Universidade Federal Fluminense

FACULDADE DE ECONOMIA

GUSTAVO DE OLIVEIRA VITAL

APLICAÇÕES DE TÉCNICAS DE ANÁLISE DE SENTIMENTOS ÀS ATAS DO COMITÊ DE POLÍTICA MONETÁRIA: COMPARAÇÃO DOS PERÍODOS DE 2003 À 2018

NITERÓI – RJ

2019

GUSTAVO DE OLIVEIRA VITAL

APLICAÇÕES DE TÉCNICAS DE ANÁLISE DE SENTIMENTOS ÀS ATAS DO COMITÊ DE POLÍTICA MONETÁRIA: COMPARAÇÃO DOS PERÍODOS DE 2003 À 2018

Monografia apresentada ao curso de Bacharelado em Ciências Econômicas da Universidade Federal Fluminense como requisito parcial para conclusão do curso.

Orientador:
Prof. Dr. Jesús Alexei Luizar Obregon

Niterói – RJ

2019

Ficha catalográfica automática - SDC/BEC
Gerada com informações fornecidas pelo autor

V836a Vital, Gustavo de Oliveira
Aplicações de Técnicas de Análise de Sentimentos às Atas do Comitê de Política Monetária : Comparação dos períodos de 2003 à 2018 / Gustavo de Oliveira Vital ; Jesús Obregon, orientador. Niterói, 2019.
52 f. : il.

Trabalho de Conclusão de Curso (Graduação em Ciências Econômicas)-Universidade Federal Fluminense, Faculdade de Economia, Niterói, 2019.

1. Analise de sentimentos. 2. Mineração de texto. 3. COPOM. 4. Kullback-Liebler. 5. Produção intelectual. I. Obregon, Jesús, orientador. II. Universidade Federal Fluminense. Faculdade de Economia. III. Título.

CDD -

Bibliotecária responsável: Cláudia Maria Gomes Curi - CRB7/4192

GUSTAVO DE OLIVEIRA VITAL

APLICAÇÕES DE TÉCNICAS DE ANÁLISE DE SENTIMENTOS ÀS ATAS DO COMITÊ DE POLÍTICA MONETÁRIA: COMPARAÇÃO DOS PERÍODOS DE 2003 À 2018

Monografia apresentada ao curso de Bacharelado em Ciências Econômicas da Universidade Federal Fluminense como requisito parcial para conclusão do curso.

Trabalho aprovado em 03 de dezembro de 2019

BANCA EXAMINADORA

Prof. Dr. Jesús Alexei Luizar Obregon
Orientador
Universidade Federal Fluminense

Prof. Dr^{a}. Danielle Carusi Machado
Universidade Federal Fluminense

Prof. Dr. Emmanoel de Oliveira Boff
Universidade Federal Fluminense

AGRADECIMENTOS

Agradeço a todas as pessoas que me apoiaram, sabendo que o fizeram ou não.

Em especial, agradeço aos meus pais, Cesar e Katia por sempre terem me incentivado, apoiado, e por todas as oportunidades que me deram, proporcionando que eu pudesse estudar em uma universidade pública, gratuita e de qualidade. Agradeço também a todos aqueles que me acompanharam e acompanham desde o meu primeiro período na faculdade. Agradeço ao Programa de Educação Tutorial que, por mais de uma vez, me serviu como base para continuar na faculdade – em especial aos dois tutores que tive nesse projeto: Renaut Michel e Lérida Povoleri, que infelizmente nos deixou neste ano de 2019.

Não poderia deixar de agradecer, além disso, à professora Danielle Carusi, que sempre me ajudou e aconselhou quando eu precisava, bem como pela oportunidade que me deu em contribuir para a Revista Econômica. Por fim, agradeço ao meu orientador, professor Jesús Alexei, por todo incentivo e ajuda que me deu, esperando sempre o melhor de mim e estando sempre presente para qualquer dúvida – antes mesmo de ser meu orientador.

RESUMO

Este estudo apresenta uma comparação das expressões do Comitê de Política Monetária, levando em consideração as publicações das atas disponíveis após cada reunião, para o período de 2003 à 2018 – isto é, do período de gestão de Meirelles ao período de gestão Goldfajn. A estrutura metodológica e analítica deste trabalho apresenta as técnicas de coleta e mineração de dados; a divergência de Kullback-Liebler; índices de sentimentos; e modelos de vetores auto-regressivos, bem como a função de resposta ao impulso. Os resultados deste trabalho indicam que efetivamente a forma de expressão da política monetária é relacionada com o período a ser analisado, sendo demonstrado que, de fato, a as distribuições de termos e palavras diferem de acordo com o cenário político brasileiro (PT-PMDB), bem como índices de sentimentos podem ser considerados em modelos macroeconômicos para predizer fenômenos econômicos.

Palavras-chave: Analise de sentimentos; Mineração de texto; COPOM; Kullback-Liebler.

