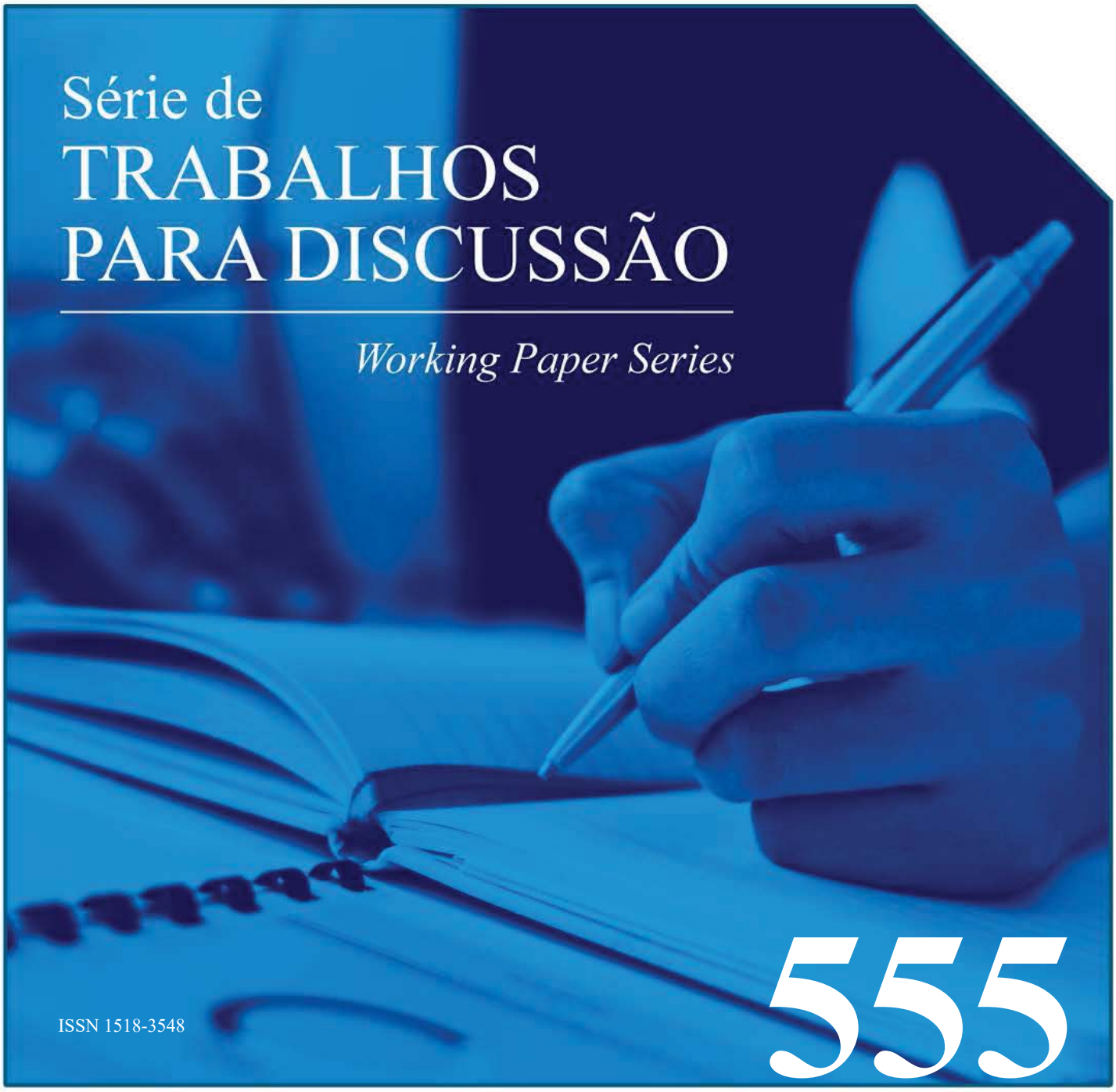


AUGUST 2021

Seeing the Forest for the Trees: using hLDA models to evaluate communication in Banco Central do Brasil

Angelo M. Fasolo, Flávia M. Graminho, Saulo B. Bastos

BANCO CENTRAL DO BRASIL

ISSN 1518-3548
CGC 00.038.166/0001-05

Working Paper Series	Brasília	no. 555	Agosto	2021	p. 1-35

Working Paper Series

Edited by the Research Department (Depep) – E-mail: workingpaper@bcb.gov.br

Editor: Rodrigo Barbone Gonzalez

Co-editor: José Valentim Machado Vicente

Head of the Research Department: André Minella

Deputy Governor for Economic Policy: Fabio Kanczuk

Citizen Service Division

Banco Central do Brasil

Deati/Diate

SBS – Quadra 3 – Bloco B – Edifício-Sede – 2º subsolo

70074-900 Brasília – DF – Brazil

Toll Free: 0800 9792345

Fax: +55 (61) 3414-2553

Internet: http//www.bcb.gov.br/?CONTACTUS

Non-technical Summary

This paper uses computational linguistic techniques to analyze the content and tone of the statements and minutes of the Monetary Policy Committee (Copom) of the Banco Central do Brasil. Sentiment indexes that measure the perception of the Copom on inflation, economic activity and uncertainty are built based on dictionary methods applied to a hierarchical Latent Dirichlet Allocation (hLDA) model combined with feature selection techniques.

The hLDA model allows for an endogenous selection of the number of topics and organizes results in a tree, which favors interpretability and provides relations between themes without previous intervention by the researcher. The use of feature selection as a preliminary step assures that every topic in the tree contain meaningful words that allow for proper document analysis. Sentiment indexes are then constructed based on the frequency of "positive" and "negative" words in each theme, according to predefined dictionaries.

Sentiment indexes are compared with actual measurements of inflation, inflation expectations, economic activity and uncertainty. Results show that the Copom's documents accurately reflect the state of the economy, although this correlation varied over time, with significant changes after July 2016. The coherence between statements and minutes after July 2016 is evaluated. We find that these documents usually deliver the same message, with minutes presenting detailed information first offered in statements.

Sumário Não Técnico

Este artigo utiliza técnicas de linguística computacional para analisar o conteúdo e o tom dos comunicados e atas do Comitê de Política Monetária (Copom) do Banco Central do Brasil. São construídos índices de sentimento que medem a percepção do Copom em relação à inflação, à atividade econômica e à incerteza, baseados em métodos de dicionário aplicados a um modelo de *hierarchical Latent Dirichlet Allocation* (hLDA) combinado com técnicas de *feature selection*.

O modelo hLDA permite a seleção endógena do número de tópicos e organiza os resultados em uma árvore, o que favorece a interpretação e fornece relações entre temas sem necessidade de intervenção prévia pelo pesquisador. O uso de *feature selection* como passo preliminar assegura que todos os tópicos contêm palavras relevantes para a análise. Os índices de sentimento são construídos com base na frequência de palavras "positivas" e "negativas" em cada tema, conforme dicionários de palavras pré-definidas.

Os índices de sentimento são comparados com observações da inflação, expectativas de inflação e atividade econômica, e índices de incerteza comuns na literatura. Resultados sugerem que os documentos do Copom refletem o estado da economia de forma acurada, embora a correlação varie com o tempo, com mudanças significativas a partir de julho de 2016. A coerência entre comunicados e atas no período após julho de 2016 é avaliada. De um modo geral, ambos os documentos transmitem a mesma mensagem, com as atas detalhando informações previamente divulgadas nos comunicados.

Seeing the Forest for the Trees: Using hLDA Models to Evaluate Communication in Banco Central do Brasil*

Angelo M. Fasolo† Flávia M. Graminho‡ Saulo B. Bastos§

August 2, 2021

Abstract

Central bank communication is a key tool in managing inflation expectations. This paper proposes a hierarchical Latent Dirichlet Allocation (hLDA) model combined with feature selection techniques to allow an endogenous selection of topic structures associated with documents published by Banco Central do Brasil's Monetary Policy Committee (Copom). These computational linguistic techniques allow building measures of the content and tone of Copom's minutes and statements. The effects of the tone are measured in different dimensions such as inflation, inflation expectations, economic activity, and economic uncertainty. Beyond the impact on the economy, the hLDA model is used to evaluate the coherence between the statements and the minutes of Copom's meetings.

Keywords: communication, monetary policy, latent dirichlet allocation, Brazil, Central Bank

JEL Code: E02, E21, E22.

*We are grateful for the helpful comments and suggestions given by Michael McMahon, Juan M. Londono, José de Gregorio and the participants of the BIS-CCA Research Network on "Monetary policy frameworks and Communication", of the Computing in Economics and Finance 27th International Conference, and of the Seminários Acadêmicos do Banco Central do Brasil.

†Research Department – BCB. E-mail: angelo.fasolo@bcb.gov.br.

‡Research Department – BCB. E-mail: flavia.graminho@bcb.gov.br.

§Department of Banking Operations and Payments System – BCB. E-mail: saulo.benchimol@bcb.gov.br

1 Introduction

There is a widespread consensus that central bank communication influences the expectations of economic agents and that increasing transparency enhances the effectiveness of monetary policy. Central bank communication has moved beyond simply informing agents on the perceived current and future state of the economy to becoming a critical instrument to anchor inflation expectations. Transparency in central bank communication improves private sector short- and long-term interest rates forecasts while reducing the bias and variation of inflation rate expectations (see Swanson (2006), Neuenkirch (2012) and Jitmaneeroj, Lamla & Wood (2019)). Therefore, it is essential to assess whether the central bank is efficiently communicating the intended signals to market agents. This paper addresses the communication of the Monetary Policy Committee (Copom) of the Banco Central do Brasil (BCB) using a hierarchical Latent Dirichlet Allocation (hLDA) model (Griffiths, Jordan, Tenenbaum & Blei (2004)) to extract the content and measure the tone of statements and minutes on different aspects of the economic situation. The topic structure obtained after the hLDA model's estimation provides the basis to build indexes measuring the perception of Copom on different aspects of the economic situation, allowing for assessments on whether communication is accurately reflecting the state of the economy and on the coherence between statements and minutes of the Copom meetings.

The first contribution of this paper is the use of the hLDA model combined with feature selection techniques before estimation to describe monetary policy communication. After estimation, the model draws on the methodology of Hansen & McMahon (2016), Labondance & Hubert (2017) and Shapiro & Wilson (2019) to extract the content and measure the tone of statements and minutes on interest rate decisions. It tries to solve two important issues in computational linguistics that are closely associated with the characterization of central bank communication. First, the hLDA model estimates the number of topics and the distribution of words on each topic. Second, the hLDA model organizes results in a tree of topics, which favors interpretability and provides abstraction measures on a given topic, as the same topic might be discussed in different contexts. In the context of central bank communication, especially under a fully-fledged inflation targeting regime, it is expected that a significant share of the official documents be related to policy objectives, even when discussing secondary issues.

The use of feature selection, although not a novelty in computational linguistics in general, is of critical importance as a preliminary step in the estimation of hLDA models, as it allows for more interpretable topics at the root of the estimated tree. In their seminal hLDA paper, Griffiths et al. (2004) apply the hLDA algorithm in databases without previous treatment and only from the second level and beyond does the hierarchy proposed by the model become helpful for analytical purposes. Results show that feature selection as a preliminary step in the estimation of the hLDA model generates a topic at the root of the tree with meaningful words for the document as a whole.

On the empirical side, this paper contributes to the literature on text analysis of BCB's statements and minutes, providing a methodology to evaluate the consistency of the tone of Copom's documents with respect to inflation, economic activity and uncertainty. The sentiment of the BCB expressed in Copom's minutes is correlated with actual inflation, inflation expectations, economic activity and uncertainty. However, this correlation varied over time, with significant changes after a structural break in communication took place in July 2016. Restricted to the period after July 2016, BCB's communication is usually coherent, with minutes presenting detailed information first offered in statements.

A number of authors have used Latent Semantic Analysis (LSA) or its Bayesian counterpart, the Latent Dirichlet Allocation (LDA) model (Blei, Ng & Jordan (2003)), to extract topics from central bank documents or from market news to assess themes perceived as having the greatest impact

on the economy. Boukus & Rosenberg (2006) and Hendry & Madeley (2010) analyzed topics from the Federal Open Market Committee (FOMC) minutes and Bank of Canada (BoC) statements, respectively, using LSA and found these topics are correlated with current and future economic conditions. Hendry (2012) performed a similar analysis encompassing market news stories released five minutes after the BoC statement to capture possible second-order effects based on market analyst commentaries on those statements. Hansen & McMahon (2016) estimated an LDA model using content from FOMC statements and combined it with dictionary methods to quantify Fed's sentiment. After the LDA model identified 15 topics addressed by statements, the authors built an economic situation index by counting the words that reflect a positive and negative tone. This paper departs from the subjectivity of LSA and LDA models using a hLDA model for the BCB where the number of topics is endogenously determined during estimation.

