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

Angelo M. Fasolo[*]   Flávia Graminho[†]   Saulo B. Bastos[‡]

January 31, 2021

Abstract

Central Bank communication is a key tool in managing inflation expectations. This paper estimates a hierarchical Latent Dirichlet Allocation (hLDA) model to allow for an endogenous selection of topic structure associated with documents published by Banco Central do Brasil's Monetary Policy Committee (Copom). These techniques from computational linguistics allow for building measures about the content and tone of Copom's minutes and statements. The effects of the tone are measured in different dimensions, like 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.

This working paper should not be reported as representing the views of the Banco Central do Brasil. The views expressed in the paper are those of the author(s) and do not necessarily reflect those of the Banco Central do Brasil.

[*] Research Department, Banco Central do Brasil. E-mail: angelo.fasolo@bcb.gov.br.

[†] Research Department, Banco Central do Brasil. E-mail: flavia.graminho@bcb.gov.br.

[‡] Department of Banking Operations and Payments System, Banco Central do Brasil. E-mail: saulo.benchimol@bcb.gov.br

1 Introduction

There is a widespread understanding that Central Bank communication influences economic agents' expectations 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 become 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 expectations (see Swanson (2006), Neuenkirch (2012) and Jitmaneeroj, Lamla, and Wood (2019)).

The literature on Central Bank communication has focused on measuring the content of communication using unsupervised algorithms to learn signals about monetary policy decisions and measuring its impact on the economy. These algorithms, however, usually create independent "bags of words", without considering the degree of abstraction of words for a given topic or offering insights about the structure of the document. Critical characteristics of the algorithms are predefined by the researcher, without generating additional inference from the database.

This paper addresses some of the issues described above using a hierarchical Latent Dirichlet Allocation (hLDA) model (see Griffiths, Jordan, Tenenbaum, and Blei (2004)) to describe the effectiveness of communication in Banco Central do Brasil (BCB). The estimated hLDA model draws on the methodology of Hansen and McMahon (2016) to extract the content and measure the tone of statements and minutes on interest rate decisions. As a novelty, the topic structure obtained after the estimation of the hLDA model provides the basis to build indexes measuring the perception of the monetary policy committee (MPC) on different aspects of economic situation. These indexes are used to measure the influence of BCB's communication in inflation and inflation expectations. Using the same structure and technique, the paper provides an evaluation of the coherence in the communication of the BCB between statements and minutes of the MPC meetings.

The hLDA model applied to monetary policy communication, to the best of our knowledge, is one of the novelties of this paper. It tries to solve two important issues in computational linguistics, both closely associated with the characterization of Central Bank communication. First, the hLDA model estimates the number of topics and the distribution of words in each topic. Conventional LDA models estimate such distribution of words for a given, predefined number of topics. Estimation results might be sensitive to the choice of the number of topics. Second, the organization of results in a tree of topics provides measures of abstraction of a given topic – the same topic might be discussed in different contexts – and an inference on the relationship among topics. 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 is related to policy objectives, even when discussing secondary issues. For instance, the minutes of MPC meetings usually present lengthy discussions on labor markets or international financial conditions, even in situations when such subjects are of secondary importance to describe inflation dynamics in the economy. The hLDA model is able to provide, in this example, an endogenous characterization of the relation between topics related with labor markets or international financial conditions with topics associated with prices, without a previous intervention by the researcher.

In terms of computational linguistics, this paper also provides a contribution with the use of feature selection techniques before the estimation of the hLDA model. In Griffiths et al. (2004), the authors apply the algorithm in databases without previous treatment, resulting in less interpretable topics at the root of the tree of topics. In fact, only from the second level and beyond that the hierarchy proposed by the model becomes helpful for analytical purposes. Results show that feature selection as a preliminary step in the estimation of hLDA model generates a topic at the root of the tree with meaningful words for the document as a whole.

The paper is organized as follows: Section 2 briefly reviews the literature on text analysis in Central Bank communication. Section 3 describes the hLDA model and the computation of economic situation indexes, while section 4 provides details on data used. The following sections describe the estimation of hLDA model and the resulting indexes of economic situation inferred from the estimated model. Section 6 concludes.

2 Text Analysis in Monetary Policy

Credible Central Bank communication through statements, minutes and speeches is an important tool to inform economic agents both on the current and future states of the economy, and on the probable reaction of monetary policy authority to future events, thus helping on the management of expectations. Text analysis has become increasingly popular in studying Central Bank communication. Transcripts, minutes and statements of the FOMC have been recently unraveled in search for patterns that may explain aspects such as deliberation and sentiment, associating their effects on market and real variables.

Gürkaynak, Sack, and Swanson (2005) investigate the response of asset prices to FOMC announcements and find that the surprise change in interest rates is only one of the factors that affects prices. There is another latent factor related with FOMC statements that is capable of affecting asset prices, probably through their influence on market expectations on future policy actions. Authors argue that this hidden factor is the language used in Central Bank documents.

Bailey and Schonhardt-Bailey (2008) and Acosta et al. (2015) study deliberation in FOMC meetings by analyzing transcripts of the meetings. Bailey and Schonhardt-Bailey (2008) uses Alceste, a popular text analysis software in social sciences, to assess how Chairman Volcker managed to convince his colleagues to engage in tighter, but politically unpopular, policies to control inflation. Acosta et al. (2015) investigates whether the degree of deliberation, or the amount of public disagreement across FOMC members, changed once meetings' verbatim transcripts were made public. In addition, topics in FOMC transcripts (published five years after the meeting) are compared to minutes (published in a shorter lag) in order to quantify transparency, using Latent Semantic Analysis (LSA).

A number of authors also use LSA to extract topics from Central Bank documents and assess which specific themes are perceived as the most relevant and therefore have the greatest impact on financial markets. Boukus and Rosenberg (2006) extract topics from the FOMC minutes using LSA and find that these topics are correlated with current and future economic conditions. However, as the authors note, topics generated by LSA are difficult to interpret, requiring some degree of subjectivity. Hendry and Madeley (2010) analyze Bank of Canada (BoC) statements and find that forward looking statements have significant effects on bond market returns. Hendry (2012) performs a similar analysis encompassing market news stories released five minutes after the BoC statement, to capture possible second-order effects based on market analysts commentaries on those statements. Results show that official BoC communication reduces market volatility, while market news tend to increase it.

Hansen and McMahon (2016) use the Bayesian version of the LSA model – the Latent Dirichlet Allocation model (LDA) – in order to extract the content of FOMC statements, and combine it with dictionary methods to quantify central bank sentiment. After the LDA model identifies 15 themes addressed by statements, the authors build an economic situation index by counting the words which reflect positive and negative tone. This index is then related to market and real variables.

Lucca and Trebbi (2009) also measure the tone of FOMC statements and discussions from the media, by building two semantic orientation scores based on Google searches and the Dow Jones Factiva database. In the first score, a sentence is defined as "hawkish" or "dovish" depending on the relative frequency with which the sentence and the words "hawkish" or "dovish" jointly occur, respectively. The joint frequencies are computed based on Google searches. The second score uses the Dow Jones Factiva database to gather news with headlines involving the Federal Reserve or the FOMC, around times of meetings. A sentence is defined as hawkish or dovish using the same reasoning as described above. The authors then relate these semantic scores to interest rates, inflation and economic activity.

Tone indices are also computed by Labondance and Hubert (2017) and Shapiro and Wilson (2019), using the "bag of words" or lexical approach to FOMC publications. The indices are calculated based on the number of positive/negative words contained in FOMC publications, according to different dictionaries. Labondance and Hubert (2017) regress the tone index as a function of economic fundamentals, defining a proxy for the information set of the central bank, and the monetary policy shock, and define "sentiment" as the residual of this regression. The authors find that central bank sentiment has some impact on macroeconomic variables. Shapiro and Wilson (2019) use a negativity index based on FOMC transcripts, minutes and members' speeches to proxy for the loss function of the Committee, and estimate its implicit inflation target.