ABSTRACT

This study compares the expressions of the Monetary Policy Committee, taking into account the publications of the minutes available after each meeting, from 2003 to 2018 - that is, from the Meirelles management period to the Goldfajn management period. The methodological and analytical structure of this paper presents the techniques of data collection and mining; the Kullback-Liebler divergence; indices of feelings; and autoregressive vector models, as well as the impulse response function. The results of this work indicate that the form of monetary policy expression is effectively related to the period to be analyzed, showing that, in fact, the distributions of terms and words differ according to the Brazilian political scenario (PT-PMDB), as well as sentiment indices can be considered in macroeconomic models to predict economic phenomena.

Keywords: Sentiment analysis; Text mining; COPOM; Kullback-Liebler.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

É sabido que a tomada de decisões em diversas áreas econômicas têm, por vezes, como referência boletins e resoluções de bancos centrais. Como todo documento textual, atas e boletins contém informações objetivas e subjetivas quantificáveis possibilitando uma relação com a atividade econômica; política econômica ou mesmo uma análise específica em relação à análise de conjuntura de um certo período - no que se refere as informações objetivas. É possível, através de modelagem econométrica, mapear posições de agentes econômicos, bem como trazer expectativas dos respectivos comportamentos.

Naturalmente, pelas características das informações dos agentes, é esperado que o mapeamento destes fenômenos econômicos não siga o mesmo padrão diante diferentes momentos conjunturais.

Em geral, análise de sentimentos pode ser considerada uma técnica computacional de manipulação e análise de dados que consiste em extrair e classificar informações contidas em textos naturais. O objetivo é encontrar opiniões, expressões, e mensagens que um ou mais textos podem transmitir - dado um texto, classificá-lo como positivo ou negativo, por meio de identificações de padrões e características desse texto.

Ainda, é possível melhor entender a abordagem relativa a um texto dado um cenário econômico, social ou geo-político. Até então, pouco utilizada no âmbito econômico, essa técnica é comumente relacionada aos campos das ciências políticas e de marketing. Um dos principais motivos de não se trabalhar com análise de sentimentos em questões econômicas é não ser óbvio - a princípio - que textos podem ser analisados e classificados como dados quantitativos, o que é uma contradição frente ao aspecto de bancos centrais utilizarem ferramentas e técnicas estatísticas que permitem esse feito (BHOLAT et al., 2015).

Um outro fator a ser discutido é a possibilidade de realização dessa técnica. Como se trata, majoritariamente, de uma técnica computacional e comumente é relacionado ao campo de *big data*, seria inviável a utilização dessa sem uma adequada capacidade de processamento e armazenamento de dados, o que explica um crescimento recente no uso da análise de sentimentos - visto que cada vez mais computadores são capazes de processar e armazenar informações.

A presente monografia visa estudar no contexto de análise de sentimentos as expressões das atas do Comitê de Política Monetária (COPOM) no período de 2003-2018, Lula-Temer. Dessa forma, a análise feita é referente aos três diferentes períodos de presidência do Banco Central: Henrique Meirelles; Alexandre Tombini; e

Ilan Goldfajn. É verificada uma mudança nos padrões das distribuições relativas aos períodos das palavras que mais aparecem. Para a contestação das mudanças referentes as distribuições, utiliza-se a divergência de Kullback-Leibler (também conhecida como entropia relativa) (KULLBACK; LEIBLER, 1951).

O segundo capítulo deste trabalho contém essencialmente uma explicação mais aprofundada sobre análise de sentimentos e *text mining*, evidenciando a importância dessas duas técnicas como parte fundamental e específica do campo da ciência de dados, bem como suas aplicabilidades. Ainda, é feita uma revisão dos tipos e características das formas de realizar essa técnica, seja ela com o uso - ou não - de dicionários de *stopwords*, ou algoritmos de *machine learning*. Neste capítulo, discute-se, também, a abordagem da análise de sentimentos por bancos centrais e pesquisas com base em trabalhos empíricos já realizados.

O terceiro caítulo se inicia com uma explicação das principais técnicas utilizadas nesta monografia. São elas: o *web scraping* empregado, bem como o funcionamento de um algoritmo de *web scraping*; tratamento dos dados e a metodologia utilizada para o tratamento dos mesmos; o dicionário de *stopwords* utilizado e o porquê de utiliza-lo; e a metodologia utilizada em relação a análise de sentimentos em si. Além disso, é explicado o porquê da utilização da linguagem R na utilização desse trabalho, bem como os pacotes dessa linguagem contidos no projeto.

O quarto capítulo, por sua vez, apresenta os resultados empíricos do trabalho. Entre esses podemos citar comparações entre as frequências de palavras; a ocorrência geral das palavras corrigidas pelos tamanhos das atas; comparações das principais palavras frentes aos momentos de conjunturas analisados; e a análise de sentimentos nas atas do COPOM.