The use of text analysis across different documents to evaluate specific features of communication relates this paper to Acosta (2015). The author uses LSA to investigate whether the degree of deliberation or the amount of public disagreement across FOMC members changed once meeting verbatim transcripts were made public. In addition, topics in FOMC transcripts (published five years after the meeting) are compared to minutes (published within a shorter lag) in order to quantify transparency. In Brazil, transcripts of Copom meetings are not available to the public. Right after the Copom meeting is finished, BCB publishes statements containing the monetary policy decision and a brief summary of its motivations, and one week later minutes are released with more detailed information about the economy and policy framework. Due to this temporal structure of publication, the Copom may strategically use minutes to smooth or reinforce messages previously sent by statements, depending on the agents reaction in between release dates. Indeed, Shapiro & Wilson (2019) repeat the computation of their negativity index using FOMC public documents, considering that *"[M]inutes and speeches provide an additional window into central bank preferences, albeit one potentially reflecting additional considerations beyond their internal policy deliberations"*. Correa, Garud, Londono & Mislang (2020) present evidence of central banks' communication reaction to recent developments associated with financial stability, while simultaneously offering information on the reaction to possible future states of the economy.

Finally, there is a growing research using text analysis on BCB's statements and minutes, such as Carvalho, Cordeiro & Vargas (2013), Rosa & Verga (2007), García-Herrero, Girardin & Dos Santos (2017), Cabral & Guimaraes (2015), Montes, Oliveira, Curi & Nicolay (2016) and Chague, De-Losso, Giovannetti & Manoel (2015). These papers usually focus on the semantic orientation of BCB communication, or the degree of "hawkishness"/"dovishness", optimism, and clarity of documents. However, to the best of our knowledge, no paper has provided an analysis of consistency across BCB's documents. This paper fills this gap.

The paper is organized as follows: Section 2 describes the hLDA model and the computation of economic situation indexes, while Section 3 provides details on the data used. The following section describes the estimation of hLDA model and the resulting indexes of economic situation inferred from the estimated model with Section 5 bringing the conclusions.

2 Methodology

2.1 From LDA to hLDA model

Early work on topic modeling derived from LSA and its Bayesian counterpart, LDA, in which the meaning of a text is a function of the words it contains. The intuition is that there is an underlying latent semantic structure to which any text can be mapped. The problem consists in reducing documents in a corpus to a vector of real numbers (word counts) whose dimensional space

resembles the latent semantic space. The LDA model is a probabilistic model of a corpus in which documents (observed variables) are represented as random mixtures over latent topics (not observed) where each topic is defined to be a probability distribution across words from a vocabulary. The core of the problem is to use documents to infer the hidden topic structure and thus compute the conditional (posterior) distribution of the hidden variables given the documents.[1] For our purposes, it suffices to note some key features of the model:

- The number of topics is assumed to be known and fixed, which requires a model selection procedure and subjectivity by the researcher to choose the adequate number of topics;
- The distribution of topics in LDA is independent and identically distributed, conditional on the underlying latent structure, neglecting the order of words in a document (exchangeability);
- Words can be allocated to multiple topics;
- Topics are a flat set of probability distributions, without relationship between topics.

The hierarchical LDA model (hLDA) developed by Griffiths et al. (2004) is an unsupervised Bayesian nonparametric model that deals with the first issue by inferring a distribution on topologies. Topics are organized according to a hierarchy tree with more general topics common to all documents being placed near the root, while more specialized topics are located near the leaves. Therefore, the nodes in this tree reflect the shared terminology of their children.

Different from the LDA model, there is no predetermined number of topics. The number of topics is jointly estimated during posterior inference and new documents can exhibit previously unseen topics, so the number of parameters can grow as the corpus grows. Moreover, in LDA, topics may be difficult to interpret since each theme is essentially a weighted sum of all of the words in a document. On the other hand, topics are usually more well-defined in the hLDA model because they are assigned along a single path in a hierarchy.

Technically, a tree can be viewed as a nested sequence of partitions. Each topic, seen again as a probability distribution across words, is associated with a node in the tree, and therefore each path is associated with an infinite collection of topics. Given a path, the probability distribution on the topics along this path is determined by a Griffiths-Engen-McCloskey (GEM) distribution. Given a draw from the GEM distribution, a document is generated by selecting topics based on that draw, and then by drawing words from the probability density function defined by its selected topic. Following the notation in Blei, Griffiths & Jordan (2010), let c_d denote the path through the tree for the dth document and $nCRP(\gamma)$ denote the stochastic process based on the "Nested Chinese Restaurant Process". The $nCRP$ is the distribution defining the probability that a certain element is part of an infinitely deep tree both in terms of number of branches and levels. Hyperparameter γ controls the frequency that a new word is moved to a new topic in a given level of the tree. Defining additional hyperparameters η, m, and π and $Z \sim Discrete(\theta)$ as the distribution setting $Z = i$ with probability θ_i, documents in a corpus are assumed drawn from the following process in the hLDA model:

- For each node $k \in T$ in the infinite tree,
  - Draw a topic $\beta_k \sim \text{Dirichlet}(\eta)$.
- For each document $d \in \{1, 2, \ldots, D\}$,
  - Draw $c_d \sim nCRP(\gamma)$.
  - Draw a distribution over levels in the tree $\theta_d|\{m, \pi\} \sim GEM(m, \pi)$.

[1] For details of the model, please refer to Blei et al. (2003) and Blei (2012).

– For each word,

  * Choose level $Z_{d,n}|\theta_d \sim Discrete(\theta_d)$.
  * Choose word $W_{d,n}|\{z_{d,n}, c_d, \beta\} \sim Discrete(\beta_{c_d}[z_{d,n}])$, which is parametrized by the topic in position $z_{d,n}$ on the path c_d.

Notice that the hLDA model provides inference not only on the allocation of words in each topic, but also with respect to the structure of the document. The model is still an unsupervised learning approach, but it demands more information from data used for estimation. This is one of the main difficulties in working with hLDA, as the possibility of different structures of the tree equally characterizing the document generates several local maxima in the likelihood function of the model. Compared to the LDA model, the Gibbs sampler procedure used to estimate the model demands a significantly larger number of iterations for convergence.

Another issue with estimation using the hLDA model is related to the inference of hyperparameters. Blei et al. (2010) propose a Metropolis-Hastings step between iterations of the Gibbs sampler. In this paper, the priors for the hyperparameters are set as in Blei et al. (2010), but it is worth noting the Metropolis-Hastings step naturally leads to additional autocorrelation between draws of the Gibbs sampler. Not only does the estimation problem become more complex, with more estimated parameters, but also the structure of the estimation procedure requires additional time for inference on the posterior distribution due to the slower convergence of the algorithm.

2.2 Creating indexes of economic perception from minutes

In order to quantify changes in perception by the Monetary Policy Committee about the economic situation, the estimated hLDA model is used to build indexes characterizing the tone of the message from the Committee through its official documents. The standard procedure in literature for building indexes on the economic situation, now adapted to the context of the hLDA model, requires four steps:

- First, given the estimated hLDA model, associate each final tree leaf (or, equivalently, a tree path) with a target subject. As an example, some leaves might be formed by words related to prices and inflation, while others might be formed by words related to economic activity (employment, production, etc), and so on. Let $\mathcal{C}_{\text{subj}}$ be the set of paths associated to a specific target subject;

- Second, using the model, locate all sentences associated with each target subject. If the tuple (m, s) identifies the sentence s of Copom's minute m, define $a_{(m,s)} = 1$ if the sentence belongs to $\mathcal{C}_{\text{subj}}$ and zero otherwise, that is:

$$a_{(m,s)} = \begin{cases} 1 & \text{if } \mathbf{c}^*_{(m,s)} \in \mathcal{C}_{\text{subj}} \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

  where $\mathbf{c}^*_{(m,s)} = \{c^1_{(m,s)}, \ldots, c^L_{(m,s)}\}$ is the path of the sentence (m, s) that has the highest probability among all possible paths and $c^l_{(m,s)}$ is the topic in level l;

- Third, given a previously defined dictionary characterizing the sentiment, locate keywords for every sentence associated with the target subject, i.e. with $a_{(m,s)} = 1$;

- Fourth, establish a metric comparing the frequency of dictionary terms found in each sentence.

Formally, the economic situation index is the net result of the application of positive and negative dictionaries over the words of a given set of the document:

$$\text{SubjSit}_m = \frac{\sum_{(m,s)} a_{(m,s)} \left(\text{Pos}_{(m,s)} - \text{Neg}_{(m,s)} \right)}{\sum_{(m,s)} a_{(m,s)} \text{Tot}_{(m,s)}} \tag{2}$$

where $\text{Pos}_{(m,s)}$ ($\text{Neg}_{(m,s)}$) is the number of positive (negative) words that appear in the sentence (m, s) and $\text{Tot}_{(m,s)}$ is the total words in the sentence (m, s).

As an additional feature of the indexes of economic perception, the presence of some words in a sentence is capable of reversing the sentiment with respect to a given subject. This is a deviation from the procedure described in Shapiro & Wilson (2019) where words preceded by "n't" or "not" are ignored in the situation metric. In the indexes presented here, the presence of "not" as well as other words listed in Appendix C transforms the sign of the sentiment in the sentence. As an example, the following sentences were published in the minutes' fifth paragraph of meeting 202 (October, 2016) describing the evolution of inflation and inflation expectations:[2]

> *"5. Returning to the domestic economy, recent inflation figures came in more* ***favorable*** *than expected, partly due to the* ***reversal*** *of food price* ***increases****. These results contributed to a* ***decrease*** *in expectations for 2016 IPCA inflation measured by the Focus survey, which stood at around 7.0%. As for 2017, IPCA inflation expectations reported in the same survey have* ***declined*** *to around 5.0% and remain* ***above*** *the inflation target of 4.5%. Expectations for 2018 and more distant horizons are already around this level."*

This paragraph is correctly identified by the hLDA model as discussing the evolution of inflation and inflation expectations. According to the dictionary proposed for the situation index on inflation, there are clearly three positive words ("favorable", "decrease", and "declined", highlighted in blue) and one negative word ("above", highlighted in red). By the dictionary's definition, the word "increases" (in green) should be considered a negative word with respect to inflation. However, the presence in the same sentence of the word "reversal" (in orange) changes the polarity of "increases". As mentioned before, in Shapiro & Wilson (2019) the word "increases" would be ignored, while here it is considered as another positive word on inflation.

Another example but with a different context in terms of sentiment is the sentence in the fifth paragraph of meeting 186 (October, 2014) discussing economic activity:

> *"5. (...) The PMI of the industrial sector, on its turn, indicates in September a* ***reversion*** *of the* ***expansion*** *seen in August."*

According to the dictionary proposed for the situation index on economic activity, the word "expansion" (in green) should be considered a positive word. Again, the presence of the word "reversion" (in orange) changes the polarity of "expansion". Thus, when computing the situation index on economic activity, this sentence accounts for one negative word.

Two observations about this procedure to build indexes. First, it should be clear that a new dictionary of positive and negative words must be defined for every target subject. The same word might have a different meaning from a sentiment perspective depending on the target subject. For instance, in the target subject "inflation", words such as "increase", "acceleration", and "rise" will usually have a negative sentiment associated with the topic, while words such as "retreat", "slowdown" and "fall" will be related to a positive sentiment. The same set of words are usually associated with the opposite sentiment if the target subject is defined as "employment" or "production".

[2] Unless otherwise noted, quotes from minutes and statements used in this paper are from the official records of the BCB's website in English.

Second, the same process can be applied to build indexes related to different outcomes associated with the target subject. Indeed, using a previously defined dictionary to find words associated with economic policy uncertainty in newspapers was the main idea in Baker, Bloom & Davis (2016). In other context, Lucca & Trebbi (2009) defined a dictionary to evaluate if the FOMC offered statements with a more "hawkish" or "dovish" tone. Thus, the idea of building dictionaries to evaluate statements are not restricted to a "positive" or "negative" tone, but also to other characteristics of communication. Formally, a monetary policy uncertainty index uses a selection of tree paths and a single uncertainty dictionary to measure the frequency of words associated with uncertainty that appear in the text:

$$\text{EconUnc}_m = \frac{\sum_{(m,s)} a_{(m,s)} \text{Unc}_{(m,s)}}{\sum_{(m,s)} a_{(m,s)} \text{Tot}_{(m,s)}} \tag{3}$$

where $\text{Unc}_{(m,s)}$ is the number of uncertainty words that appear in the sentence (m, s) and $\text{Tot}_{(m,s)}$ is defined as before.