In Brazil, research using text analysis on BCB's statements and minutes is very recent. Carvalho, Cordeiro, and Vargas (2013) build on the methodology of Lucca and Trebbi (2009), using Google searches to assess the semantic orientation of sentences in BCB's statements. The authors find that, during the period prior to Governor Tombini's tenure, changes in the informational content of Copom's statements have significant effects in bond yields in the short and medium terms. Different versions of the methodology of Rosa and Verga (2007) – creating dummy variables based on the degree of "hawkishness" or "dovishness" of BCB's publications – have been applied by a number of authors (García-Herrero, Girardin, and Dos Santos (2017), Cabral and Guimaraes (2015), among others). Montes, Oliveira, Curi, and Nicolay (2016) compute clarity scores based on the number of words, sentences and syllables, finding that greater transparency is correlated with lower levels of disagreement about inflation expectations. Chague, De-Losso, Giovannetti, and Manoel (2015) construct an Optimism Factor, based on the Harvard IV dictionary and on the minutes of Copom's meeting. However, to the best of our knowledge, no paper has provided an analysis of consistency across BCB's documents. This paper fills this gap.

3 Methodology

3.1 From LDA to hLDA model

Early work on topic modeling derived from Latent Semantic Analysis (LSA), 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 Bayesian approach to this problem led to the Latent Dirichlet Allocation (LDA) method Blei, Ng, and Jordan (2003). 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), and 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, that is, to 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.

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

- 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 which deals with the first issue by inferring a distribution on topologies. Topics are organized according to a hierarchy tree: more general topics (common to all documents) are placed near the root, while more specialized topics are located near the leaves. Therefore, nodes in a tree reflect the shared terminology of their children.

Different from the LDA model, there is no predefined number of topics. The number of topics is jointly estimated during posterior inference, and new documents can exhibit previously unseen topics – the number of parameters can grow as the corpus grow. 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 defined in hLDA model because they are assigned along single paths 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 GEM distribution defines a probability distribution on the topics along this path. Given a draw from a GEM distribution, a document is generated by selecting topics based on that draw, and then drawing words from the PDF defined by its selected topic. Following the notation in Griffiths et al. (2004), let c_d denote the path through the tree for the dth document and *nCRP*(γ) 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 level $k \in T$ 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, $\theta_d|\{m, \pi\} \sim GEM(m, \pi)$.
  - 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, indeed, 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 of the hLDA model is related to the inference of hyperparameters. Blei, Griffiths, and Jordan (2010) propose a Metropolis-Hastings step between iterations of the Gibbs sampler. In this paper, the priors for the hyperparameters are set just like Blei et al. (2010), but it is worth noting that the Metropolis-Hastings step naturally leads to additional autocorrelation between simulation of the Gibbs sampler. Not only the estimation problem becomes more complex, with more estimated parameters, but the structure of the estimation procedure requires additional time for inference on the posterior distribution, due to the slower convergence of the algorithm.

3.2 Creating indexes of economic perception from minutes

In order to quantify changes in perception of 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; 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 $\mathbf{c}^*_{(m,s)} = \{c^l_{(m,s)}, \ldots, c^L_{(m,s)}\}$ 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 $a_{(m,s)} = 1$.

- Finally, 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 of 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 and Wilson (2019), where words preceded by "n't" or "not" are ignored in the situation metric. In indexes presented here, the presence of "not" – as well as other words listed in Appendix C – transform 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 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 and 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, **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, for every target subject, a new dictionary of positive and negative words must be defined. 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 like "increase", "acceleration" and "rise" will usually have a negative sentiment associated with the topic, while words like "retreat", "slowdown" and "fall" will be related with 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".

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 is the main idea in Baker, Bloom, and Davis (2016). In other context, Lucca and Trebbi (2009) define 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 present 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.

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

Thus, situation indexes require predefined dictionaries together with specific paths of the tree. Given dictionaries and the solution of the model – in the case here, described by the tree – it is necessary to find and count the number of words fitting the appropriate criteria.

4 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.

4.1 Textual data

There are two documents published by BCB after every Copom's meeting: statements[3] and minutes[4]. Statements disclose the Copom's decision on the policy interest rate (Selic) and are released after the second day of 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, the minutes of Copom are longer, in terms of number of words and paragraphs, than the statements. Minutes are released on Tuesday after the meetings (thus, within six business days after the meetings). For estimation purposes, data on both documents are collected in Portuguese, with estimation results translated to English.

One important detail that is unique with respect to the documents of Copom's meeting, compared to similar documents in other Central Banks: both statements and minutes are published with domestic financial markets closed. Statements are published early in the night of the last day of the meeting, while the minutes are published on Tuesday morning after the meeting. There is an explicit effort from BCB to adjust publication schedule in order to keep this characteristic of the documents[5].

Text of both documents was preprocessed based on the following steps:

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

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

3. Common stopwords were excluded, like 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].

[3] Historical statements of Copom. Available at: https://www.bcb.gov.br/controleinflacao/comunicadoscopom.

[4] Historical minutes of Copom. Available at: https://www.bcb.gov.br/publicacoes/atascopom/cronologicos.

[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.

[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 accent mark providing an inaccurate stemmed term.

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

The stemming process of the original words used the algorithm proposed by Orengo and 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 reduce overfitting, according to Baeza-Yates and Ribeiro-Neto (2008). The procedure is applied by removing (beyond stopwords) words based on the document frequency. This is a simple procedure and, according to Yang and Pedersen (1997), it is as efficient as more sophisticated methods (such as information gain, mutual information and χ^2 statistic). Removing 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) are formed by prepositions, pronouns and articles[8]; 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 that cutting the vocabulary by a small document frequency (such as only one or two sentences in the whole dataset) provided compact trees, with fewer nodes, but with words that do not always seemed 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 whole 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 results. Given these results, the choice was to cut the vocabulary by keeping words that are present in at least tree sentences.

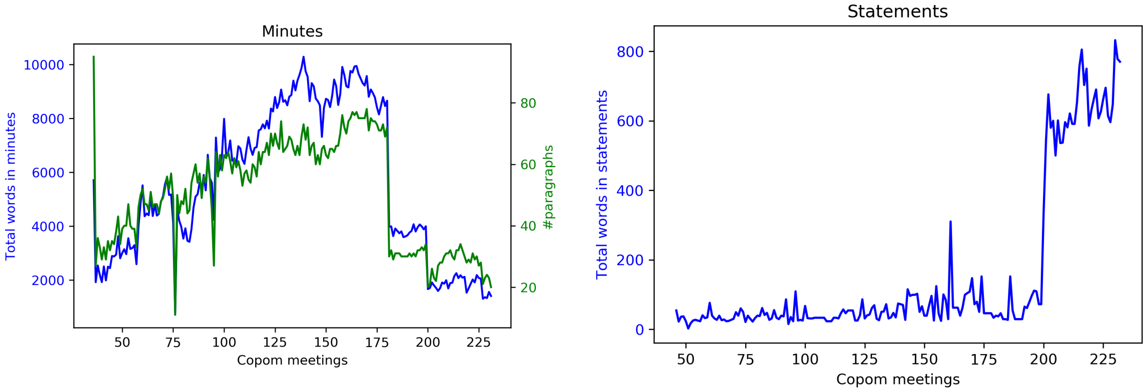
The hLDA model is estimated using only the minutes from July-1999 (when Inflation Targeting regime was implemented in Brazil) until May-2020. The estimation procedure uses only certain sections of minutes, as early minutes used to present a long summary of data analyzed during meetings, without any qualitative 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 its respective indication if it was removed or not from the estimation procedure. It's worth noting that, depending on the period described in the appendix, sections with the same name were removed from estimation. That 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 structure of communication for Copom meeting number 200, in July-2016, few months after 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. 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 also shows some significant fluctuations in the number of words in minutes. Table 1 splits the available sample in four periods, consistent with these fluctuations. Early in the Inflation Targeting regime, minutes provided 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 for longer statements and minutes. However, as mentioned before, a significant part of the minutes was a summary of data analyzed during meetings. That structure lasted

[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).