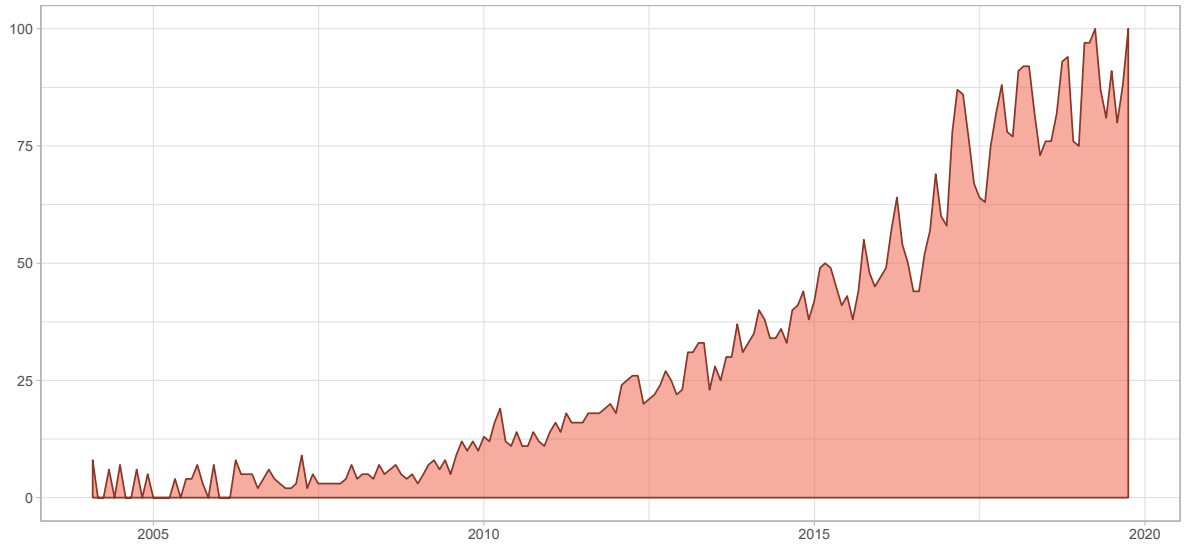
Finalizamos com o quinto capítulo, evidenciando as considerações finais do trabalho e as conclusões referentes ao que foi feito.

2 REVISÃO DA LITERATURA

2.1 Preliminar

A crescente pesquisa por análise de sentimentos e *text mining* tende a corroborar, de forma geral, para um melhor entendimento do que acontece no cenário mundial econômico, por meio de *proxys* que por muitas vezes não podem ser captadas senão de forma computacional. Compreender melhor a demanda por alguma coisa torna-se mais fácil, no quesito em que é possível automatizar resoluções de problemas econômicos e não econômicos. Como um próprio exemplo disso, podemos nos fazer valer da própria busca pelo termo *sentiment analysis*. A Figura 1 apresenta a evolução da busca pelo termo *sentiment analysis*, desde 2004 até o mês de outubro de 2019 – A busca foi realizada no dia primeiro de novembro de 2019, com os seguintes parâmetros: 1 - termo de pesquisa: sentiment analysis; 2 - foi aplicado um período de busca de 2004 até o mês de outubro de 2019; Para todo o mundo, todas as categorias, e pesquisa na web. No volume de busca, 100 representa o máximo volume como referência.

Figura 1 – Evolução da busca pelo termo *sentiment analysis*, de janeiro de 2004 até outubro de 2019



Fonte: Google trends. Elaboração própria.

Ainda, acrescenta-se, mesmo que atualmente o volume de buscas por esse termo seja quase insignificante para o Brasil (cerca de 4), países como a Índia ou a China apresentam volumes de buscas muito maiores: 81 e 22, respectivamente. A Índia é o pais que mais busca por esse termo. Tal qual, pode ser uma relação direta na crescente potencialidade tecnológica e de pesquisa deste país.

2.2 Mineração Textual

Processamento de linguagem natural, conhecido também como *Text Minning*, é um conjunto de procedimentos que utiliza ferramentas computacionais e técnicas estatísticas que permitem quantificar textos. Este processo de quantificação permite intuir algum significado textual – indubitavelmente, por meio de um computador a análise e organização dos significados textuais podem ser feitas de formam muito mais precisa e automática do que de forma manual. O *text mining*, ainda, possibilita a extração de *significados* de textos, que podem ser difíceis de se identificar quando analisados sem o auxílio de um computador – nos permite detectar padrões pouco prováveis ante às perspectivas humanas.

Mesmo quando muito usada em outros campos científicos, como ciências políticas e marketing, a técnica de *text minning* ainda é pouco utilizada nas ciências econômicas como enuncia Bholat et al. (2015)

> "Primeiro, pode não ser óbvio que o texto possa ser descrito e analisado como dados quantitativos. Como resultado, provavelmente existe uma falta de familiaridade nos bancos centrais com as ferramentas e técnicas estatísticas que tornam isso possível. Segundo, mesmo que os banqueiros centrais tenham ouvido falar em mineração de texto, eles já têm acesso a outros dados quantitativos prontamente disponíveis. A oportunidade e outros tipos de custos de transformar textos em dados quantitativos e aprender novas ferramentas e técnicas para analisar esses dados podem ser vistos como superando os benefícios esperados. [1]" (BHOLAT et al., 2015, p.1)

Assim, a inserção do *text mining* no campo da ciência econômica se apresenta de forma tardia, devido a obstáculos computacionais e por vezes desconfiança por parte dos agentes econômicos acostumados com informações quantitativas tradicionais. Uma análise a partir destes pode, ainda mais, incluir aspectos não convencionais em modelos micro e macroeconômicos, como possíveis *proxys* de conjuntura política, interesses populacionais ou mesmo opiniões públicas. É notável, então, que a utilização de técnicas de *text mining* poderia ser utilizada de maneira extremamente proveitosa por, por exemplo, bancos centrais. Isso é, extrair informações de formas não usuais de fontes que permitam, de alguma forma, avaliar estabilidade monetária e financeira de maneira quantitativa. Ou seja, a partir de notícias, jornais, artigos científicos, relatos de inteligência de mercado, ou mesmo relatórios empresariais, afim de quantificar estes dados para uma melhor obtenção da situação conjuntural vigente.