Thus, situation indexes are computed by finding and counting the number of words fitting appropriate criteria, given the solution of the model described by the tree and suitable dictionaries.

3 Data

This section describes the two sources of data used for analysis: the textual data and the quantitative variables used to measure the impact of BCB's communication. The description of textual data also provides details on structural breaks observed in the main documents of Copom's meetings over time.

3.1 Textual data

There are two documents published by the BCB after every Copom meeting: statements[3] and minutes[4]. The statements disclose Copom's decision on the policy interest rate (Selic) and are released on the second day of the meetings. Minutes provide a detailed description of the reasons behind the decision on interest rates: recent economic developments, prospects for the Brazilian and global economies, and related balance of risks. Thus, Copom's minutes are longer in terms of number of words and paragraphs than the statements. The minutes were released on Wednesday after the meetings, and since July 2016, they are released on Tuesday after the meetings, so within five business days after the meetings. For estimation purposes, data on both documents are collected in Portuguese with estimation results translated into English.

One important detail that is unique with respect to the documents of Copom's meeting compared to similar documents in other central banks is that both the statements and minutes are published with domestic financial markets closed. Statements are published early in the evening of the last day of the meeting, while the minutes are published on Tuesday morning after the meeting. There is an explicit effort from the BCB to adjust publication schedule in order to keep this characteristic of the documents.[5] Therefore, minutes may serve as an instrument for reinforcing or smoothing information previously released in statements, depending on market reaction in between release dates.

The text of both documents was preprocessed based on the following steps:

[3] Historical statements of Copom available at: https://www.bcb.gov.br/controleinflacao/comunicadoscopom (in Portuguese) and at: https://www.bcb.gov.br/en/monetarypolicy/copomstatements (in English).

[4] Historical minutes of Copom available at: https://www.bcb.gov.br/publicacoes/atascopom/cronologicos and at: https://www.bcb.gov.br/en/publications/copomminutes (in English).

[5] For instance, there is a periodic adjustment of the schedule of publication of statements during daylight saving time season, as financial markets adjust to it.

1. First, all non-alphanumeric characters were removed except for periods at the end of sentences, and all text was put in lowercase with accent marks removed to avoid misspelling. The original data was kept for posterior stemming;[6]

2. The tokenization process splits sentences with periods and then the words by the whitespace character, and it considers a list of compound words in order to handle expressions such as "gross domestic product" (*produto interno bruto* in Portuguese) as a unique term instead of separated words. The list of compound words is partially displayed in Appendix B, where inflation indexes and treasury bill abbreviations were also added, Brazilian states and capitals, and all members of Copom meetings;

3. Common stopwords were excluded such as months, days of the week, cardinal and ordinal numbers (numeric characters and written in full as well), Brazilian states and capitals, and members of Copom meetings.[7]

The stemming process of the original words used the algorithm proposed by Orengo & Huyck (2001) due to the smaller understemming and overstemming errors observed when compared to other algorithms. Compound words, inflation indexes, and treasury bill abbreviations were kept without stemming.

Finally, one last procedure applied to textual data is feature selection. Feature selection is a technique to reduce the dimensionality of the vocabulary, a common practice to ease computational processing and also to reduce overfitting, according to Baeza-Yates & Ribeiro-Neto (2008). Hansen, McMahon & Prat (2018) applied feature selection using term frequency-inverse document frequency (TF-IDF) before estimating an LDA model with data from FOMC transcripts. The procedure here removes other words beyond stopwords based on document frequency. This procedure is easier to interpret and according to Yang & Pedersen (1997) it is as efficient as more sophisticated methods such as information gain, mutual information, and χ^2 statistic. While feature selection has an important role in the estimation of LDA models, the removal of words without a significant meaning due to low abstraction content is a critical step for estimation of the hLDA model. Roots of the trees estimated in Griffiths et al. (2004) and in Blei et al. (2010) were formed by prepositions, pronouns, and articles[8], while feature selection avoided this result and provided an additional layer of structure for the analysis of central bank communication.

The minimum document frequency affected the number of nodes in each level of the tree and also the most relevant words in each node. There is, however, an important choice with respect to the cutoff frequency to remove vocabulary. Our estimations suggest cutting the vocabulary by a small document frequency, such as only one or two sentences in the entire dataset, provided compact trees with fewer nodes, but with words that do not always seem the right fit for the node. On the other hand, cutting the vocabulary by a large document frequency, such as eight or ten sentences in the entire dataset, resulted in a slower convergence of the algorithm and provided oversized trees with too many nodes and with a level of detail that seemed too excessive, hurting comprehension of the results. Given these results, the choice was to cut the vocabulary by keeping words that are present in at least tree sentences.

The hLDA model is estimated using only the minutes from July 1999, when the Inflation Targeting regime was implemented in Brazil, until May 2020. The estimation procedure uses only certain sections of the minutes as early minutes used to present a long summary of data analyzed

[6]The stemming process in Portuguese needs to run on original data due to the effect of accent marks. For example, the stemmed term *deflacionar* ("to deflate" in English) is *deflac*, and so is *deflacionarão* ("will deflate"). But the stemmed result of the misspelled word *deflacionarao* is *deflacionara*, with the word without an accent mark providing an inaccurate stemmed term.

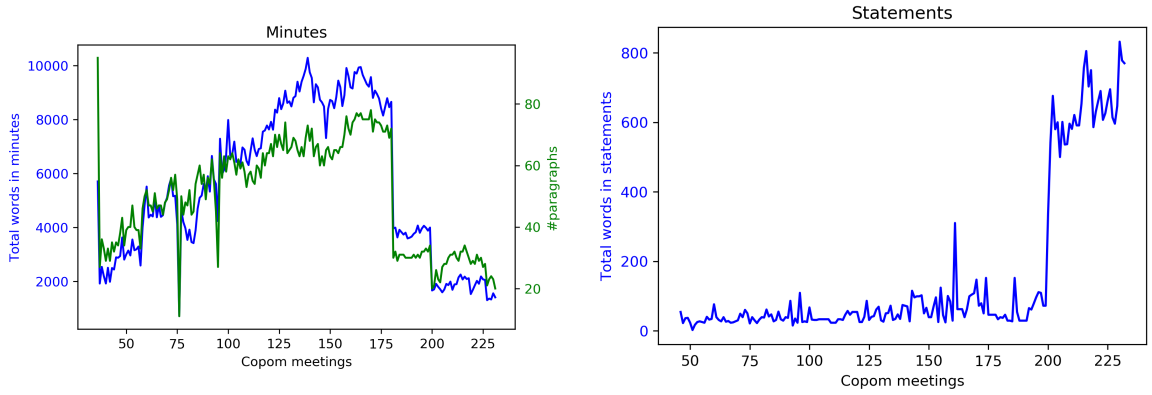
[7]See Natural Language Toolkit (NLTK) Python package, available at: `https://www.nltk.org/`.

[8]See Figure 5 in Griffiths et al. (2004); and Figures 1 (page 5), 7 (page 22), and 8 (page 23) in Blei et al. (2010).

during meetings, without any qualitative analysis by the Committee. Removing these sections results in a clear dataset for building situation indexes. Appendix A shows all sections comprised in Copom minutes with their respective indication if it was removed or not from the estimation procedure. It is worth noting that depending on the period described in the appendix, sections with the same name were removed from estimation, which is usually a consequence of changes in Copom's members, mainly the Chairman, resulting also in changes in the content of each section.

Data from the statements are not used in estimation, and the coherence analysis comprises only a small part of the sample. For the sake of the exercise, statements are used as out-of-sample information for the estimated model. The reason for that is the significant change in the structure of communication of Copom meetings as of July 2016 (meeting number 200), a few months after the new governor Ilan Goldfajn started his mandate. Before that meeting, statements were very short, with few sentences and usually without a reasoning for the decision[9]. Figure 1 shows the number of words in minutes and statements and the number of paragraphs in minutes for every meeting in the sample for estimation. That is the most basic metric available to characterize this structural change.

Figure 1: Statistics of minutes and statements



(a) (b)

Figure 1 also shows some significant changes in the number of words in the minutes. Table 1 splits the available sample into four periods that are consistent with those changes. Early in the Inflation Targeting regime, the minutes provided a qualitative analysis in almost every section using an average of approximately 3,700 words in 43 paragraphs. In the early months of Governor Henrique Meirelles, there was a change to longer minutes. However as mentioned before, a significant part of the minutes was a summary of data analyzed during the meetings. That structure lasted more than 10 years, from the minutes of meeting 82 of March 2003 to the minutes of meeting 180 of January 2014. Starting in February 2014, the entire section called "Summary of data analyzed by Copom" was removed from the main document. In statistical terms, that meant a reduction in the number of words in minutes from almost 7,900 to close to 3,800, but still keeping the same writing style with respect to the number of words per paragraph and per sentence. Statements remained short with the main objective of informing the monetary policy decision.

The structural break in July 2016 is significant in every statistic comparing the basic structure of both documents. While the previous breaks did not change the structure and the role of statements, justifying the absence of these periods in the coherence exercise and the use of statements in estimation overall, the July 2016 structural break altered the role of both documents in monetary

[9] A typical statement before July 2016 (available at: https://www.bcb.gov.br/en/pressdetail/2117/nota) contains the monetary policy decision and a disclaimer if the decision was unanimous, as in the April 2006 meeting:

> *"In the April Meeting, the Monetary Policy Committee (Copom) unanimously decided to reduce the Selic target by 75 basis points to 15.75 percent, without bias."*

Table 1: Statistics of minutes and statements of Banco Central do Brasil

		Copom meeting			
		36 (Jun/1999) – 81 (Feb/2003)	82 (Mar/2003) – 180 (Jan/2014)	181 (Feb/2014) – 199 (Jun/2016)	200 (Jul/2016) – 231 (Jun/2020)
Minutes	#words	3727.9 ± 1195.3	7887.9 ± 1593.6	3851.4 ± 151.8	1831.5 ± 265.9
	#paragraph	43.1 ± 11.5	64.1 ± 8.3	31.0 ± 1.2	27.4 ± 3.9
	#sentence	138.6 ± 38.6	255.7 ± 46.3	119.9 ± 5.1	68.4 ± 10.6
	$\frac{\#words}{\#paragraph}$	86.3 ± 53.1	122.9 ± 71.9	124.2 ± 58.1	66.6 ± 39.9
	$\frac{\#words}{\#sentence}$	26.8 ± 13.8	30.8 ± 13.6	32.1 ± 12.2	26.7 ± 11.7
Statements	#words	32.9 ± 13.2	56.0 ± 38.8	60.7 ± 35.7	630.5 ± 96.0
	#sentence	1.6 ± 0.7	1.7 ± 1.2	2.0 ± 1.1	24.1 ± 4.4

policy communication. The size of the statements grew substantially, adding information to the document published right at the end of the meeting. On the other hand, minutes became more succinct and more analytical, working as an extension and a complement of information provided in the statements. As shown in Table 1, the number of words in minutes was reduced to just under half, and the sentences became shorter on average, which facilitates readability.

3.2 Quantitative data

Indexes of economic perception generated from the estimation of hLDA models are used in this paper to measure the impact of BCB's communication on a set of economic variables measuring inflation, inflation expectations, and economic activity. In terms of prices, the reference inflation rate for the Brazilian inflation targeting regime is the IPCA (Extended National Consumer Price Index).[10] It is computed by the IBGE (Brazilian Institute of Geography and Statistics) and is based on the consumption basket of families with an income between 1 and 40 minimum wages. Inflation expectations for IPCA are gathered from the Focus survey carried out by BCB, which compiles daily forecasts of about 140 banks, asset managers, and other institutions regarding main Brazilian economic variables. Inflation expectations are provided in monthly and annual frequencies.[11] The system collecting information provides a sequence of checks for information providers in order to ensure consistency. For inflation, the system informs if inflation expectation for the next 12 months is consistent with the partial results for each month. Provision of information is not mandatory for participants of the survey, but the BCB discloses monthly and annual rankings of the survey's best forecasters in order to induce participation.

In terms of economic activity, we use monthly information about industrial production and wholesale trade indexes, both computed by the IBGE, because of its higher frequency when compared to the National Accounts.

[10] Open data from Banco Central do Brasil available at: http://dadosabertos.bcb.gov.br/.

[11] Open data from Banco Central do Brasil available at: https://www3.bcb.gov.br/expectativas/publico/consulta/serieestatisticas.