Figure 1: Statistics of minutes and statements



(a) (b)

more than 10 years, from the minutes of meeting 82 of March-2003 to minutes of meeting 180 of January-2014. Starting in February-2014, the whole 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.

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 (Fev/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
	$\frac{\#\text{words}}{\#\text{paragraph}}$	86.3 ± 53.1	122.9 ± 71.9	124.2 ± 58.1	66.6 ± 39.9
	$\frac{\#\text{words}}{\#\text{sentence}}$	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

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 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 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 have become more succinct, 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.

4.2 Quantitative data

In this paper, indexes of economic perception generated from the estimation of hLDA models are used 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)[9]. It is computed by IBGE (Brazilian Institute of Geography and Statistics) and it is based on the consumption basket of families with 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[10]. 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 particpants of the survey, but BCB discloses monthly and annual rankings of the survey's best forecasters, in order to induce participation.

Finally, in terms of economic activity, information about industrial production and wholesale trade indexes, both computed by IBGE, are used due to the higher frequency, compared to the National Accounts.

5 Estimation and Results

The first part of this section discusses the estimation of hLDA model, with details about convergence of the algorithm, inference on the hyperparameters and its 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, and 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.

5.1 Estimation of hLDA model

Using the code available at the authors' webpage[11], most of the decisions in terms of estimation are related to the choice of the priors on hyperparameters described in section subsection 3.1 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 depth equal to 3, in order to facilitate visualization of results, values of η close to 1, since very small values generates oversized trees with many nodes, and prior on $m = 0.5$ so the posterior assigns more words from each sentence to

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

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

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

higher levels of abstraction. Hyperparameter γ is sampled using the efficient Montecarlo procedure described in Escobar and West (1995), p. 585[12].

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 makes 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 2: Simulation of the hLDA algorithm

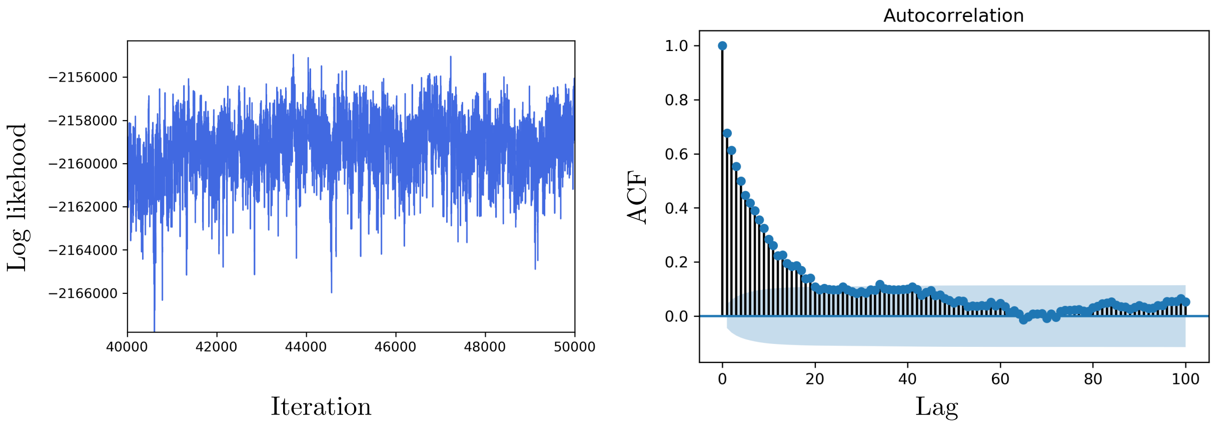


Figure 3 shows the hierarchy learned from the Copom minutes, translated to English based on the Portuguese words without stemming[13]. The root node provides a set of words commonly used in general reports in the Economics field. Internal nodes reflect the shared terminology of the documents assigned 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 Copom minutes over time[14]:

- Topic 1 (leaves in blue) focus mostly on the Copom's analysis of risks and scenarios in the context of the monetary policy decision;
- Topic 2 (leaves in yellow) is related to economic activity, such as economic growth, industrial production and the labor market;
- Topic 3 (leaves in pink) details information on prices, involving forecasts, expectations and wholesale prices;
- Topic 4 (leaves in green) is more generic, discussing other scenarios and issues on the international economy.

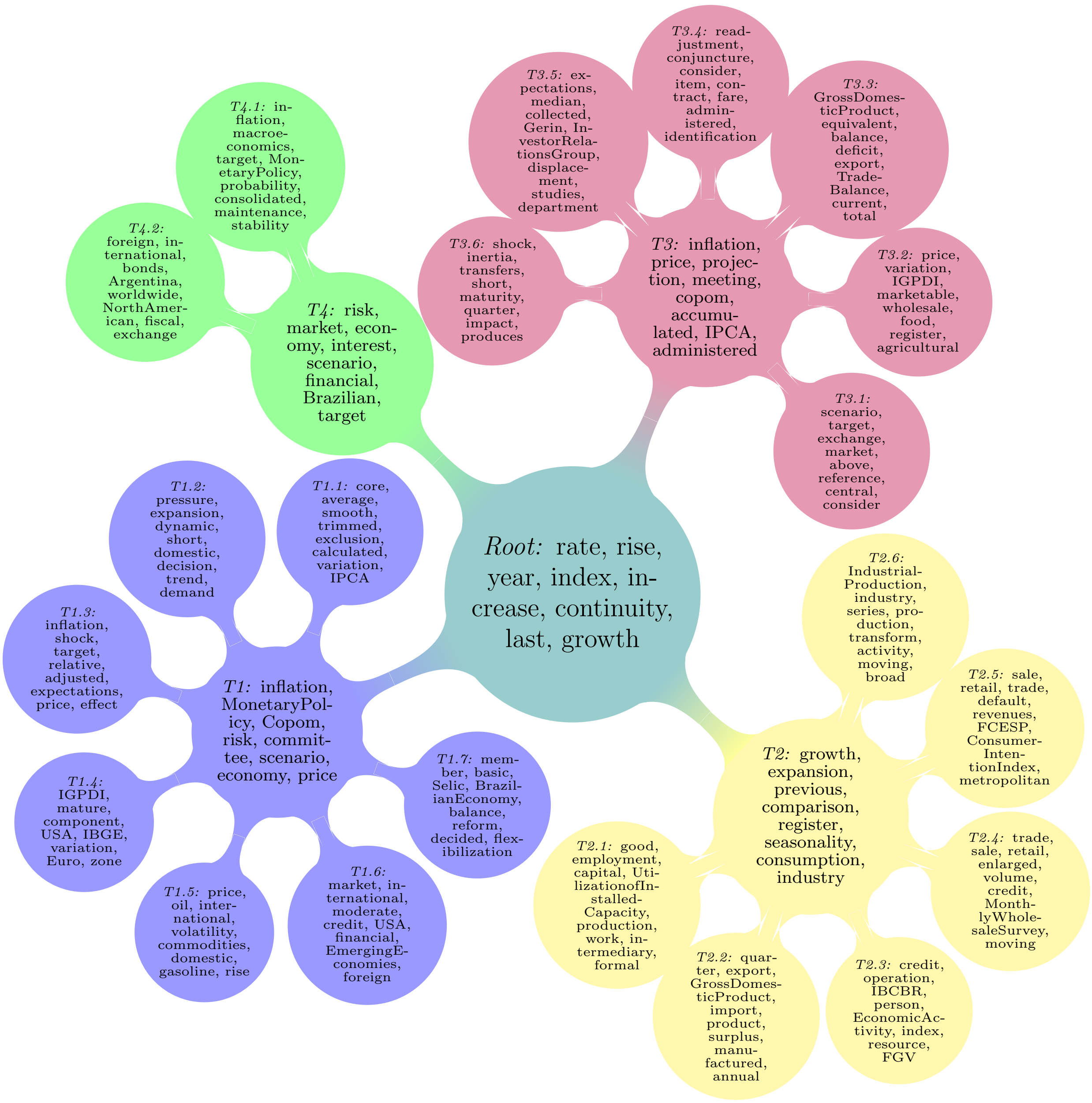
As observed in subsection 4.1, monetary policy communication had a significant structural change in 2016. A natural question is if the structural change also included changes in the content of Copom's minutes. Figure 4 shows the evolution of second-level topics over time and the relative