[1] First, it may not be obvious that text can be described and analysed as quantitative data. As a result, there is probably a lack of familiarity in central banks with the tools and statistical techniques that make this possible. Second, even if central bankers have heard of text mining, they already have access to other readily available quantitative data. The opportunity and other types of costs from transforming texts into quantitative data, and learning new tools and techniques to analyse these data, may be viewed as outweighing the expected benefits. (BHOLAT et al., 2015, p.1)

A análise de um documento, ou um conjunto de documentos (tecnicamente chamado de *corpus*) se daria, por exemplo, a partir de boletins informacionais de órgãos de pesquisa (vide por exemplo uma carta de conjuntura do Ipea - Instituto de Pesquisa Econômica Aplicada) ou mesmo atas de instituições financeiras.

Mesmo que ainda não seja uma técnica usual de análise, bancos centrais já se fazem valer de benefícios do *text mining* diariamente. Uma busca online por determinados termos econômicos, dependendo da forma como é feita, ou mesmo proteção contra *cyber-attacks* ou busca por uma base de dados de citações, no âmbito acadêmico, são exemplos usuais da funcionalidade do *text mining* (BHOLAT et al., 2015).

2.3 Análise de Sentimentos

Em um recente artigo publicado pelo Banco da Inglaterra, Bholat et al. (2015), é relatado os interesses dos bancos centrais e como estes, por meio textuais, não são devidamente aproveitados. Um exemplo de aplicação em relação a estes volumes de dados é a própria pesquisa realizada na internet, em termos econômicos ou mesmo em relação ao volume de buscas. Ainda podemos citar a correlação entre Jobseeker's Allowance (JSA) e a taxa oficial de desemprego no Reino Unido, como mostra McLaren e Shanbhogue (2011):

> "É notável que os 'empregos', que provavelmente foram pesquisados por quem está dentro e fora do emprego, não aumentaram muito durante a recessão. As pesquisas por "desempregados"aumentaram acentuadamente durante a recessão. O termo "JSA"(sigla para Subsídio de Desemprego) foi escolhido porque seus movimentos melhor se correlacionaram com os dos dados oficiais. Também é provável que seja usado por aqueles que pensam que podem em breve ficar desempregados e, portanto, buscam mais informações sobre benefícios de desemprego.[2]" (MCLAREN; SHANBHOGUE, 2011, p.136)

Faz sentido uma busca na internet por "emprego" quando a pessoa está desempregada, bem como por "auxílio desemprego" (Jobseeker's Allowance). No mesmo artigo, os autores vão além: é possível identificar, ainda, uma correlação entre "inflação imobiliária" e "agentes imobiliários". Assim, a partir de expressões como "preços de casas"; "comprar imóvel"; e "vender imóvel" é possível obter uma *proxy* para a demanda por moradia:

[2] It is notable that 'jobs', which is likely to have been searched for by both those in and out of employment, did not increase much during the recession. Searches for 'unemployed' rose markedly during the recession. The term 'JSA' (acronym for Jobseeker's Allowance) was chosen because its movementsbest correlated with those in the official data. It is also a termlikely to be used by those who think they may soon become unemployed and so search for more information on unemployment benefit. (MCLAREN; SHANBHOGUE, 2011, p.136)

> "Para o mercado imobiliário, segue-se uma abordagem semelhante à adotada acima para o mercado de trabalho. Uma proporção significativa de pesquisas relacionadas a habitação é para sites de empresas específicas. No entanto, essas pesquisas variam ao longo do tempo, dependendo da popularidade de cada site. Portanto, uma ampla variedade de termos de pesquisa genérica é considerada (incluindo "preços de casas", "casa de compra", "casa de venda", "hipoteca"e "agentes imobiliários"). Os termos de pesquisa 'comprar casa' e 'vender casa' foram inicialmente considerados, pois capturariam a demanda e a oferta de casas.[3]" (MCLAREN; SHANBHOGUE, 2011, p.137)

Em ambos os casos, os autores obtiveram exito em suas pesquisas, o artigo "Using internet search data as economic indicators" (MCLAREN; SHANBHOGUE, 2011, p.136), publicado também pelo Banco da Inglaterra, apresenta resultados econométricos significativos frente a regressões utilizando variáveis reais regredidas em variáveis obtidas por meio de volume de buscas online. Isso é, seja em relação à taxa de desemprego como em relação à inflação imobiliária, ambas as séries puderam ser explicadas por meio das séries dos volumes de pesquisa online – Nos exemplos citados no texto, o mecanismo de busca utilizado foi o *google trends*.