4 Estimation and Results

The first part of this section discusses the estimation of the hLDA model with details about the convergence of the algorithm, inference on the hyperparameters and their effects on the tree of topics' structure, and a preliminary analysis of the topics. The second part computes the indexes of economic situation and uses these indexes to measure the impact of BCB's communication on prices and economic activity. It also discusses the Central Bank's communication over time, associating changes of the economic situation index with the state of the economy, as well as the coherence of communication, comparing the indexes extracted from the hLDA model with indexes computed from the Copom's statements after monetary policy committee meetings.

4.1 Estimation of hLDA model

Using the code available at David Blei's webpage[12], most of the decisions in terms of estimation are related to the choice of the priors on hyperparameters described in subsection 2.1 and the covariance matrix of proposals for the Random-Walk Metropolis-Hastings (RWMH) algorithm estimating the distribution of some hyperparameters. The RWMH algorithm estimating the hyperparameters η, π, and m follows the strategy and the same values in Blei et al. (2010) with independent proposals and evaluations for each hyperparameter. Indeed, the procedure is implemented as one RWMH step for each hyperparameter instead of independent proposals generating sets of hyperparameters that are evaluated in one step. The choice of the priors was guided on some desired properties about the estimated model combined with the suggestions offered in Blei et al. (2010). Thus, priors were set looking for trees with a depth equal to three in order to facilitate visualization of results and values of η close to one since very small values generate oversized trees with many nodes, and prior on $m = 0.5$ so the posterior assigns more words from each sentence to higher levels of abstraction. Hyperparameter γ is sampled using the efficient Monte Carlo procedure described in (Escobar & West 1995, p. 585).[13]

In order to run the Gibbs sampler, 10,000 trees were randomly initialized, and estimation started with the one that had the highest score. The algorithm proceeded for 50,000 iterations with the first 40,000 iterations discarded as burn-in. Figure 2 (left) shows the evolution of the score over the final 10,000 iterations. Figure 2 (right) shows the autocorrelation as a function of the number of iterations between samples. Both panels make clear the need of a higher number of iterations in the Gibbs sampler compared to the baseline LDA model as the RWMH algorithm step generates significant autocorrelation of the draws, even if executed independently for each hyperparameter.

Figure 3 shows the hierarchy learned from the Copom minutes translated into English based on the Portuguese words without stemming.[14] The root node provides a set of words commonly used in general reports in the field of Economics. Internal nodes reflect the shared terminology of the documents assigned to the paths that contain them. Thus, the root node offers a set of words most with the highest global frequencies across documents applicable in the context of every other node in the tree.

The four subtopics provide insights on subjects analyzed in the Copom minutes over time:[15]

- Topic 1 (leaves in blue) focuses mostly on Copom's analysis of risks and scenarios in the context of the monetary policy decision;

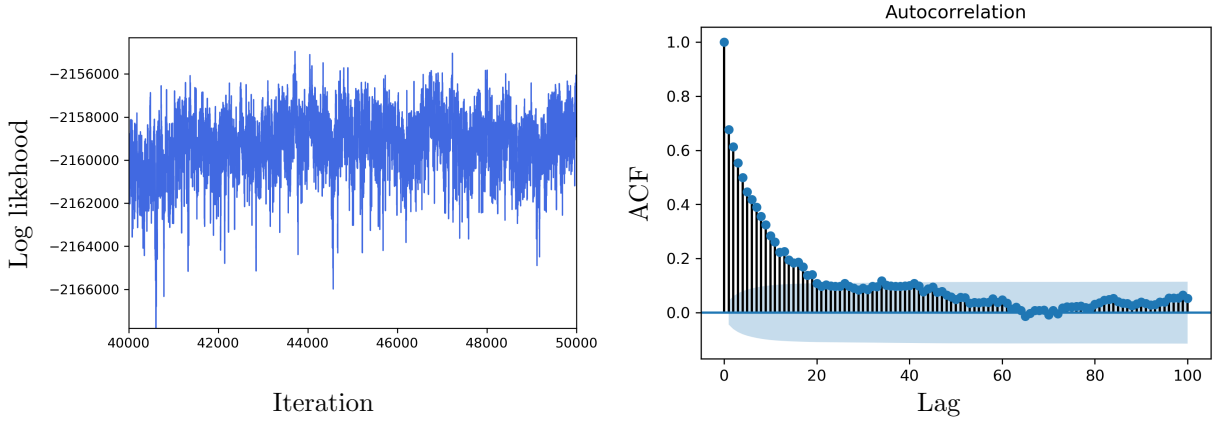
[12] "Hierarchical LDA C code implementation", available at: https://github.com/blei-lab/hlda.

[13] The presence of "hyper-hyperparameters" does not seem to influence results. Several values for the "hyper-hyperparameters" of η, γ, m, and π were tried, and all of them produced almost identical parameter values for most estimations with small changes in γ. Therefore, it is safe to say data itself led to the convergence of estimation, instead of the the choices on priors' setup.

[14] The original tree with words in Portuguese is available in Figure 11 of Appendix D.

[15] Notice that, for a given level of the tree, topics are independent from each other. Thus, the sequence of topics presented does not reflect the importance of a given topic in the sample analyzed.

Figure 2: Simulation of the hLDA algorithm



- Topic 2 (leaves in yellow) is related to economic activity variables, such as economic growth, industrial production, and the labor market;
- Topic 3 (leaves in pink) details information on prices, involving projections, expectations, and wholesale prices;
- Topic 4 (leaves in green) is more generic, discussing other scenarios and issues of the international economy.

As observed in subsection 3.1, monetary policy communication had a significant structural change in 2016. A natural question is if the structural change also included changes in the content of Copom's minutes. Figure 4 shows the evolution of second-level topics over time and the relative contribution of each topic to the document. Before July 2016, the minutes included detailed description of inflation and economic activity before a discussion on monetary policy. The relative contribution of topics 2 (economic activity) and 3 (prices) remained very stable during that period despite a spike in topic 3 around late 2002 and early 2003, while topics related to the monetary policy decision (topic 1) presented a steady increase over time. The 2002-2003 period is characterized by increases in domestic risk premium related to the fiscal policy's prospects after elections and an adverse scenario in the world's economy. Among the measures adopted to contain the crisis, an extraordinary Copom meeting was called, raising interest rates from 18%p.a. to 21%p.a.[16] The increase in risk premium resulted in a significant exchange rate devaluation and questions with respect to the evolution of prices.

After July 2016, with longer statements published after the meeting and shorter minutes one week later, information associated with topics related to economic activity (topic 2) sharply decreased both in absolute and in relative terms. The observed decrease in information associated with prices on topic 3 was mostly proportional to the reduction in the size of the minutes as the relative contribution of the topic to the document, presenting a modest reduction especially compared to the reduction of content on topic 2. The reduction of the relative share of paragraphs associated with topics 2, 3, and 4 was compensated by the increase of paragraphs associated with topic 1 on the discussion of monetary policy implementation, risks, and scenarios.

[16] Minutes for meeting number 77 in October 2002 described the crisis with the following paragraph: *"The confidence crisis derived from the uncertainties concerning the future guidance of the economic policy reduced credits to Brazil. Furthermore, the scandals related to big US corporations, the crises observed in emerging markets, the prospects of another war in the Gulf and the reduction of the likelihood of recovery of the US and European economies have been reducing the market's tolerance to risk. A series of indicators have been showing the increase in the risk perception, to levels comparable to those observed during the Russian crisis. The high correlation between the Embi+ Brazil and S&P 500 indexes shows that the increase of the country risk registered in the last months is partially due to the higher risk aversion observed in the international financial markets."*

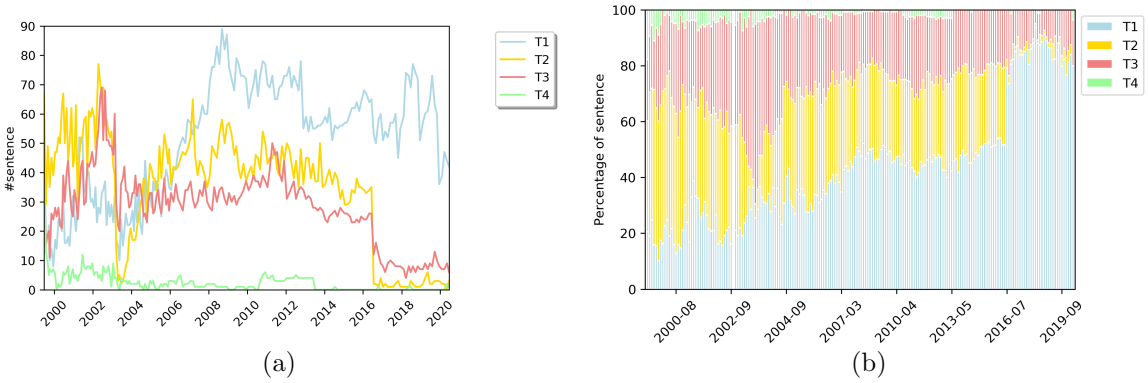
Figure 3: Hierarchy cloud from Copom minutes – Translated into English

4.2 Why not LDA? Why feature selection on hLDA?

This subsection discusses the effects of some of the techniques used to estimate the model. First, it presents a comparison of results of the hLDA model with the simple LDA model in terms of interpretation of wordclouds. The LDA model is configured to provide a clear baseline as to the output of the two models. Second, it discusses the role of feature selection in data treatment before the estimation. As briefly mentioned before, feature selection might play a critical role in the structure of the tree as its absence might concentrate at the root words without importance for content evaluation of documents.

The comparison with the baseline LDA model is that it tries to keep the estimation of both models as close as possible in terms of the dataset and the structure of the estimation. Thus, the dataset used when estimating the LDA model is exactly the same used in the hLDA model and with the same treatments, namely stemming, tokenization, and feature selection after the removal of stopwords. In terms of the structure of the model, LDA requires an exogenous definition about

Figure 4: Evolution of topics of the hLDA model



(a) (b)

the number of topics. The LDA model is estimated with 21 topics, which is the same number of topics endogenously defined during the hLDA model in the last level of the tree.

Table 2 summarizes the results of the LDA model showing in each column the five most relevant words in every cloud of the LDA model. Words highlighted in blue are the same words from the root of the tree of the hLDA model, while words in italic font are from the second level of the tree of the hLDA model, as shown in Figure 3. The sequence of columns presents the words ordered according to the weight in each cloud. The last two lines in the table sum the number of words in the column associated in the hLDA model with the root of the tree and with the tree's second level, respectively.

The first notable result from the estimation of the LDA model is related to the presence of words from the root and first level of the hLDA tree on the clouds of the LDA model. Considering only the second column of Table 2, only four words from the root of the hLDA tree[17] are the most relevant in 8 of the 21 LDA clouds of words estimated by the model. Including words from the second level of the tree, this proportion changes to 18 of the the 21 clouds. On the aggregate of results in Table 2, words from the root or from the second level of the hLDA tree are almost 75% of the total words presented.

The second notable result from Table 2 is related with the association of words with meaningful topics. On the one hand, results from the LDA model confirm the words included in the root or in the second level of the hLDA tree are indeed the most relevant in the set of documents analyzed. On the other hand, the very broad meaning of these words in the context of the documents requires additional inspection of the estimated LDA clouds in order to link them with specific topics. As examples, the words "price" and "inflation" are among the most relevant words in 9 of the 21 LDA topics (topics 2, 3, 7, 8, 9, 14, 16, 17, and 19), while the words "inflation", "Copom", and "scenario" are among the most relevant words in 3 of the 21 LDA topics (topics 2, 5, and 19). In both cases it is necessary to evaluate a larger set of words inside each cloud in order to associate the clouds with specific meaningful topics. The use in the hLDA model of the second level of the tree to define broad subjects and the last level of the tree to associate with specific topics makes the hLDA a very attractive alternative to build the situation indexes to evaluate the tone of the documents.

Another novelty of the paper is the use of feature selection as the last step in data treatment before using textual data to estimate the hLDA model. As mentioned in subsection 3.1, the objective of feature selection before model estimation is to provide content to the root of the hLDA tree, avoiding common but meaningless words such as prepositions, pronouns, and articles to dominate important words due to their high frequency in the documents. To evaluate the effect of feature selection in the model, two clouds are generated from the hLDA model. In the first version,

[17] Namely, "rate", "rise", "year", and "growth".