[12] 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's safe to say that the data itself led to the convergence of estimation, instead of the the choices on priors' setup.

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

[14] 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 3: Hierarchy cloud from Copom minutes – Translated to English

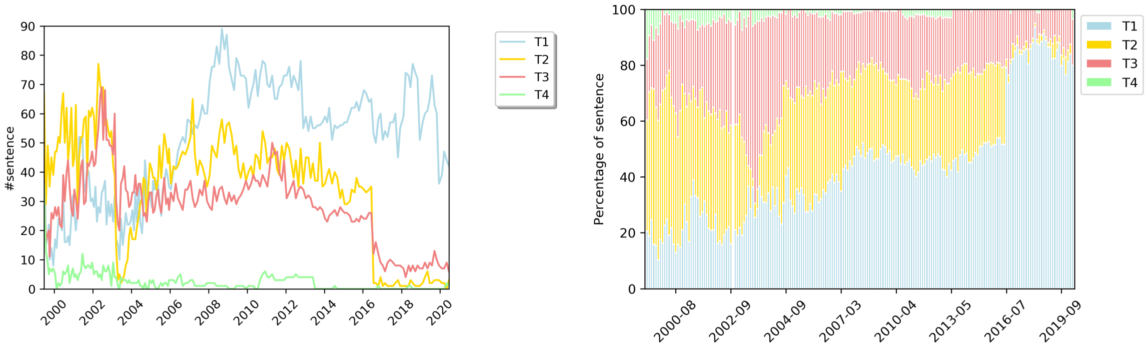

Root: rate, rise, year, index, increase, continuity, last, growth
T4: risk, market, economy, interest, scenario, financial, Brazilian, target
T4.1: inflation, macroeconomics, target, MonetaryPolicy, probability, consolidated, maintenance, stability
T4.2: foreign, international, bonds, Argentina, worldwide, NorthAmerican, fiscal, exchange
T3: inflation, price, projection, meeting, copom, accumulated, IPCA, administered
T3.5: expectations, median, collected, Gerin, InvestorRelationsGroup, displacement, studies, department
T3.4: readjustment, conjuncture, consider, item, contract, fare, administered, identification
T3.3: GrossDomesticProduct, equivalent, balance, deficit, export, TradeBalance, current, total
T3.6: shock, inertia, transfers, short, maturity, quarter, impact, produces
T3.2: price, variation, IGPDI, marketable, wholesale, food, register, agricultural
T3.1: scenario, target, exchange, market, above, reference, central, consider
T1: inflation, MonetaryPolicy, Copom, risk, committee, scenario, economy, price
T1.2: pressure, expansion, dynamic, short, domestic, decision, trend, demand
T1.1: core, average, smooth, trimmed, exclusion, calculated, variation, IPCA
T1.3: inflation, shock, target, relative, adjusted, expectations, price, effect
T1.4: IGPDI, mature, component, USA, IBGE, variation, Euro, zone
T1.5: price, oil, international, volatility, commodities, domestic, gasoline, rise
T1.6: market, international, moderate, credit, USA, financial, EmergingEconomies, foreign
T1.7: member, basic, Selic, BrazilianEconomy, balance, reform, decided, flexibilization
T2: growth, expansion, previous, comparison, register, seasonality, consumption, industry
T2.6: IndustrialProduction, industry, series, production, transform, activity, moving, broad
T2.5: sale, retail, trade, default, revenues, FCESP, ConsumerIntentionIndex, metropolitan
T2.4: trade, sale, retail, enlarged, volume, credit, MonthlyWholesaleSurvey, moving
T2.1: good, employment, capital, UtilizationofInstalledCapacity, production, work, intermediary, formal
T2.2: quarter, export, GrossDomesticProduct, import, product, surplus, manufactured, annual
T2.3: credit, operation, IBCBR, person, EconomicActivity, index, resource, FGV

contribution of each topic to the document. Before July-2016, the minutes included detailed analysis of inflation and economy 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 steadily increase over time. The 2002-2003 period is characterized by increases in domestic risk premium, related to fiscal policy's perspective after elections and an adverse scenario for 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.[15]. The

[15] Minutes for meeting number 77 – 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 marketâs 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 indices show that the increase of the country risk registered in the last months is partially due*

increase in risk premium resulted in significant exchange rate devaluation and questions with respect to the evolution of prices.

Figure 4: Evolution of topics of the hLDA model



(a)

(b)

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) has 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 presented a modest reduction, especially when 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 the topic 1, on the discussion of monetary policy implementation, risks and scenarios.

5.2 Why not LDA? Why feature selection on hLDA?

This subsection discusses the effects of some of the techniques used to estimate the model. Notably, 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 about 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 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, 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 the number of topics.The LDA model is estimated with 21 topics – 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 second level of the tree, respectively.

to the higher risk aversion observed in the international financial markets."

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[16] 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.

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

The second notable result from Table 2 is related with the association of words with meaningful topics. On the one hand, results from LDA model confirm that the words included in the root or in the second level of the hLDA tree are, indeed, the most relevant in the set of analyzed documents. 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); 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 indices 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 4.1, the objective of feature selection before model estimation is to provide content to the root of the hLDA tree, avoiding that common but meaningless words (mainly prepositions, pronouns and articles) dominate important words due to their high frequency in the documents. To evaluate the effect of

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

feature selection in the model, two clouds are generated from the hLDA model: in the first version, parameters obtained in the estimation using feature selection are fixed in the moments used for simulation − thus, 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. Thus, 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 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 selections, 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[17].

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
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*)

5.3 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), inflation (InfSit_t) and uncertainty (EconUnc_m). As described in the methodology, for each index of perception, a new dictionary is defined 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 and McMahon (2016), translated to 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 on the appendix. Dictionary words are presented in a three-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 3.2.

[17] Results available upon request.

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 the Copom's perception on the two subjects. And, 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 for inference on words weights as well. It shows the whole structure of topic 3 combined with the additional leaves from topic 1. The two additional leaves added from topic 1 discusses relevant topics on the inflation analysis of Copom: the first one is a topic discussing the evolution of core inflation measures, where words like "core", "trimmed", "smooth" and "exclusion" have a large weight in the wordcloud; the second one provides an overview about the Copom's expectations about inflation, with words like "target", "expectations" and "shock" with significant weight.