Outra maneira de se aproveitar algorítimos e volumes de buscas online é levar em consideração medidas de risco e incerteza no âmbito econômico e financeiro dado um determinado momento. Uma recente contribuição nessa direção foi apresentada por Nyman et al. (2018): é proposta uma teoria sobre "hipótese financeira emocional", a qual sustenta que "os indivíduos se convencem a assumir posições nos mercados financeiros, criando narrativas sobre os possíveis resultados de suas ações"[4]. Os autores correlacionam emoções de "exitação" com ganhos financeiros e "ansiedade" com possíveis perdas. Essas narrativas, entretanto, não são compostas de simples textos, e sim de interações sociais entre agentes. As hipóteses foram testadas a partir de três textos fundamentais: the Bank's daily market commentary (2000-2010), broker research reports (2010-2013) and the Reuters' News Archive (1996-2014). A partir disto, os autores propõe um índice de medida de sentimento textual:

$$SI_t = \frac{N_e - N_a}{N_t}$$

onde N_e e N_a representam o número de palavras correlacionadas com os estados de, respectivamente, excitação e ansiedade. N_t denota o número total de palavras do docu-

[3] For the housing market a similar approach is followed to thattaken above for the labour market. A significant proportion ofhousing-related searches are for specific companies' websites.However, these searches vary over time depending on thepopularity of each website. So a wide range of more genericsearch terms are considered (including 'house prices', 'buyhouse', 'sell house', 'mortgage' and 'estate agents'). The searchterms 'buy house' and 'sell house' were initially considered,since they would capture the demand for and supply of houses. (MCLAREN; SHANBHOGUE, 2011, p.137)

[4] "individuals gain conviction to take positions in financial markets by creating narratives about the possible outcomes of their actions" (NYMAN et al., 2018; BHOLAT et al., 2015)

mento. O sinal do índice nos fornece um indicativo do que acontece no mercado: sendo esse positivo é simbolizado crescimento financeiro; caso negativo, uma retração no mercado financeiro – respectivamente *bullish* e *bearish*. O índice, então, é comparado com outros eventos históricos e outros indicadores financeiros (BHOLAT et al., 2015).

Uma vez que a incerteza econômica é medida, esta poderia ser utilizada como uma variável explicativa em modelos de bancos centrais - essa torna-se, então, um possível indicador quanto às orientações de política monetária de bancos centrais.

A análise de sentimentos pode ser definida como o resultado de uma sequência hierarquizada de processos de "classificação". Por sua vez, a classificação é entendida como uma função, ou domínio, em um conjunto de entidades e imagens em um conjunto binário: positivo, negativo. Indo além, existem três principais níveis de classificação em análise de sentimento: documento, aspecto e sentença.

> " Em documentos o objetivo é classificar a opinião do documento como expressando um sentimento ou opinião positiva ou negativa. É considerado o documento em sua totalidade como uma unidade de informação (falando sobre um tópico). Com relação às sentenças, o objetivo é classificar o sentimento expresso em cada sentença. O primeiro passo é identificar se a sentença é subjetiva ou objetiva. Se a sentença é subjetiva a análise determinará se a sentença expressa opiniões positivas ou negativas. Já em nível de aspecto, o alvo é classificar o sentimento com relação as aspectos específicos das entidades a partir da identificação das entidades e seus aspectos dada a possibilidade de existir diferentes opiniões sobre aspectos distintos da mesma entidade (como em: A qualidade da chamada do telefone não é boa, mas a bateria dura muito tempo). " (COSTA, 2016, p.12)

Em geral, pode-se classificar algorítimos de análise de sentimentos de duas formas: baseados em *machine learning*, sendo possível dividir esses algorítimos em forma de aprendizado supervisionado e aprendizado não supervisionado; a outra forma é trabalhar por meio de dicionários (*léxicos*). Como exemplo, podemos citar o dicionário `qdap` contido no pacote de mesmo nome do R implementado por Rinker (2013) que pode determinar, segundo seus critérios de análise, se uma palavra pode ser classificada como positiva ou negativa.

Para exemplificarmos a utilização de um dicionário léxico, um trabalho recente de Costa (2016), apresenta uma análise de sentimentos para as atas do Comitê de Política Monetária para o período de 2000 à 2016. No estudo, um *corpus* é composto por 157 atas e analisado e, por meio deste, um índice é proposto, onde I_t é o índice de sentimento:

$$I_t = \frac{NP_t - NN_t}{N} \quad , \tag{2.1}$$

para cada ata divulgada em t, NP_t é a quantidade de palavras "positivas", NN_t a quantidade de palavras "negativas", e N a quantidade de palavras na ata (COSTA, 2016, p.13).

Os autores chegam a conclusão, finalmente, de uma correlação entre determinadas variáveis macroeconômicas (taxa de juros Selic, IPCA, IPCA Meta) e o índice para tomada de decisão da autoridade monetária. A correlação se apresenta de maneira mais forte no que diz respeito ao comportamento de longo prazo dessas variáveis. Em períodos de alta de inflação há uma diminuição do *score* do índice de sentimentos, o que representa uma maior cautela na expectativa econômica representada pela maior quantidade de palavras "negativas" utilizadas nas atas divulgadas no período analisado. Comportamento análago ocorre quando se compara o *score* do índice com a taxa de juros Selic.