Table 2: LDA model cloud from Copom minutes – Relevant words

Topic 0	**growth**	quarter	*price*	*projection*	*accumulated*
Topic 1	**growth**	sale	commercial	wholesale	*consumption*
Topic 2	*price*	*inflation*	*scenario*	*risk*	*Copom*
Topic 3	*price*	**index**	*inflation*	**increase**	variation
Topic 4	**rate**	**growth**	employment	**year**	**index**
Topic 5	*inflation*	*Copom*	*scenario*	*Monetary Policy*	**rate**
Topic 6	**year**	**increase**	employment	*expansion*	*industry*
Topic 7	*price*	*inflation*	*accumulated*	expected	variation
Topic 8	*price*	*inflation*	**index**	variation	**year**
Topic 9	**rise**	*price*	**rate**	*inflation*	*Copom*
Topic 10	Utilizationof InstalledCapacity	good	**rate**	*industry*	*previous*
Topic 11	**rate**	*Copom*	*scenario*	*economy*	*committee*
Topic 12	*Monetary Policy*	effect	*price*	import	should
Topic 13	**growth**	**year**	**increase**	**index**	production
Topic 14	*inflation*	*projection*	*scenario*	*price*	**rate**
Topic 15	good	producer	*consumption*	*inflation*	capital
Topic 16	*inflation*	trajectory	**rate**	*price*	**increase**
Topic 17	variation	*inflation*	average	core	*price*
Topic 18	*economy*	**index**	**growth**	*Monetary Policy*	*Economic Activity*
Topic 19	*inflation*	*price*	*meeting*	*Copom*	*scenario*
Topic 20	**year**	adjustment	*inflation*	**continuity**	**increase**
# Root	8	5	6	4	6
# Second level	10	9	9	12	9

parameters obtained in the estimation using feature selection are fixed in the moments used for simulation, keeping the tree structure mostly fixed irrespective of the dataset, and feature selection is removed from data treatment. In the second version, the model is estimated with the dataset not treated with feature selection, so it is possible to approximate the effects of using feature selection both in terms of filtering a dataset with a given model (the first case) and in estimating a new tree with new information.

Table 3 compares the main words at the root of the tree of the baseline model with the main words at the tree with the two alternative specifications without feature selection. Due to the significant number of prepositions and the different prepositions based on gender, words at the root also present the original term in Portuguese. Results without using feature selection are consistent with those presented in the applications of the hLDA model in Blei et al. (2010), with the roots of the new trees including high-frequency but low-meaning words. Interestingly, in the model with constant parameters and no feature selection, wordclouds are almost identical starting from the second level of the tree compared with the baseline model. On the other hand, the model with new parameters estimated with dataset not using feature selection eliminates one full branch from the second level, resulting in a more concise tree but with low-meaning words spread at the second level as well.[18]

4.3 Situation indexes: Inflation, economic activity, and uncertainty

In this section, branches of the tree computed by the hLDA model are combined with specific dictionaries to build indexes related to the perception (or sentiment) of Copom about different aspects of the Brazilian economy. More specifically, from the tree presented in Figure 3, topics are selected to compute indexes associated with the perception about economic activity (EconSit_t),

[18] Results available upon request.

Table 3: hLDA model root without feature selection

Baseline Model	No Feature Selection	No Feature Selection: New Parameters
rate	of (*de*)	that (*que*)
rise	the (*a*)	of (*do*)
year	at (*em*)	price (*preço*)
index	of (*do*)	of (*da*)
increase	of (*da*)	with (*com*)
continuity	and (*e*)	at (*no*)
last	the (*o*)	for (*pelo*)
growth	at (*no*)	international (*internacional*)

inflation (InfSit_t), and uncertainty (EconUnc_m). As described in the methodology, a new dictionary is defined for each index of perception to associate the words with the proper tone of the document.

Indexes on inflation and economic activity use positive and negative dictionaries based on an extended version of the words used in Hansen & McMahon (2016) translated into Portuguese. The dictionaries are extended with words based on the authors' experience. The positive and negative dictionaries of the inflation sentiment and the economic activity sentiment are presented in Table 7 and Table 8, respectively, both in the appendix. Dictionary words are presented in a 3-column table where the first column contains the stemmed word in Portuguese, the second column contains the original word in Portuguese (an example of the stemmed word or the actual word used for matching dictionary words in the absence of a stemmed word), and the third column is the word translated into English. Both tables finish with the words characterizing polarity inversion when computing the sentiment of a given sentence, as discussed in the examples of subsection 2.2.

The dynamic nature of the documents plays a key role in choosing the appropriate paths on the tree to compute the situation indexes. Indeed, topics 2 and 3 on the second level of the tree are clearly associated with economic activity and inflation, respectively. However, topic 1 related to the monetary policy decision process also contains information about Copom's perception on the two subjects. As shown in Figure 4, topic 1 has recently covered a large share of the total words in documents. Thus, specific leaves from the third level of the tree of topic 1 are included in the path for each index to properly characterize the situation indexes.

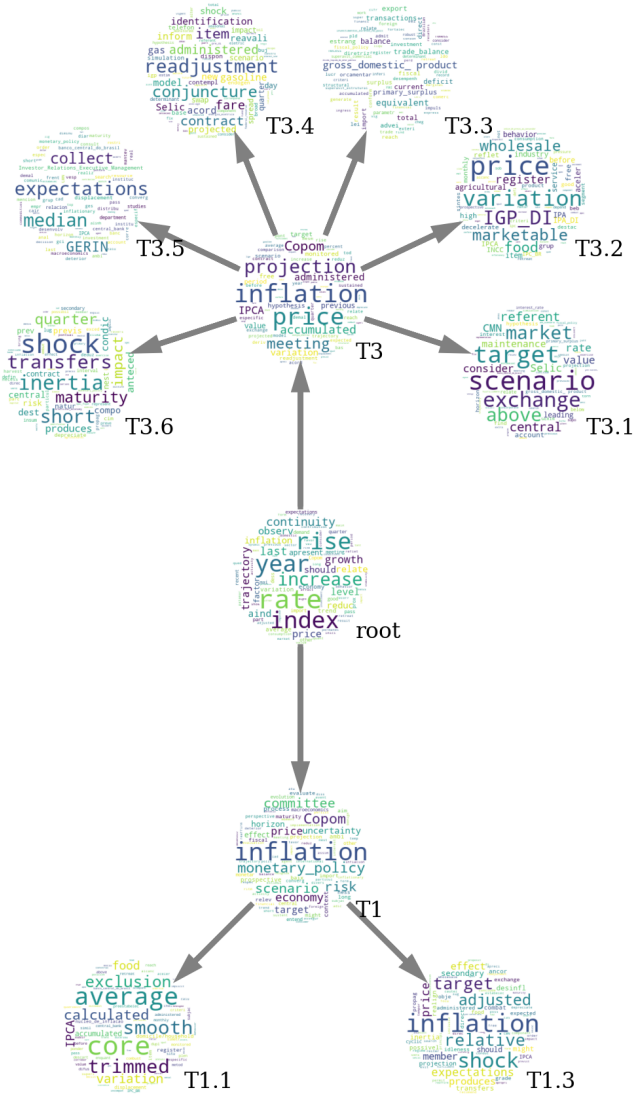
With respect to the inflation situation index, InfSit_t, Figure 5 presents the combination of topics used to compute the index in a wordcloud fashion, allowing inference on word weights as well. It shows the entire structure of topic 3 combined with the additional leaves from topic 1. The two additional leaves added from topic 1 discuss relevant topics on Copom's inflation analysis. The first one is a topic discussing the evolution of core inflation measures where words such as "core", "trimmed", "smooth", and "exclusion" have a large weight in the wordcloud. The second one provides an overview about Copom's expectations about inflation with words such as "target", "expectations", and "shock" with significant weight.

Figure 6 plots the evolution of inflation situation index (InfSit_t) with the year-on-year inflation and 12-month-ahead inflation expectations.[19] Overall, the inflation situation index shows, as expected, a negative correlation with both observables since negative words in terms of sentiment are associated with higher inflation. However, it seems that the correlation changes over time in a significant way. Indeed, the change is very significant comparing simple correlations across the entire sample with those after the structural break in communication of July 2016. For year-on-year IPCA inflation, correlation changes from -0.326 in the entire sample to -0.556 after July 2016. For inflation expectations, on the other hand, correlation changes from -0.451 to -0.274.

The same procedure applied to economic activity results in the wordcloud presented in Figure 7.

[19] Inflation expectations are the smoothed measure collected by the Focus survey, as described in section 4.2 with information provided on the last day of every Copom meeting considered in the sample. Data available since December 2001.

Figure 5: hLDA Tree Paths for Inflation Situation Index

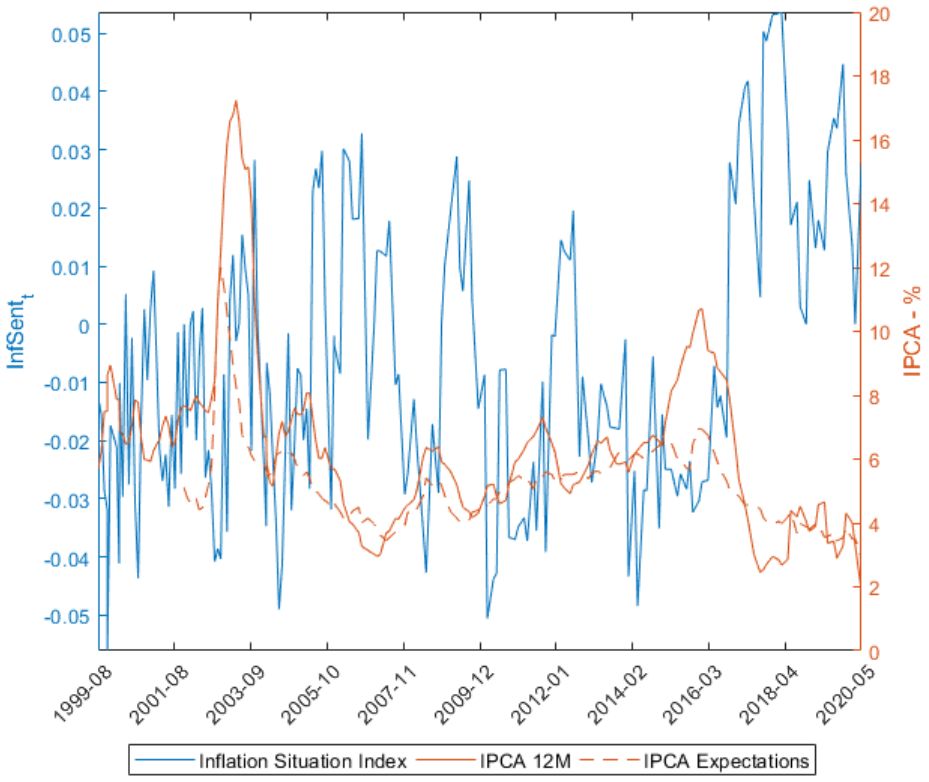


Five leaves characterize the branch with clear topics on labor market (T2.1), foreign trade (T2.2), credit and confidence (T2.3), retail and wholesale trades (T2.4 and T2.5), and industrial production and capacity utilization (T2.6). As in the case for the inflation situation index, leaves from the monetary policy topic 1 were added to provide a better characterization of the sentiment. In order to build the index of economic activity situation, one should be careful with the dictionary definition in situations of binary words. The same words describing "employment" or "production" in a positive manner usually describe "unemployment" with a negative sentiment. A natural evolution of this work is using word embedding techniques in an attempt to disentangle the context where each word enters in the document. For the sake of this paper, the index of economic activity situation is computed without the leaf related to labor market (T2.1).

Figure 8 compares the economic activity situation index, EconSit_t, with the evolution of wholesale trade volume and industrial production (both measured in 12-month changes). Contemporaneous correlation between the economic activity situation index and the real variables is significant but low in both cases: the correlation coefficient with industrial production is estimated at 0.361 and with wholesale trade volume is estimated at 0.225.

It is worth noting the increase in volatility of the economic activity situation index after the

Figure 6: Inflation situation index (InfSit_t), inflation, and inflation expectations



structural break of July 2016. The increase in volatility can be partially attributed to the significant reduction in the number of words dedicated to the topics associated with the subject after July 2016 – the denominator of the index. Figure 4 provided a hint with the significant reduction in the number of words associated with topic 2 over time. Including leaves from topic 1 did not improve the situation. Indeed, the average number of words building the economic activity situation index declined from an average of more than 1,100 words before July 2016 to only 105 words after the break. While the high volatility of the economy has also contributed to a volatile index, the structural break in Copom's communication is also important to explain the low correlation of the index with other datasets.