Figure 5: hLDA Tree Paths for Inflation Situation Index

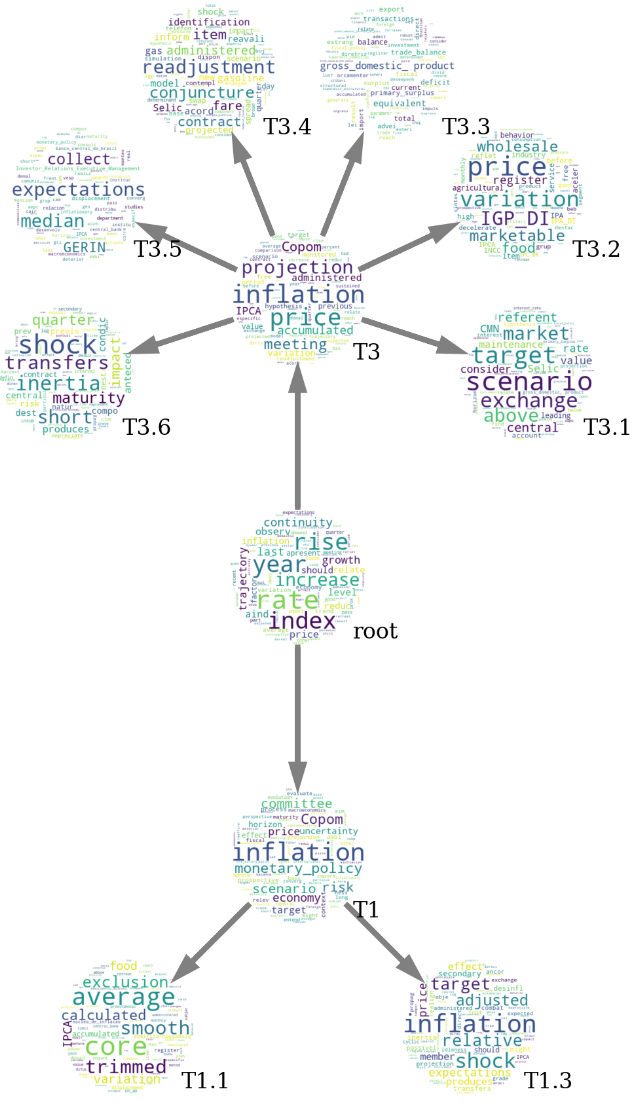
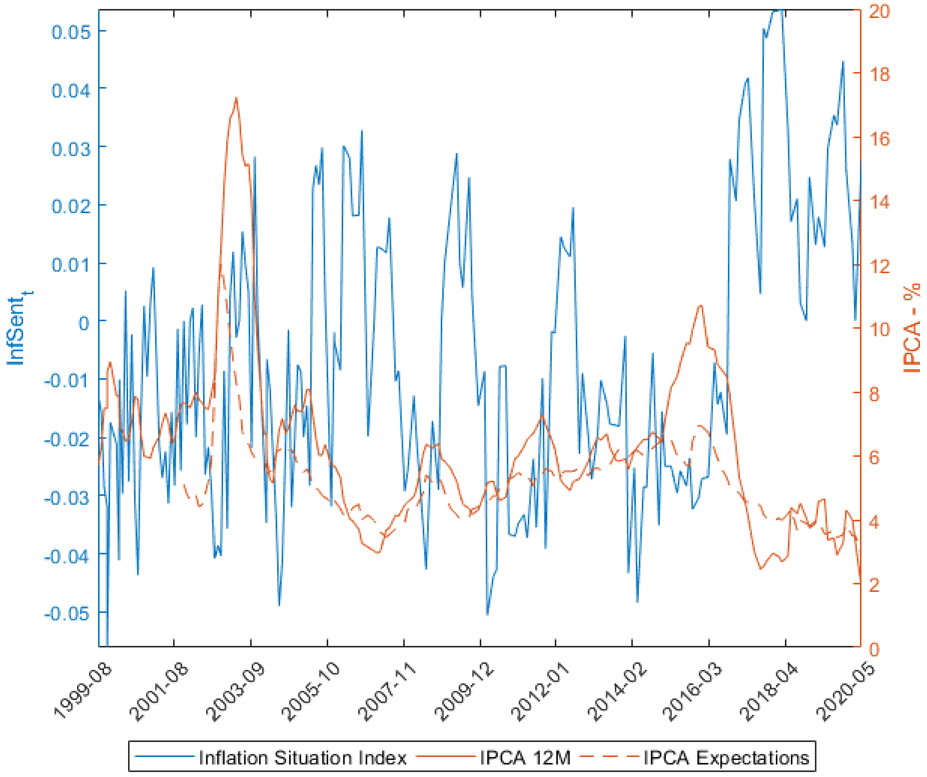


Figure 6 plots the evolution of inflation situation index (InfSit$_t$) with the 12-month accumulated inflation and inflation expectations 12 months ahead[18]. Overall, the inflation situation index shows, as expected, a negative correlation with both observables, since negative words in terms of sentiment 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 in the whole sample with those after the structural break in communication of July-2016: for 12-month IPCA inflation, correlation changes from -0.326 in the whole sample to -0.556 after July-2016; for inflation expectations, on the other hand, correlation changes from -0.451 to -0.274.

Figure 6: Inflation situation index (InfSit$_t$), inflation and inflation expectations

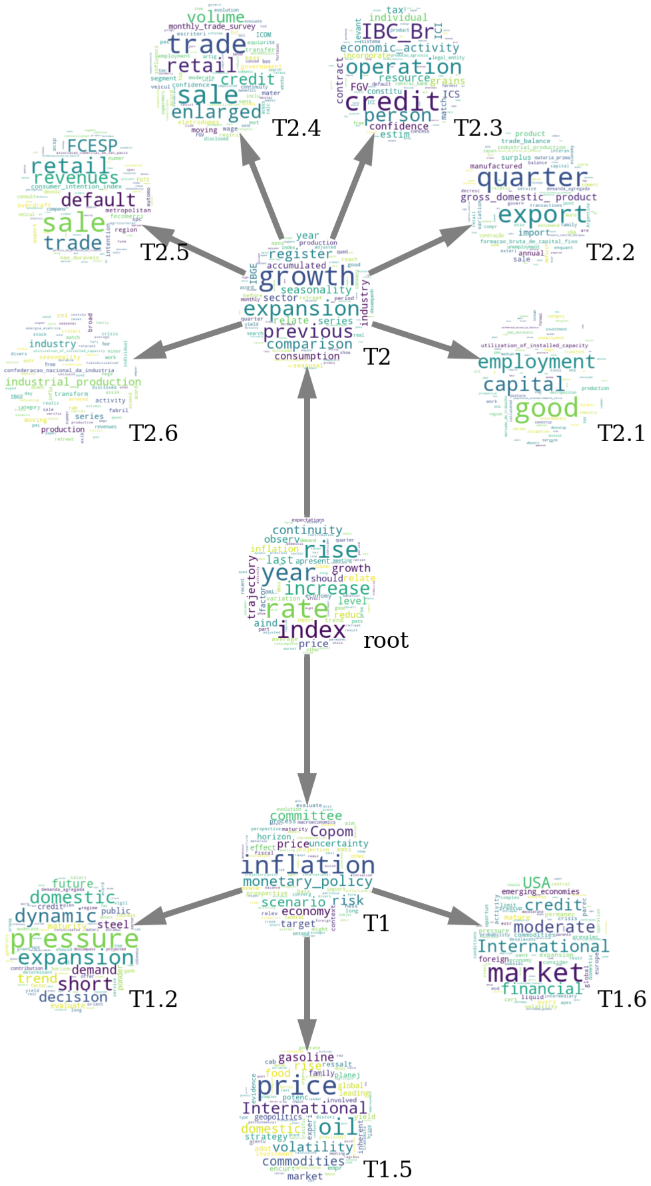


The same procedure applied to economic activity results in the wordcloud presented in Figure 7. Five leaves characterize the branch, with clear topics on labor market (T2.1), foreign trade (T2.2), credit and confidence (T2.3), retail and wholesale trades (T2.4 and T2.5) and industrial production and 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 describes "unemployment" with a negative sentiment. A natural evolution of this work is using word embedding techniques, trying to disentangle the context where each word in entering 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-months changes). Contemporaneous correlation between the economic activity situation index and the real variables is significant but low in both cases: the correlation with industrial production 0.361 and with wholesale