2.4 Aplicações Econométricas

Formuladores de políticas econômicas (*policy makers*) e aqueles que participam do mercado confiam amplamente em uma variedade de modelos que incorporam o que é chamado de informação branda. Ao contrário de informações complexas, que incluem variáveis objetivas e diretamente quantificáveis (como produção e taxa de desemprego), informações brandas influem medidas subjetivas relativas a atitudes em relação às condições econômicas atuais e futuras. Há, assim, uma ampla variedade de variáveis flexíveis disponíveis, mas sem dúvida as mais amplamente levadas em consideração são as medidas relativas às confianças de mercado e sentimento do consumidor (SHAPIRO; SUDHOF; WILSON, 2018).

No artigo "Measuring News Sentiment" de Shapiro, Sudhof e Wilson (2018), mais um índice de sentimentos foi proposto levando em consideração a estimação dos efeitos de positividade em artigos de forma mensal. Isto é, os autores consideram positividades em artigos em jornais de cunho econômico.

Este índice de sentimento proposto foi utilizado como um exercício de aplicação, relacionando-o com a atividade econômica dos Estados Unidos. Neste exercício, é avaliado se especificamente a positividade do índice surte algum efeito na atividade econômica futura. Para isso foi utilizado o o método de projeção local proposto por Jordà (2005), similar a um vetor auto-regressivo (VAR), porém, menos restritivo. De forma geral, este método analisa como um choque do novo índice de sentimentos afeta um dado nível de atividade econômica. O novo choque do índice de sentimentos é construído como um componente da nova série de sentimentos que é ortogonal a atual e a seis defasagens de atividade econômica bem como a seis defasagens de si mesmo. Isso é, para cada previsão num horizonte h, uma regressão diferente é estimada para cada valor da atividade econômica calculada (y_j) no momento respectivo e defasado

do novo índice de sentimentos e de outras quatro medidas econômicas (consumo, produção, taxa de juros real, e inflação) (SHAPIRO; SUDHOF; WILSON, 2018).

Feita a estimação, chega-se a conclusão que um choque positivo no índice de sentimentos, afeta positivamente o consumo, bem como na produção, e na taxa de juros real do FED. Houve, entretanto, uma leve redução para o nível de preços. O efeito no nível de preços é transitório, mas os efeitos no consumo, na produção e na taxa real dos fundos são mais duradouros, aumentando gradualmente até 12 meses após o choque, nas palavras de Shapiro, Sudhof e Wilson (2018).

> "Estender o horizonte ainda mais [...] indica que as respostas de consumo, produção e taxa real atingem um pico entre 12 e 18 meses após o choque, antes de diminuir gradualmente.[5]" (SHAPIRO; SUDHOF; WILSON, 2018, p.19-20)

Isso é, é possível avaliar uma notável melhora nas variáveis macroeconômicas, ainda, após os 12 meses a frente estimados no impulso resposta.

Em outro artigo Shapiro, Sudhof e Wilson (2018) tiveram como inspiração um segundo artigo de Barsky e Sims (2012) – que apresenta resultados semelhantes, foi verificado que um choque positivo de sentimentos leva a um aumento persistente em consumo, produção, e taxa real de juros; mas resulta em uma queda na inflação. A similaridades dos resultados, assim, fortalece a hipóteses que um possível índice de sentimentos tem medidas similares em impactos macroeconômicos, como por exemplo, o índice de sentimentos do consumidor.

2.5 Entendendo a Função Objetivo de um Banco Central por meio de Análise de sentimentos

Em um outro estudo recente a abordagem em relação ao *score* de sentimentos foi diferente. O objetivo do artigo "Taking the Fed at its Word: A New Approach to Estimating Central Bank Objectives using Text Analysis" (SHAPIRO; WILSON, 2019) é entender, por meio de publicações, qual é a função objetivo de um banco central – sendo essa uma importante questão macroeconômica a ser tratada. A literatura atual, por exemplo Walsh (2017) pressupõe que a abordagem canônica, frente aos modelos macroeconômicos, assumem uma forma quadrática ao tratamento da inflação, e meta inflacionária.

> "Embora exista amplo consenso sobre como deve ser a função do objetivo do banco central, com base na ampla literatura sobre política monetária ideal, houve muito pouco estudo sobre o que realmente é a

[5] Extending the horizon out further [...] indicates that the responses of consumption, output, and the real rate peak between 12 and 18 months after the shock before gradually waning.

> função do objetivo do banco central na prática[6]" (SHAPIRO; WILSON, 2019, p.2)

A escassez de análises positivas da função objetivo do banco central é surpreendente, considerando que é implicitamente a base subjacente às regras de política monetária (SHAPIRO; WILSON, 2019). Ainda, a dispersão de análises não se deve a uma crença de que a função objetiva de um banco central é bem entendida. Um exemplo disto é a forma própria forma funcional, mesmo considerando os parâmetros, não é muito bem aceita. Segundo Blinder (1997):

> "macroeconomistas acadêmicos tendem a usar funções de perda quadrática por razões de conveniência matemática, sem pensar muito em suas implicações substantivas. A suposição não é inócua ... banqueiros centrais e acadêmicos práticos se beneficiariam de um pensamento mais sério sobre a forma funcional da função de perda[7](BLINDER, 1997, p.6)"

O autor propõe, assim, uma nova abordagem na estimação dos parâmetros objetivos de um banco central – a partir de um índice de negatividade construído por meio das discussões internas do U.S. Federal Open Market Committee's (FOMC). A medida de negatividade foi baseada fundamentalmente em dicionários (léxicos) criados especificamente para economia/finanças Loughran e McDonald (2011), contém milhares de palavras e termos econômicos.

