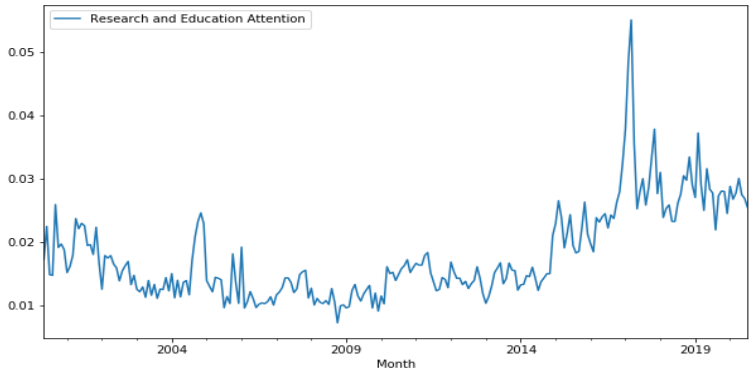
Keywords: “cheque”, “quit”, “receit feder”, “juiz”, “débit”, “renegoci”, “abon salar”, “devoluç”, “ins”, “sentença”, “reembols”, “audit”, “ipv”, “irregul”, “bloquei”, “auxíli”, “advocac geral”, “suplemente”, “uni agu”, “ved”, “procurad geral”, “plaus”, “tcu”, “de”, “equal”, “duplicat”, “spe”, “extingu”, “rubr”, and “capit gir”.



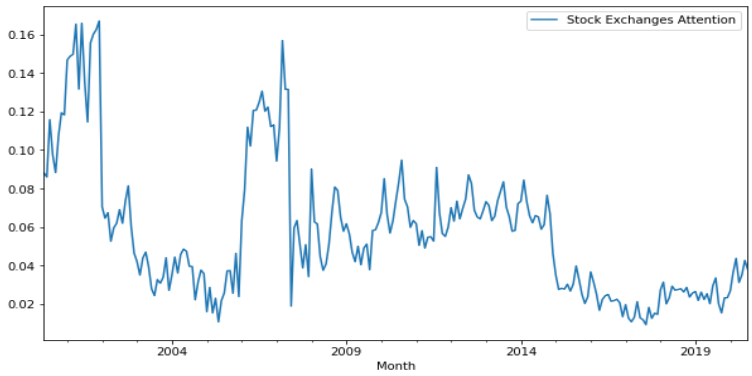
Keywords: “imóvel”, “aluguel”, “lop”, “condomínio”, “loc”, “apart”, “viaj”, “percent”, “hotel”, “além”, “reg metropolit”, “férias”, “dormitório”, “vil”, “bairr”, “paulist”, “imobiliár”, “prédi”, “cor”, “edifíci”, “secov sp”, “aven”, “gom”, “metr quadr”, “morad”, “shopping cent”, “sindicat habit”, “alug”, “arquitet”, and “franqu”.



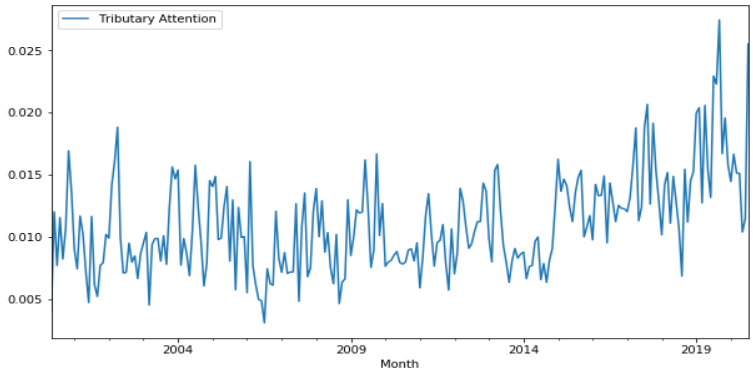
Keywords: “solicit”, “correi”, “conselh administr”, “aneel”, “stf”, “telefôn”, “inden”, “cláusul”, “protocol”, “eletrobr”, “nº”, “cont luz”, “irregular”, “mora”, “unanim”, “estatut”, “tent convenc”, “prévi”, “termelétr”, “obrigatór”, “amor”, “clt”, “criv”, “júni”, “segurança juríd”, “andre”, “mac”, “advog”, “desburocr”, and “coleg”.



Keywords: “sit”, “univers”, “escol”, “laboratório”, “educ”, “usp”, “preench”, “médic”, “genér”, “unid”, “hom”, “aul”, “pal”, “barb”, “seleç”, “cânc”, “vie”, “mulh”, “lar”, “faix etár”, “étic”, “negr”, “concurs”, “voz”, “infantil”, “gêner”, “passaporte”, “process sele”, “feminin”, and “cens”.



Keywords: “nasdaq”, “ibovesp”, “on”, “hsbc”, “barril”, “us barril”, “realiz lucr”, “madr”, “milhã barril”, “estrateg chef”, “sul core”, “exchang”, “frankfurt”, “blu chip”, “bols londr”, “temer”, “ptax”, “petrobr pn”, “nymex”, “brad pn”, “mercantil exchang”, “brent”, “terren posi”, “seul”, “sul afric”, “ice”, “on pn”, “dow jon”, “gerent compr”, and “nov york”.



Keywords: “contribuint”, “tribut”, “impost rend”, “alíquota”, “isent”, “isenç”, “pi pasep”, “restitu”, “incid”, “icm”, “tributar”, “provisor”, “reform tributár”, “encarg”, “repique”, “incid sobr”, “iof”, “carg tributár”, “deduç”, “irpf”, “cofim”, “subsecret”, “deduz”, “folh pag”, “fisc”, “pi cofim”, “alíquot impost”, “iva”, “produt industri”, and “ipi”.