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

Figure 7: hLDA Tree Paths for Economic Activity Index

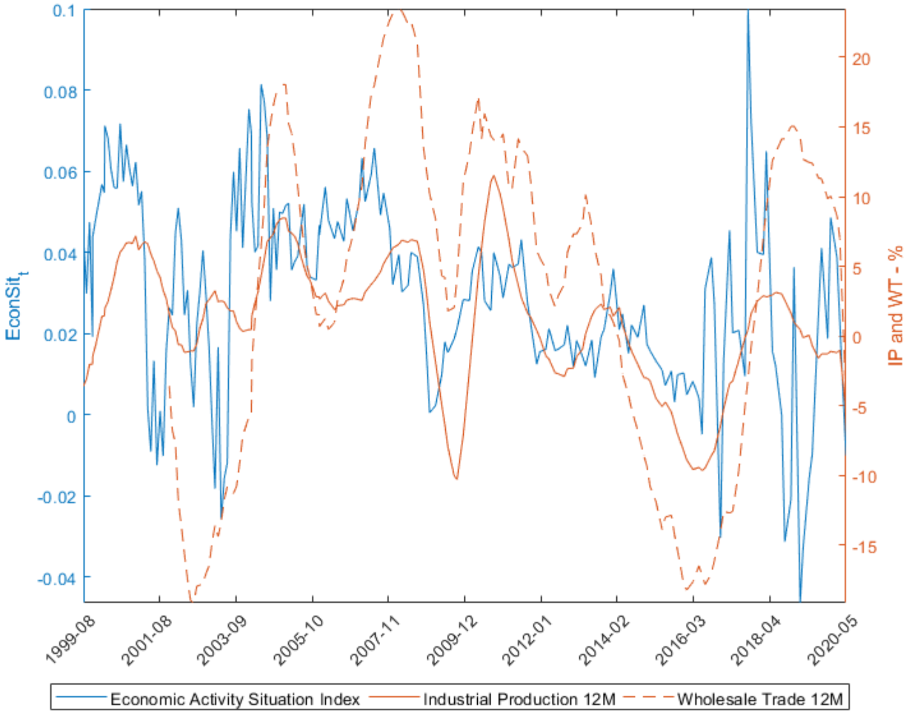


trade volume is 0.225.

It is worth noting the increase in volatility of the economic activity situation index after the 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 minutes of Copom's meetings, as previously mentioned, is one example of sentiment index targeting different features of Central Bank 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

Figure 8: Economic activity situation index (EconSit$_t$), industrial production and wholesale trade



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 over the whole tree. There are two main reasons to justify using the whole tree. First, the use of the whole tree helps smoothing the indicator over time, facilitating the comparison with other indexes measuring uncertainty[19]. Second, specific episodes responsible for increasing the uncertainty might be described outside of the tree path related to the monetary policy decision. Notably, topic 4, discussing mostly international financial markets issues related to the decision might contain information on the degree of uncertainty during the Great Financial Crisis (2008-09). Thus, eliminating this path might actually underestimate uncertainty in that period.

The dictionary, as it is standard in the literature, is the uncertainty word list of of Loughran and McDonald (2011)[20], 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 on Appendix C.

Figure 9 compares the economic uncertainty index with the Economic Policy Uncertainty Index for Brazil, based on calculations from Baker et al. (2016)[21]. The correlation of the economic uncertainty index with EPU is 0.4838 for the whole sample period, but it falls to 0.1497 after July-2016. The coefficient of variation suggests that 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 EPU.

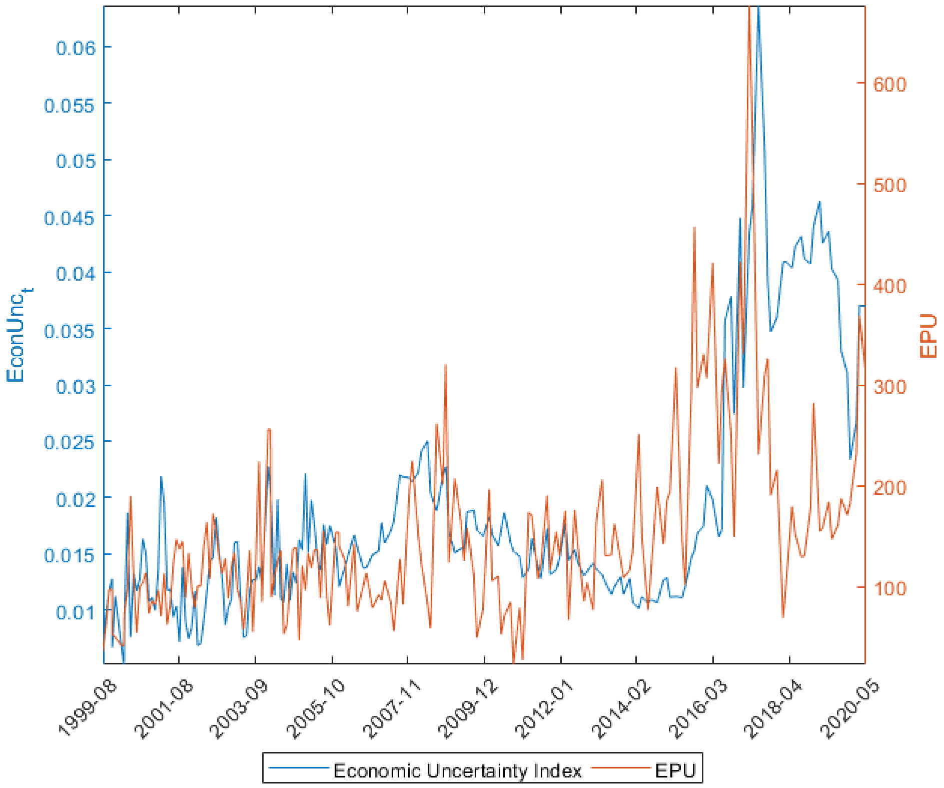
The increase in uncertainty after July-2016 and the decrease in correlation between the economic uncertainty index with EPU must be carefully considered. While both indexes show a build up in

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

[20] Loughran-McDonald Sentiment Word Lists. Available at: https://sraf.nd.edu/textual-analysis/resources/

[21] Data available at https://www.policyuncertainty.com/brazil_monthly.html.

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



uncertainty during 2016, as a consequence of the political crisis, the significant reduction in EPU between 2018-2019 not followed by the economic uncertainty index suggests that 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 that during the peak of the crisis, in 2016, both indexes moved closely. Other factors, like the start of the term of Governor Roberto Campos Neto in February 2019, could also influence this correlation.