Desta forma, para cada expressão de cada encontro do FOMC foi construído uma medida de negatividade baseada na frequência utilidade de palavras positivas e negativas. Para medir as variáveis que potencialmente entram na função de perda de curto prazo do FOMC, os autores usam dados em tempo real nas previsões da equipe do Federal Reserve (Greenbook) sobre as principais inações de variáveis econômicas reais.

Assim, o exercício proposto questiona dois pontos cruciais sobre as preferências do FOMC:

1. Foi indicado que o FOMC tinha como meta de inflação cerca de $1\frac{1}{2}$% em média no período de 2000-2013. Foi estimado que nesse período, a meta de inflação está significantemente abaixo de 2%. Dessa forma, chega-se a um *gap* inflacionário positivo. Além disso, está abaixo, também, da própria meta anunciada pelo

[6] Although there is broad consensus on what the central bank objective function should look like based on the large literature on optimal monetary policy, there has been very little study of what the central bank objective function actually is in practice (SHAPIRO; WILSON, 2019, p.2).

[7] academic macroeconomists tend to use quadratic loss functions for reasons of mathematical convenience, without thinking much about their substantive implications. The assumption is not innocuous...practical central bankers and academics would benefit from more serious thinking about the functional form of the loss function (BLINDER, 1997, p.6).

"Statement on Longer-Run Goals and Monetary Policy Strategy", que também corresponde ao valor de 2%.

Dada essa diferença entre o valor estimado e a meta para inflação de $1\frac{1}{2}$% e o valor convencional de 2%, é completada a regressão com uma análise narrativa que identifica e tabula os casos em que os participantes do FOMC declararam uma preferência para a meta de inflação. Embora a preferência para a meta declarada seja conceitualmente distinta da meta de inflação implícita consistente com o tom geral das discussões do comitê, foi chegado ao consenso que a preferência para a meta era de $1\frac{1}{2}$% para a maior parte do período analisado. Entretanto, também foi documentada uma mudança de 2% no final da Grande Recessão, tal qual o consenso para o período teria sido, realmente, uma meta para inflação de 2% (SHAPIRO; WILSON, 2019)[8].

2. Em contraste às típicas formulações de função de perda do banco central, que a perda do FOMC está monotonicamente reduzindo a atividade econômica. Especificamente, os resultados apontam que a perda em FOMC decresce em relação ao crescimento[9] e performance no mercado financeiro. Assim, uma função objetivo, formulada por Barro e Gordon (1983), e descrita por Walsh (2017), tem uma importante implicação:

> "o banco central está disposto a negociar uma diferença positiva de inflação em troca de uma atividade real mais alta[10]"(SHAPIRO; WILSON, 2019, p.4)

Isso é, um hiato positivo da inflação no estado estacionário é teoricamente consistente com preferências lineares sobre a atividade real. Os resultados empíricos, portanto, coincidem com as previsões de um modelo simples novo keynesiano com uma função de perda Shapiro e Wilson (2019).

Percebe-se, então, que uma abordagem utilizando uma análise textual acaba por complementar o estudos das preferências do banco central. Anteriormente as análises

[8] The upward shift after 2008 seen in the narrative analysis begs the question of whether the FOMC's implicit inflation target also increased around that time. Though the ability to identify a post-2008 break in the inflation target is somewhat limited by the short time series dimension [. . .]

[9] We find little evidence that the loss function depends on the level of slack or the quadratic of slack. While the finding that the FOMC appears to care more about output growth than slack may seem surprising given that slack is commonly assumed to be part of the FOMC's loss function (while growth in the loss function is somewhat less common), it is consistent with narrative evidence from the FOMC's public communications. Thornton (2011) documents that from 1991 until 2009 the FOMC's policy directive, announced to the public after each FOMC meeting, stated "The Federal Open Market Committee seeks monetary and financial conditions that will foster price stability and promote sustainable *growth in output*". Thornton further notes that neither "maximum sustainable employment nor the unemployment rate" is mentioned in these directives.

[10] the central bank is willing to trade of a positive inflation gap in exchange for higher real activity (SHAPIRO; WILSON, 2019, p.4).

foram baseadas em inferências indiretas derivando as preferências dos banqueiros centrais, a partir dos votos observados nas taxas de juros, ou nas declarações sobre as taxas de juros desejadas, vistas através de uma regra de juros estimada.

Dessa forma, é possível concluir que a ferramenta de *text minning*, mais especificamente o uso dessa como um potencial mecanismo para entendimento do que acontece co cenário macroeconômico torna-se cada vez mais eficaz. A utilização desta para avaliações de expressões dos agentes econômicos, bem como ferramenta potencial para melhor compreensão da atividade econômica por meio de bancos centrais, vem crescendo e sendo incorporada em modelos econométricos para aprimorar os modelos econômicos já existentes.