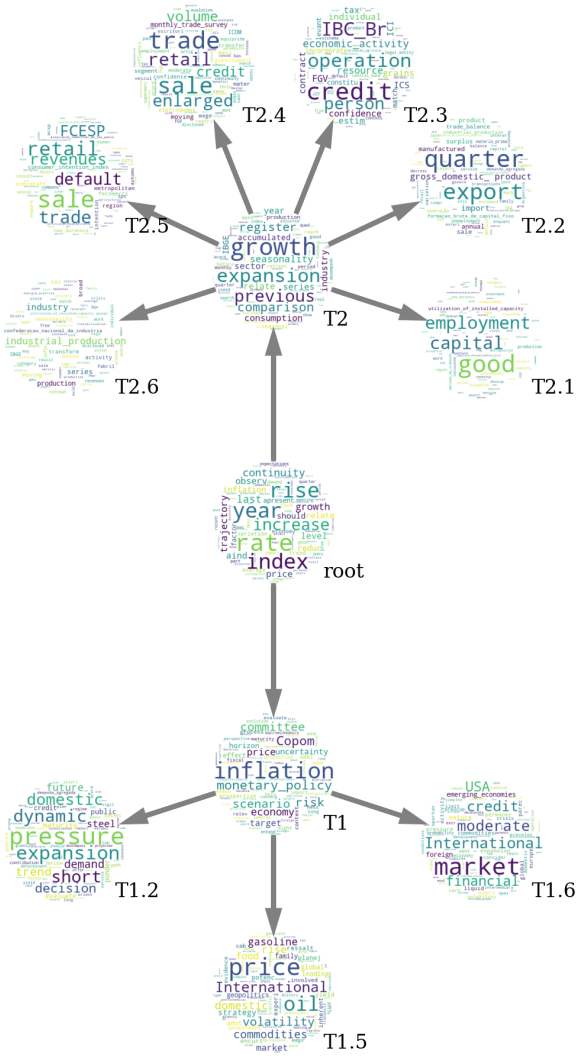
The economic uncertainty index (EconUnc_m) computed from the Copom meeting minutes, as previously mentioned, is one example of the sentiment index targeting different features of the Central Bank's communication. Instead of providing inference on a tone (positive/negative) with respect to the subject, the economic uncertainty index measures the degree of uncertainty expressed by Copom when justifying its decision on monetary policy.

Instead of relying on specific choices of tree paths, as in the case of the inflation and economic activity sentiment indexes, the economic uncertainty index is evaluated across the entire tree. There are two main reasons to justify using the entire tree. First, using the entire tree helps smoothing the indicator over time, facilitating comparison with other indexes measuring uncertainty.[20] Second, specific episodes responsible for increasing the uncertainty might be described outside of the tree path related to the monetary policy decision, such as topic 4 discussing mostly international financial market issues related to the decision may contain information on the degree of uncertainty during the Global Financial Crisis (2008-09). Thus, eliminating this path might actually underestimate uncertainty in that period.

As it is standard in the literature, the dictionary is the uncertainty word list of Loughran &

[20] Objectively, the denominator in Equation 3 becomes larger, smoothing the time series of the index.

Figure 7: hLDA Tree Paths for Economic Activity Index



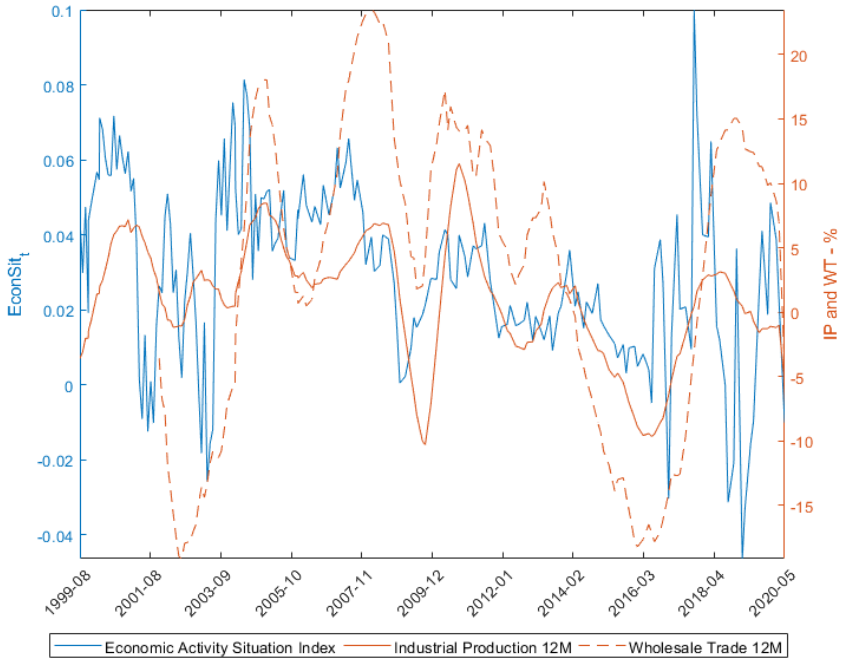
McDonald (2011)[21] after the exclusion of a few words that were out of context or could provide erroneous interpretations when discussing monetary policy. The complete dictionary used to build EconUnc_m is available in Table 9 under Appendix C.

Figure 9 compares the economic uncertainty index with the Economic Policy Uncertainty (EPU) index for Brazil based on calculations from Baker et al. (2016).[22] The correlation of the economic uncertainty index with the EPU is 0.4838 for the entire sample period, but falls to 0.1497 after July 2016. The coefficient of variation suggests the economic uncertainty index (0.5672) is slightly less volatile than the EPU (0.6212). Both indexes show similar changes during the electoral crisis of 2002-03, the Great Financial Crisis, and after January 2020 when the Covid-19 pandemic hit the economy. In the last episode, however, the increase in the economic uncertainty index is smaller compared to the EPU.

The increase in uncertainty after July 2016 and the decrease in correlation between the economic uncertainty index with EPU must be carefully considered. While both indexes show a build-up in uncertainty during 2016 as a consequence of the political crisis, the significant reduction in the EPU between 2018-2019 not followed by the economic uncertainty index suggests the structural break in Copom's communication after July 2016 might have affected the correlation of the two indexes. It is hard to disentangle the effects of the structural break in a small sample especially considering

[21] Loughran-McDonald Sentiment Word Lists. Available at: https://sraf.nd.edu/textual-analysis/resources/
[22] Data available at https://www.policyuncertainty.com/brazil_monthly.html.

Figure 8: Economic activity situation index (EconSit_t), industrial production, and wholesale trade



that both indexes moved closely during the peak of the crisis in 2016.

4.4 Statement and minute coherence

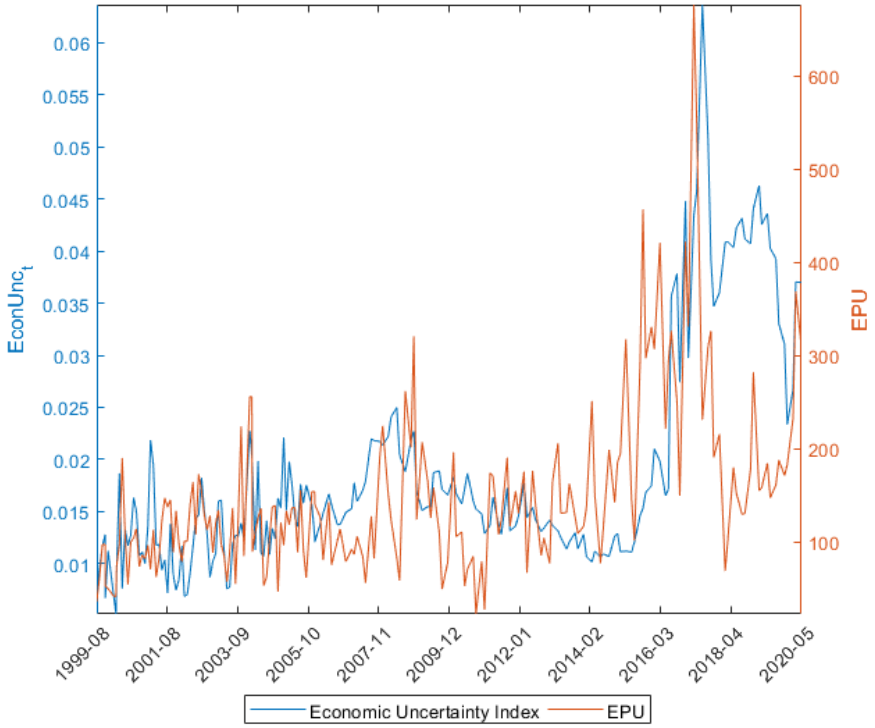
In this section we explore the temporal structure of documents published after a Copom meeting to evaluate the degree of coherence between statements and minutes. After the July 2016 structural break in communication, the statement of a given meeting (published right after the meeting is finished) became a sort of reduced version of the minutes (published in the next week) in the sense that economic agents usually expect more details about Copom's view of the economy in the minutes.[23] This temporal structure of documents allows for a simultaneous analysis both in terms of a given Copom meeting and across consecutive meetings. Conditional on the market reaction after the release of a statement, the Copom may strategically use the minutes to lead the market in another direction. Note, again, that the minutes in Banco Central do Brasil's monetary policy framework do not consist of literal transcripts from the Copom meeting, thus allowing for interventions in the tone and sentiment transmitted in the document.

In order to build the exercise, the same hLDA model and its cloud of words from the previous section were applied to each sentence of the statements, generating indexes of inflation, economic situation, and economy uncertainty based on the statements. Thus, the new time series of indexes from statements is built using only data from outside the sample used in the estimation of the cloud. The significant increase in the number of words and variety of topics discussed in the statement after July 2016 allows for a proper construction of both the economic situation and inflation sentiment indexes for the document.

Figure 10 compares the time series of the inflation and economic activity situation indexes for minutes after the July 2016 meeting with the out-of-sample indexes from the statements for each meeting. Indexes from the statements are usually more volatile compared to those of the minutes partly due to the smaller overall number of words used in statements described in detail in Table 1.

[23] Before this change, the statement was very short, usually just one sentence, with the explanation about the decisions restricted mainly to the minutes. Therefore, the analysis of coherence between statements and minutes does not apply to this period.

Figure 9: Economic uncertainty index (EconUnc_t) and EPU Index

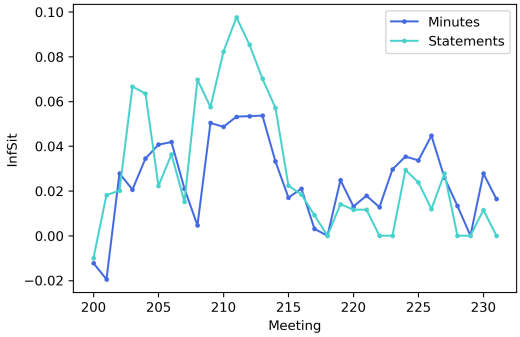


Since statements and minutes usually contain most of the relevant words to characterize the state of the economy, a smaller number of words in a paragraph induces larger variations in indexes from statements.

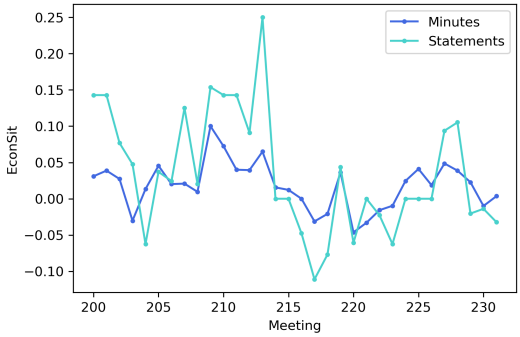
There are three hypotheses to be tested when comparing the time series of indexes from the statements and from the minutes. First, if for a given Copom meeting the sentiment expressed in the minute is consistent with the sentiment expressed in the previous week in the statement. Second, if the sentiment in the previous Copom meeting influences the sentiment in the current meeting. Finally, what other factors might influence the sentiment in a given document. As mentioned before, the structure of documents allows for a simultaneous test of all three hypothesis. A system of simultaneous equations linking situation indexes offers a chance to use information across equations to properly estimate the necessary parameters for hypothesis testing.

After defining Ind_t and $\widehat{\text{Ind}}_t$ as the situation index from the minutes and from the statements, respectively, and Weight_t and $\widehat{\text{Weight}}_t$ as the share of words in the document associated with the situation index from the minutes and from the statements, respectively, published after Copom

Figure 10: Minutes and statement indexes



(a) Inflation situation



(b) Economy situation

meeting t, the system has the following structure:

$$\text{Ind}_t = \alpha_0 + \alpha_1 \widehat{\text{Ind}}_t + \alpha_2 \text{Ind}_{t-1} + \alpha_3 \text{Weight}_t + \alpha_4 \text{X}_t + \epsilon_{1,t} \tag{4}$$

$$\widehat{\text{Ind}}_t = \beta_0 + \beta_1 \text{Ind}_{t-1} + \beta_2 \widehat{\text{Ind}}_{t-1} + \beta_3 \widehat{\text{Weight}}_t + \beta_4 \widehat{\text{X}}_t + \epsilon_{2,t} \tag{5}$$

The system of equations described above, without considering the control variables, consists of a comparison of information sets available to build the document after meeting t. In the first equation, the sentiment in minutes is associated with sentiment in statements (again, published the week before the minutes) and the minutes of the previous meeting. The second equation offers the same rationale for sentiment in statements, associating sentiment in statements of meeting t with the minutes of the previous meeting (the latest document available on the last meeting) and the sentiment in the statement of the previous meeting. The second term in both equations implicitly assume that there might be a specific language or word usage in each document (statements and minutes) that is persistent across meetings.