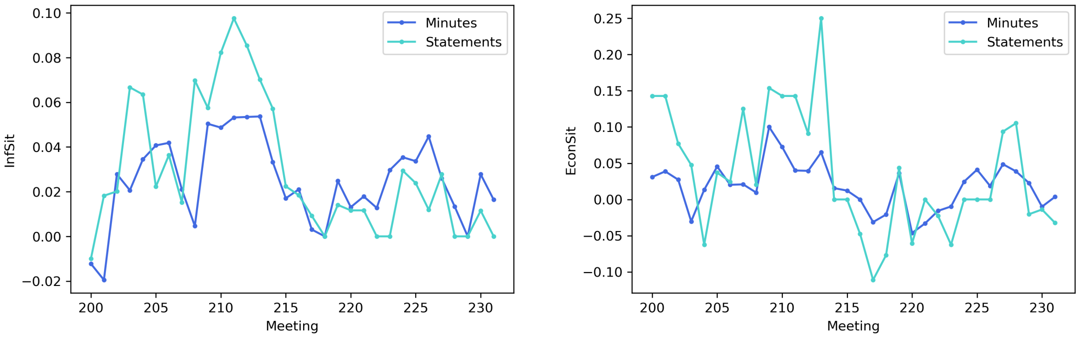
5.4 Statement and minute coherence

In this section, we explore the temporal structure of documents in Banco Central do Brasil's framework 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 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. This the temporal structure of documents allows for a simultaneous analysis, both in terms of a given MPC meeting and across consecutive meetings.

In order to build the exercise, the same hLDA model and its cloud of words from the previous section was applied to each sentence of the statements, generating indexes of inflation and economic situations 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 details in Table 1. 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.

Figure 10: Minutes and statement's indexes



(a) Inflation situation

(b) Economy situation

There are three hypothesis to be tested in the comparison of time series of indexes from the statements and from the minutes. First, if, for a given MPC meeting, the sentiment expressed in the minutes is consistent with the sentiment expressed in the previous week in statements. Second, if the sentiment in the previous MPC 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 allow 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. Defining Ind_t and $\widehat{\text{Ind}}_t$ as the situation index from minutes and from statements, respectively, and Weight_t and $\widehat{\text{Weight}}_t$ as the share of words in the document associated with the situation index from minutes and from statements, respectively, published after meeting t, the system has the following structure:

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

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

In equations above, 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 since 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. It also includes the (log) differences of nominal exchange rates and futures of interest rates between the days of the current and the last document of reference.[22]

Table 4 shows the baseline results (columns 2 and 3) and the estimation with three different controls for the inflation and economic activity situation indexes: the change in swap rates (columns 4 and 5), the (log) difference in nominal exchange rates (columns 6 and 7) and the economic uncertainty index (columns 8 and 9). The first notable result is the significance of parameter α_1 across all estimations. It means that the situation indexes from the statements play a key

[22] As an example, for equation describing the sentiment of the minutes, change in this variable is measured between the day after publication of the statement and the day before 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.

role in explaining the situation indexes in the minutes, as expected in a consistent communication procedure. Furthermore, the fact that α_2 is not significant in any of the estimations show that the previous minutes do not influence the current published document.

Table 4: Coherence of communication – SUR estimation

	InfSit_t	EconSit_t	InfSit_t	EconSit_t	InfSit_t	EconSit_t	InfSit_t	EconSit_t
			δSwap	δSwap	δER	δER	EconUnc	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%.

When analyzing the statements, a slightly different picture is shown: 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. On the other hand, β_2 is significant for all models with the inflation situation index, meaning that the previous statement contains information about the sentiment of inflation in the current meeting. Together, these results show that the language used in the statement is key in describing the Copom's sentiment with respect to inflation.

6 Conclusions

This paper estimated a hierarchical Latent Dirichlet (hLDA) model to analyze minutes from the Banco Central do Brasilâs 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 analysis of documents.

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 âpositiveâ and ânegativeâ 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 Brasilâs 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 whole 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 Index (EPU) for Brazil based on calculations from Baker et al. (2016). While both indexes show similar changes during the electoral crisis of 2002-03, the Great 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 that, 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

M. Acosta et al. FOMC responses to calls for transparency. Technical report, Board of Governors of the Federal Reserve System (US), 2015.

R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley Publishing Company, USA, 2nd edition, 2008. ISBN 9780321416919.

A. Bailey and C. Schonhardt-Bailey. Does deliberation matter in FOMC monetary policymaking? the volcker revolution of 1979. *Political Analysis*, 16(4):404–427, 2008.

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

D. M. Blei. Probabilistic topic models. *Commun. ACM*, 55(4):77–84, Apr. 2012. ISSN 0001-0782. doi: 10.1145/2133806.2133826. URL `http://doi.acm.org/10.1145/2133806.2133826`.

D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

D. M. Blei, T. L. Griffiths, and M. I. Jordan. The nested Chinese restaurant process and bayesian nonparametric inference of topic hierarchies. *J. ACM*, 57(2), Feb. 2010. ISSN 0004-5411. doi: 10.1145/1667053.1667056. URL `https://doi.org/10.1145/1667053.1667056`.

E. Boukus and J. V. Rosenberg. The information content of FOMC minutes. *Available at SSRN 922312*, 2006.

R. Cabral and B. Guimaraes. O comunicado do Banco Central. *Revista Brasileira de Economia*, 69 (3):287–301, 2015.

C. Carvalho, F. Cordeiro, and J. Vargas. Just words?: A quantitative analysis of the communication of the Central Bank of Brazil. *Revista Brasileira de Economia*, 67(4):443–455, 2013.

F. Chague, R. De-Losso, B. Giovannetti, and P. Manoel. Central bank communication affects the term-structure of interest rates. *Revista Brasileira de Economia*, 69(2):147–162, 2015.

M. Escobar and M. West. Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90(430), 06 1995. doi: 10.2307/2291069.

A. García-Herrero, E. Girardin, and E. Dos Santos. Follow what I do, and also what I say: monetary policy impact on Brazil's financial markets. *Economia*, 17(2):65–92, 2017.

T. L. Griffiths, M. I. Jordan, J. B. Tenenbaum, and D. M. Blei. Hierarchical topic models and the nested Chinese restaurant process. In *Advances in neural information processing systems*, pages 17–24, 2004.

R. S. Gürkaynak, B. Sack, and E. T. Swanson. Do actions speak louder than words? the response of asset prices to monetary policy actions and statements. *International Journal of Central Banking*, 2005.

S. Hansen and M. McMahon. Shocking language: Understanding the macroeconomic effects of central bank communication. *Journal of International Economics*, 99:114–133, 2016.

S. Hendry. Central bank communication or the media's interpretation: What moves markets? Technical report, Bank of Canada Working Paper, 2012.

S. Hendry and A. Madeley. Text mining and the information content of Bank of Canada communications. *Staff Working Papers, Bank of Canada*, 2010. URL https://www.bankofcanada.ca/2010/11/working-paper-2010-31/.

B. Jitmaneeroj, M. J. Lamla, and A. Wood. The implications of central bank transparency for uncertainty and disagreement. *Journal of International Money and Finance*, 90:222–240, 2019.

F. Labondance and P. Hubert. Central Bank sentiment and policy expectations. Sciences Po publications 648, Sciences Po, Mar. 2017.

T. Loughran and B. McDonald. When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *Journal of Finance*, 66(1):35–65, 2011. URL http://onlinelibrary.wiley.com/doi/10.1111/j.1540-6261.2010.01625.x/abstract.

D. O. Lucca and F. Trebbi. Measuring central bank communication: an automated approach with application to FOMC statements. Technical Report 15367, National Bureau of Economic Research, 2009.

G. Montes, L. Oliveira, A. Curi, and R. Nicolay. Effects of transparency, monetary policy signalling and clarity of central bank communication on disagreement about inflation expectations. *Applied Economics*, 48(7):590–607, 2016.

M. Neuenkirch. Managing financial market expectations: the role of central bank transparency and central bank communication. *European Journal of Political Economy*, 28(1):1–13, 2012.

V. M. Orengo and C. Huyck. A stemming algorithm for the portuguese language. In *Proceedings Eighth Symposium on String Processing and Information Retrieval*, pages 186–193, Nov 2001. doi: 10.1109/SPIRE.2001.989755. URL https://pdfs.semanticscholar.org/e9d9/ea5fc73013ff9d408b95c744d668896eb31b.pdf.