In the equations, X_t and $\widehat{\text{X}}_t$ are control variables included in different estimations of the system for robustness. Control variables are not added at once in the estimation since the sample from the meeting of July 2016 is rather small. Among others, these controls include information about the term structure of interest rates and nominal exchange rates on the day before the publication of the documents. They also include the (log) differences of nominal exchange rates and futures of interest rates between the days of the current and the latest document of reference.[24]

Table 4 shows the baseline results (columns 1 and 2) and the estimation with three different controls for the inflation and economic activity situation indexes: the change in swap rates (columns 3 and 4), the (log) difference in nominal exchange rates (columns 5 and 6), and the economic uncertainty index (columns 7 and 8). The first notable result is the significance of parameter α_1 across all estimations, supporting the first hypothesis stated earlier. It means that the situation indexes from the statements play a key role in explaining the situation indexes in the minutes, as expected in a consistent communication procedure.

With respect to the second hypothesis, the fact that α_2 is not statistically significant is an evidence that past minutes do not directly affect sentiment in the current minute. On the other hand, an analysis of the statements (equation (5)) shows a significant β_2 for all models with the inflation situation index, meaning that the previous statement contains information about the sentiment of inflation in the current meeting. This result, together with the consistency between statements and minutes, imply that there is an indirect effect of the inflation sentiment of the previous minute on the current one.

A slightly different picture is shown when analyzing the effect of the previous minutes on the current statements. Given the values for parameter β_1, the effect of the minutes from the previous meeting is marginal to describe the economic activity situation index, while it also is not significant for the inflation situation index.

Finally, with respect to the third hypothesis, it is worth noting the significance of α_3 with a negative sign in regressions for the inflation sentiment index in minutes. Combined with the negative correlation between inflation sentiment and inflation measures, this result implies that the minutes allocate a larger number of words when the sentiment on inflation deteriorates. Combined with the lack of significance of β_3 across all regressions and the temporal structure of publications, as discussed before, results show the minutes are used to expand the view of Copom about inflation, especially when the sentiment deteriorates. Also, the use of control variables does not seem to alter

[24] As an example, for the equation describing the sentiment of the minutes, a change in this variable is measured between the day after publication of the statement and the day before the publication of the minutes, approximately one week. For the equation describing the sentiment of the statement, change is measured between the day after publication of the last minutes and the day before publication of the new statement, approximately 45 days.

Table 4: Coherence of communication – SUR estimation

	InfSit_t	EconSit_t	InfSit_t δSwap	EconSit_t δSwap	InfSit_t δER	EconSit_t δER	InfSit_t EconUnc	EconSit_t EconUnc
α_0	0.024**	0.007	0.030**	0.007	0.025**	0.007	0.042**	0.067**
	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.014)	(0.025)
α_1	0.430**	0.288**	0.513**	0.277**	0.431**	0.309**	0.417**	0.279**
	(0.097)	(0.0558)	(0.097)	(0.056)	(0.097)	(0.055)	(0.094)	(0.051)
α_2	0.154	0.071	0.054	0.125	0.158	0.023	0.184	0.038
	(0.147)	(0.140)	(0.148)	(0.145)	(0.146)	(0.139)	(0.143)	(0.129)
α_3	-0.060*	0.005	-0.082**	-0.016	-0.064**	0.020	-0.060**	-0.064
	(0.030)	(0.067)	(0.031)	(0.068)	(0.032)	(0.066)	(0.029)	(0.067)
α_4			2.380*	-3.009	-0.001	-0.004	-0.480	-1.302**
			(1.373)	(2.452)	(0.002)	(0.003)	(0.290)	(0.518)
β_0	0.016	0.033	-0.008	0.033	0.016	0.031	0.005	0.072
	(0.016)	(0.027)	(0.018)	(0.026)	(0.016)	(0.027)	(0.021)	(0.071)
β_1	0.041	0.969	0.310	0.986*	0.057	1.044*	0.001	0.944
	(0.266)	(0.580)	(0.280)	(0.574)	(0.272)	(0.588)	(0.268)	(0.577)
β_2	0.723**	0.159	0.690**	0.169	0.712**	0.158	0.713**	0.171
	(0.155)	(0.226)	(0.146)	(0.224)	(0.160)	(0.225)	(0.154)	(0.225)
β_3	-0.027	-0.311	0.017	-0.419	-0.026	-0.278	-0.035	-0.341
	(0.045)	(0.283)	(0.047)	(0.309)	(0.046)	(0.286)	(0.046)	(0.284)
β_4			-1.789**	-2.378	0.000	-0.001	0.282	-0.692
			(0.791)	(2.965)	(0.001)	(0.002)	(0.343)	(1.196)
Wald: (H0)								
$\alpha_1 = \beta_1 = 0$	19.64**	29.50**	27.97**	27.34**	19.65**	34.35**	20.05**	32.95**
$\alpha_2 = \beta_2 = 0$	22.06**	0.76	22.45**	1.31	20.25**	0.53	22.19**	0.66
$\alpha_3 = \beta_3 = 0$	5.02*	1.21	7.11**	1.89	5.14*	1.03	5.75*	2.31
N	30	30	30	30	30	30	30	30
$R^2(\alpha)$	0.468	0.579	0.489	0.599	0.473	0.605	0.505	0.650
$R^2(\beta)$	0.541	0.301	0.574	0.315	0.541	0.309	0.548	0.309

Note: Standard-deviation in parenthesis. (**) significant at 5%, (*) significant at 10%.

the main results obtained in the baseline estimations. Indeed, the use of control variables allows for an additional characterization of the structure of statements and minutes, instead of altering the main results of the estimation.

5 Conclusions

This paper estimated a hierarchical Latent Dirichlet (hLDA) model to analyze minutes from the Banco Central do Brasil's Monetary Policy Committee (Copom). Compared to other text analysis models, the hLDA model allows for an endogenous selection of topic structure and for measures of abstraction of a given topic, thus providing relations between topics without previous intervention by the researcher. The additional use of feature selection as a preliminary step to the estimation assures that topics contain meaningful words that allow for proper document analysis.

The estimated model was then used to compute indexes characterizing the tone of Copom's message regarding inflation, economic activity, and uncertainty. Each tree path was associated with a target subject (inflation/economic activity) and indexes were built based on the frequency of "positive" and "negative" words according to a predefined dictionary. Overall, the comparison between the situation indexes and economic variables was affected by the significant changes in Banco Central do Brasil's communication in July 2016. The structural break in communication affected not only the correlation between the indexes and observables, but also their own volatility. The increase in volatility can be partially attributed to changes in the average number of words dedicated to topics associated with a specific subject.

The uncertainty index did not show the problem of changes in volatility due to a smaller number of words since it was evaluated over the entire tree. The economic uncertainty index, which measures the degree of uncertainty expressed by Copom when justifying its decision on monetary policy, was compared to the Economic Policy Uncertainty (EPU) index for Brazil based on calculations from Baker et al. (2016). While both indexes show similar changes during the electoral crisis of 2002-03, the Global Financial Crisis, and after January 2020, the economic uncertainty index is less volatile than the EPU.

Last, the coherence of Banco do Central do Brasil's communication was evaluated using statement data after July 2016 as out-of-sample data. Inflation situation and economic situation indexes from statements computed from the same hLDA model are usually more volatile compared to those of the minutes, partly again due to the smaller overall number of words used in statements. Results show, despite the fact that statements do not share the same information with all details present in minutes, they both transmit the same information.

Future work should explore more dimensions of Copom minutes such as measures related to monetary policy and the degree of forward guidance built in Banco Central do Brasil's communication.

References

Acosta, M. (2015), FOMC Responses to Calls for Transparency, Technical report, Board of Governors of the Federal Reserve System (US).

Baeza-Yates, R. & Ribeiro-Neto, B. (2008), *Modern Information Retrieval*, 2nd edn, Addison-Wesley Publishing Company, USA.

Baker, S. R., Bloom, N. & Davis, S. J. (2016), 'Measuring Economic Policy Uncertainty', *The Quarterly Journal of Economics* **131**(4), 1593–1636.

Blei, D. M. (2012), 'Probabilistic Topic Models', *Commun. ACM* **55**(4), 77–84.

Blei, D. M., Griffiths, T. L. & Jordan, M. I. (2010), 'The Nested Chinese Restaurant Process and Bayesian Nonparametric Inference of Topic Hierarchies', *J. ACM* **57**(2).

Blei, D. M., Ng, A. Y. & Jordan, M. I. (2003), 'Latent Dirichlet Allocation', *Journal of Machine Learning Research* **3**(Jan), 993–1022.

Boukus, E. & Rosenberg, J. V. (2006), 'The Information Content of FOMC Minutes', *Available at SSRN 922312* .

Cabral, R. & Guimaraes, B. (2015), 'O Comunicado do Banco Central', *Revista Brasileira de Economia* **69**(3), 287–301.

Carvalho, C., Cordeiro, F. & Vargas, J. (2013), 'Just Words?: A Quantitative Analysis of the Communication of the Central Bank of Brazil', *Revista Brasileira de Economia* **67**(4), 443–455.

Chague, F., De-Losso, R., Giovannetti, B. & Manoel, P. (2015), 'Central Bank Communication Affects the Term-Structure of Interest Rates', *Revista Brasileira de Economia* **69**(2), 147–162.

Correa, R., Garud, K., Londono, J. M. & Mislang, N. (2020), 'Sentiment in Central Banks' Financial Stability Reports', *Review of Finance* **25**(1), 85–120.

Escobar, M. & West, M. (1995), 'Bayesian Density Estimation and Inference Using Mixtures', *Journal of the American Statistical Association* **90**(430).

García-Herrero, A., Girardin, E. & Dos Santos, E. (2017), 'Follow What I Do, and Also What I Say: Monetary Policy Impact on Brazil's Financial Markets', *Economia* **17**(2), 65–92.

Griffiths, T. L., Jordan, M. I., Tenenbaum, J. B. & Blei, D. M. (2004), Hierarchical Topic Models and the Nested Chinese Restaurant Process, *in* 'Advances in Neural Information Processing Systems', pp. 17–24.

Hansen, S. & McMahon, M. (2016), 'Shocking Language: Understanding the Macroeconomic Effects of Central Bank Communication', *Journal of International Economics* **99**, 114–133.

Hansen, S., McMahon, M. & Prat, A. (2018), 'Transparency and Deliberation within the FOMC: a Computational Linguistics Approach', *The Quarterly Journal of Economics* **133**(2), 801–870.

Hendry, S. (2012), Central Bank Communication or the Media's Interpretation: What Moves Markets?, Technical report, Bank of Canada Working Paper.

Hendry, S. & Madeley, A. (2010), 'Text Mining and the Information Content of Bank of Canada Communications', *Staff Working Papers, Bank of Canada* .

Jitmaneeroj, B., Lamla, M. J. & Wood, A. (2019), 'The Implications of Central Bank Transparency for Uncertainty and Disagreement', *Journal of International Money and Finance* **90**, 222–240.

Labondance, F. & Hubert, P. (2017), Central Bank Sentiment and Policy Expectations, Sciences Po publications 648, Sciences Po.

Loughran, T. & McDonald, B. (2011), 'When Is a Liability not a Liability? Textual Analysis, Dictionaries, and 10-Ks', *Journal of Finance* **66**(1), 35–65.

Lucca, D. O. & Trebbi, F. (2009), Measuring Central Bank Communication: an Automated Approach with Application to FOMC Statements, Technical Report 15367, National Bureau of Economic Research.

Montes, G., Oliveira, L., Curi, A. & Nicolay, R. (2016), 'Effects of Transparency, Monetary Policy Signalling and Clarity of Central Bank Communication on Disagreement About Inflation Expectations', *Applied Economics* **48**(7), 590–607.

Neuenkirch, M. (2012), 'Managing Financial Market Expectations: the Role of Central Bank Transparency and Central Bank Communication', *European Journal of Political Economy* **28**(1), 1–13.

Orengo, V. M. & Huyck, C. (2001), A Stemming Algorithm for the Portuguese Language, *in* 'Proceedings Eighth Symposium on String Processing and Information Retrieval', pp. 186–193.

Rosa, C. & Verga, G. (2007), 'On the Consistency and Effectiveness of Central Bank Communication: Evidence from the ECB', *European Journal of Political Economy* **23**(1), 146–175.

Shapiro, A. H. & Wilson, D. J. (2019), Taking the Fed at its Word: A New Approach to Estimating Central Bank Objectives Using Text Analysis, Working Paper Series 2019-2, Federal Reserve Bank of San Francisco.

Swanson, E. T. (2006), 'Have Increases in Federal Reserve Transparency Improved Private Sector Interest Rate Forecasts?', *Journal of Money, Credit, and banking* **38**(3), 791–819.

Yang, Y. & Pedersen, J. O. (1997), A Comparative Study on Feature Selection in Text Categorization, *in* 'Proceedings of the Fourteenth International Conference on Machine Learning', ICML '97, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 412–420.

Appendices

A Sections of minutes used in estimation

Table 5: Minutes structure.

Copom meeting	Sections	Removed?
36 (June/1999) to 81 (February/2003)	Agregados Monetários e Crédito	No
	Ambiente Externo	No
	Atividade Econômica	No
	Avaliação Prospectiva das Tendências da Inflação	No
	Balanço de Pagamentos	No
	Demanda e Oferta Agregadas	No
	Diretrizes da Política Monetária	No
	Evolução do Mercado de Câmbio Doméstico e Posição das Reservas Internacionais	No
	Finanças Públicas	No
	Liquidez Bancária	No
	Preços	No
	Preços e Nível de Atividade	No
	Mercado Monetário e Operações de Mercado Aberto	Yes
82 (March/2003) to 180 (January/2014)	Evolução Recente da Atividade Econômica	No
	Evolução Recente da Economia	No
	Avaliação Prospectiva das Tendências da Inflação	No
	Implementação da Política Monetária	No
	Ambiente Externo	Yes
	Atividade Econômica	Yes
	Comércio Exterior	Yes
	Comércio Exterior e Alguns Resultados do Balanço de Pagamentos	Yes
	Comércio Exterior e Balanço de Pagamentos	Yes
	Comércio Exterior e Itens do Balanço de Pagamentos	Yes
	Comércio Exterior e Reservas Internacionais	Yes
	Crédito	Yes
	Crédito e Inadimplência	Yes
	Economia Mundial	Yes
	Evolução Recente da Inflação	Yes
	Expectativas e Sondagens	Yes
	Inflação	Yes
	Mercado de Trabalho	Yes
	Mercado Monetário e Operações de Mercado Aberto	Yes
	Setor Externo	Yes
	Sondagens e Expectativas	Yes
181 (February/2014) to 199 (June/2016)	Evolução Recente da Economia	No
	Avaliação Prospectiva das Tendências da Inflação	No
	Implementação da Política Monetária	No
200 (July/2016) to 230 (May/2020)	Atualização da Conjuntura Econômica e do Cenário Básico do Copom	No
	Riscos Em Torno do Cenário Básico Para a Inflação	No
	Discussão Sobre a Condução da Política Monetária	No
	Decisão da Política Monetária	

B Compound words

Table 6: Compound words.

Compound word in Portuguese	Meaning in English
Central de Custódia e de Liquidação Financeira de Títulos	Custody and Securities Settlement Center
índice de Preços por Atacado - Disponibilidade Interna	Wholesale Price Index - Internal Availability
índice de Preços ao Consumidor - Disponibilidade Interna	Consumer Price Index - Internal Availability
Federação do Comércio do Estado de São Paulo	São Paulo State Trade Federation
Notas do Tesouro Nacional - série D	National Treasury Notes - D Series
Notas do Banco Central - série E	Central Bank Notes - E Series
índice Nacional de Preços ao Consumidor Amplo	Broad National Consumer Price Index
índice Geral de Preços - Disponibilidade Interna	General Price Index - Internal Availability
índice de Preços ao Consumidor - Brasil	Consumer Price Index - Brazil
Imposto sobre a Renda Retido na Fonte	Withholding Income Tax
Fundo de Garantia por Tempo de Serviço	Lifetime Warranty Fund
Banco Nacional de Desenvolvimento Econômico e Social	National Bank for Economic and Social Development
Instituto Brasileiro de Geografia e Estatística	Brazilian Institute of Geography and Statistics
índice Nacional de Preços ao Consumidor	National Consumer Price Index
índice de Renda Fixa de Mercado	Market Fixed Income Index
índice de Preços ao Consumidor Harmonizado	Harmonized Consumer Price Index
índice de Confiança do Empresário Industrial	Confidence Index of Industrial Entrepreneur
Gerência Executiva de Relacionamento com Investidores	Investor Relations Executive Management
Serviço de Proteção ao Crédito	Credit Protection Service
National Association of Purchasing Managers	National Association of Purchasing Managers
índice Nacional da Construção Civil	National Index of Civil Construction
índice de Preços por Atacado	Wholesale Price Index
índice de Preços ao Produtor	Producer Price Index
índice de Preços ao Consumidor	Consumer Price Index
índice de Intenções do Consumidor	Consumer Intent Index
Emerging Market Bond Index Plus	Emerging Market Bond Index Plus
Contribuição sobre o Lucro Líquido	Contribution on Net Income
Bolsa de Mercadorias & Futuros	Commodities & Futures Exchange
Associação Comercial de São Paulo	São Paulo Commercial Association
Adiantamento de Contrato de Câmbio	Advance of Exchange Contract
Secretaria do Tesouro Nacional	National Treasury Secretariat
Letras Financeiras do Tesouro	Treasury Bills
Letras do Tesouro Nacional	National Treasury Bills
Imposto sobre a Renda	Income Tax
Federal Open Market Committee	Federal Open Market Committee
Contribuição sobre Movimentação Financeira	Contribution on Financial Transactions
Confederação Nacional da Indústria	National Confederation of Industry
Certificado de Depósito Interfinanceiro	Interbank Certificate of Deposit
Produto Interno Bruto	Gross Domestic Product
População Economicamente Ativa	Economically active population
Fundo Monetário Internacional	International Monetary Fund
Fundação Getúlio Vargas	Getúlio Vargas Foundation
Forward Rate Agreement	Forward Rate Agreement
Federal Reserve System	Federal Reserve System
Instituição Financeira	Financial institution
Depósito Interfinanceiro	Interbank Deposit
risco país	country risk
prêmio de risco	risk premium
produção industrial	industrial production
atividade econômica	economic activity
crescimento econômico	economic growth
economia brasileira	Brazilian economy
demanda agregada	aggregate demand
política monetária	monetary policy
política fiscal	fiscal policy
sistema financeiro	financial system
estabilidade financeira	financial stability
balança comercial	trade balance
superávit primário	primary surplus
energia elétrica	electricity
economia emergente	emerging economy
taxa de juros	interest rate
Banco Central	central bank
Banco Central Europeu	European central bank
Estados Unidos	United States

C Word lists

Table 7: Word lists related to the inflation situation index

Positive		
Stemmed word	**Original word**	**Translation**
adequ	*adequado*	adequate
arrefec	*arrefecimento*	cooling
baix	*baixo*	low
abaix	*abaixo*	below
benign	*benigno*	benign
-	*compatível, compatíveis*	compatible
-	*confortável, confortáveis*	confortable
-	*contração, contrações*	contraction
desaceler	*desaceleração*	slowdown
diminu	*diminuição, diminuir*	decrease
favor	*favorável*	favorable
frac	*fraco, fracamente*	weak
lent	*lento, lentamente*	slow
perd	*perda*	loss
progress	*progresso*	progress
qued	*queda*	fall
recu	*recuo*	retreat
recuper	*recuperação*	recovery
reduc	*redução*	reduction
reduz	*reduzir, reduz*	reduce or cut
Negative		
Stemmed word	**Original word**	**Translation**
aceler	*aceleração, acelerar*	acceleration
acim	*acima*	above
-	*alto, alta, altos, altas*	high
aquem	*aquém*	below
aument	*aumento, aumentar*	increase
cresc	*crescimento*	growth
desfavor	*desfavorável*	unfavorable
deterior	*deteriorar, deteriorado*	deteriorate
elev	*elevação, elevado*	elevation
expans	*expansão*	expansion
fort	*forte, fortemente*	strong
ganh	*ganho, ganhar*	gain
rapid	*rápido, rapidamente*	fast
sub	*subir, sobe, ...*	to rise
Polarity inversion		
Stemmed word	**Original word**	**Translation**
-	*desinflação*	disinflation
-	*não*	no, not
-	*reversão*	reversion

Table 8: Word lists related to the economic situation index

Positive		
Stemmed word	**Original word**	**Translation**
aceler	*aceleração, acelerar*	acceleration
adequ	*adequado*	adequate
acim	*acima*	above
-	*alto, alta, altos, altas*	high
aument	*aumento, aumentar*	increase
benign	*benigno*	benign
-	*compatível, compatíveis*	compatible
-	*confortável, confortáveis*	confortable
cresc	*crescimento*	growth
elev	*elevação, elevado*	elevation
expans	*expansão*	expansion
favor	*favorável*	favorable
fort	*forte, fortemente*	strong
ganh	*ganho, ganhar*	gain
progress	*progresso*	progress
rapid	*rápido, rapidamente*	fast
recuper	*recuperação*	recovery
sub	*subir, sobe, ...*	to rise
Negative		
Stemmed word	**Original word**	**Translation**
arrefec	*arrefecimento*	cooling
aquem	*aquém*	below
baix	*baixo*	low
abaix	*abaixo*	below
-	*contração, contrações*	contraction
desaceler	*desaceleração*	slowdown
desfavor	*desfavorável*	unfavorable
deterior	*deteriorar, deteriorado*	deteriorate
diminu	*diminuição, diminuir*	decrease
frac	*fraco, fracamente*	weak
lent	*lento, lentamente*	slow
perd	*perda*	loss
qued	*queda*	fall
recu	*recuo*	retreat
reduc	*redução*	reduction
reduz	*reduzir, reduz*	reduce or cut
Exclusion		
Stemmed word	**Original word**	**Translation**
infl	*inflação*	inflation
preç	*preço, preços*	price
Polarity inversion		
Stemmed word	**Original word**	**Translation**
-	*não*	no, not
-	*reversão*	reversion

Table 9: Word lists related to the economic uncertainty index

Stemmed word	Original word	Translation
incert	*incerteza, incerto*	uncertain, uncertainly, uncertainties, uncertainty
cautel	*cautela, cauteloso, cautelosa*	cautious, cautiously, cautiousness
aparent	*aparente, aparentemente*	apparent, apparently
confund	*confundir, confundido*	confuses, confusing, confusingly
-	*poderia, poderiam*	might, could
-	*pode, podem*	may
-	*depende, dependem, dependerá, dependerão*	depend, depended, dependence, dependencies, dependency, dependent, depending, depends
desvi	*desvio, desvios*	deviate, deviated, deviates, deviating, deviation, deviations
flutu	*flutuação, flutuações*	fluctuate, fluctuated, fluctuates, fluctuating, fluctuation, fluctuations
imprecis	*imprecisão*	imprecision
instabil	*instabilidade*	instability, instabilities
-	*possível, possivelmente*	possible, possibly
porvent	*porventura*	perhaps
talv	*talvez*	maybe
prelimin	*preliminar, preliminares*	preliminary, preliminarily
probabil	*probabilidade*	probability, probabilities, probabilistic
prov	*provável*	probable, probably
-	*reavaliado, reavaliada, reavaliados, reavaliadas, reavaliação, reavaliações*	reassess, reassessed, reassesses, reassessing, reassessment, reassessments
	revisão, revisões, revisar, revisado, revisada, revisados, revisadas	revise, revised
risc	*risco*	risk, risked, riskier, riskiest, riskiness, risking, risks, risky
-	*parece, parecem*	seems
especul	*especulativa, especulativo*	speculate, speculated, speculates, speculating, speculation, speculations, speculative, speculatively
esporád	*esporádico*	sporadic, sporadically
indef, indefin	*indefinido, indefinição, indefinições*	undefined
inesper	*inesperado, inesperada*	unexpected, unexpectedly
imprevist	*imprevisto*	unforseen, unexpected, unexpectedly, unpredictable
volatil	*volátil, volatilidade*	volatile, volatilities, volatility
antecip	*antecipado, antecipada*	anticipated
-	*temporário, temporária, temporários, temporárias*	temporary

D hLDA model results – Stemmed words in Portuguese

Figure 11: Hierarchy cloud from Copom minutes – Original in Portuguese