C. Rosa and G. Verga. On the consistency and effectiveness of central bank communication: Evidence from the ECB. *European Journal of Political Economy*, 23(1):146–175, 2007.

A. H. Shapiro and D. Wilson. Taking the Fed at its word: Direct estimation of central bank objectives using text analytics. Federal Reserve Bank of San Francisco, Federal Reserve Bank of San Francisco, 2019.

E. T. Swanson. Have increases in Federal Reserve transparency improved private sector interest rate forecasts? *Journal of Money, Credit, and banking*, 38(3):791–819, 2006.

Y. Yang and J. O. Pedersen. A comparative study on feature selection in text categorization. In *Proceedings of the Fourteenth International Conference on Machine Learning*, ICML '97, pages 412–420, San Francisco, CA, USA, 1997. Morgan Kaufmann Publishers Inc. ISBN 1558604863.

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 Letters
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

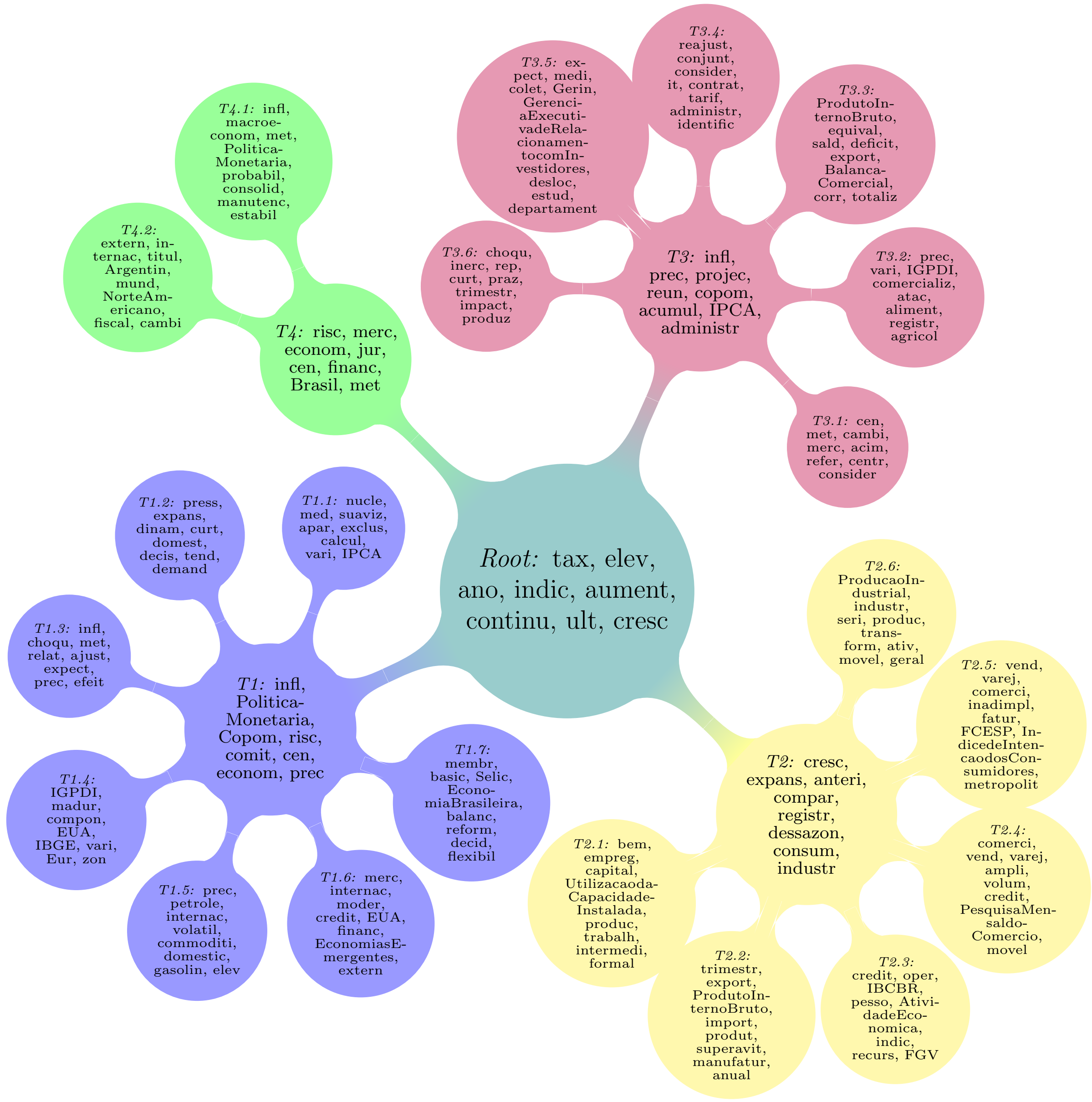
Positive		
Stemmed word	**Original word**	**Translation**
aceler	*aceleração, acelerar*	acceleration
adequ	*adequado*	adequate
acim	*acima*	above
-	*alto, alta, altos, altas*	high
aument	*aumento, aumentar*	increase
benign	*benigno*	benign
-	*compatível, compatíveis*	compatible
-	*confortável, confortáveis*	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


Root: tax, elev, ano, indic, aument, continu, ult, cresc
T1: infl, PoliticaMonetaria, Copom, risc, comit, cen, econom, prec
T1.1: nucle, med, suaviz, apar, exclus, calcul, vari, IPCA
T1.2: press, expans, dinam, curt, domest, decis, tend, demand
T1.3: infl, choqu, met, relat, ajust, expect, prec, efeit
T1.4: IGPDI, madur, compon, EUA, IBGE, vari, Eur, zon
T1.5: prec, petrole, internac, volatil, commoditi, domestic, gasolin, elev
T1.6: merc, internac, moder, credit, EUA, financ, EconomiasEmergentes, extern
T1.7: membr, basic, Selic, EconomiaBrasileira, balanc, reform, decid, flexibil
T2: cresc, expans, anteri, compar, registr, dessazon, consum, industr
T2.1: bem, empreg, capital, UtilizacaodaCapacidadeInstalada, produc, trabalh, intermedi, formal
T2.2: trimestr, export, ProdutoInternoBruto, import, produt, superavit, manufatur, anual
T2.3: credit, oper, IBCBR, pesso, AtividadeEconomica, indic, recurs, FGV
T2.4: comerci, vend, varej, ampli, volum, credit, PesquisaMensaldoComercio, movel
T2.5: vend, varej, comerci, inadimpl, fatur, FCESP, IndicedeIntencaodosConsumidores, metropolit
T2.6: ProducaoIndustrial, industr, seri, produc, transform, ativ, movel, geral
T3: infl, prec, projec, reun, copom, acumul, IPCA, administr
T3.1: cen, met, cambi, merc, acim, refer, centr, consider
T3.2: prec, vari, IGPDI, comercializ, atac, aliment, registr, agricol
T3.3: ProdutoInternoBruto, equival, sald, deficit, export, BalancaComercial, corr, totaliz
T3.4: reajust, conjunt, consider, it, contrat, tarif, administr, identific
T3.5: expect, medi, colet, Gerin, GerenciaExecutivadeRelacionamentocomInvestidores, desloc, estud, departament
T3.6: choqu, inerc, rep, curt, praz, trimestr, impact, produz
T4: risc, merc, econom, jur, cen, financ, Brasil, met
T4.1: infl, macroeconom, met, PoliticaMonetaria, probabil, consolid, manutenc, estabil
T4.2: extern, internac, titul, Argentin, mund, NorteAmericano, fiscal, cambi