3 METODOLOGIA

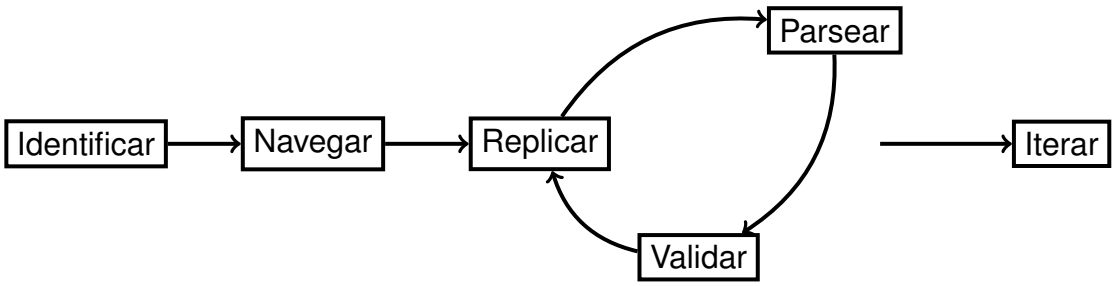
Neste capítulo, apresentaremos as ferramentas que adotamos para a análise de sentimentos dos dados obtidos nos *corpus* textuais; técnicas de *web scraping* e *text mining*; critérios de frequência de distribuições empíricas de probabilidades; e a análise de sentimento. Para esse efeito, dividimos esse capítulo em quatro partes: a primeira contém as ferramentas de *web scraping* e *text mining*; a segunda o critério de divergência de Kullback-Leibler; na terceira, será feito uma descrição de como funciona a análise de sentimento por meio de um dicionário léxico; por fim, apresentamos introdutoriamente o funcionamento de um vetor auto-regressivo e da função resposta ao impulso .

3.1 Coleta de Dados (*Web Scraping*)

Coleta de Dados (ou *web scraping*) pode ser definida como a técnica de coleta de informação (dados) a partir de um ambiente web - rede mundial de computadores - possibilitando uma automatização de obtenção de dados como um todo. Esta técnica tem como sustento as características homogêneas que existem no armazenamento das informações dos computadores. Bem como o fato dos computadores serem máquinas com alta capacidade de processamento de informações. Dizemos, então, que o *web scraping* compreende um conjunto de algorítimos afim de processar informações na rede.

Podemos demonstrar o processo de *web scraping* por meio de um circuito iterativo (Figura 2). Dado o momento em que podemos coletar um dado específico de um site, por exemplo uma classe específica de um ambiente HTML (*Hypertext Markup Language*), é possível a generalização na coleta de dados que acabam por seguir o mesmo padrão – dado uma conexão com o servidor deste site.

Figura 2 – Fluxo de funcionamento de um web scraping



Fonte: Elaboração própria

A possibilidade da reprodução de algorítimos por meio de *web scraping* nos

permite – por vezes – realizarmos coletas volumosas de dados da internet, como por exemplo, dados econômicos estruturados e não-estruturados.

O *web scraping* cada vez mais vem sendo utilizado no mundo empresarial e em políticas públicas. A partir deste, é possível monitorar atividades em tempo real, bem como variações de preços ou acontecimentos políticos de forma automatica.

De forma mais generalizada, exemplos de aplicações do *web scraping* podem ser dados como: 1 - **área comercial e de vendas**, visto que é possível qualificar base de dados de maneira automática bem como adicionar informações às mesmas - como informações de clientes, ou algo que possa interessar; 2 - **monitorar preços de concorrência**, manter listado e atualizado a preços reais variações de preços, bem como lista de vendas de setores específicos; 3 - ***marketing* e investigação de mercado**, investigar compradores, tendências, monitorar marcas. Neste ponto, o *web scraping* pode ser utilizado para - desde a monitoração de fóruns *onlines* até - investigações de tendências em redes sociais (SEMSEO, 2017).

3.1.1 O Fluxo de Funcionamento do *web scraping*

Utilizando termos mais técnicos, o funcionamento do *web scraping* é baseado num fluxo partindo da identificação dos dados a serem coletados; a *navegação* web; um fluxo interativo (replicação, *parseamento* e validação); e finalmente a Iteração dos dados.(CURSO-R, 2018)

O primeiro passo, quando nos referimos a identificação, é definirmos o que queremos coletar. No presente trabalho, o foco se restringe as atas do COPOM (*minutes*), disponíveis em inglês. Definem-se as páginas no site do Banco Central, e a partir delas, busca-se as informações necessárias para a coleta. (COSTA, 2016). Tendo identificado nosso objetivo, foi feita uma seleção das atas necessárias para o presente trabalho. Como o escopo do estudo foram as atas do ano de 2003 ao ano de 2018, por meio de um algorítimo[1] definimos onde deveríamos trabalhar. A replicação dos dados ocorre por meio de um script no software R, no qual realizamos todo o processo anterior, bem como o download das atas. A iteratividade realizada nos permite baixar todas as atas em formato pdf, bem como armazená-las para o estudo posterior.

3.2 A Divergência de Kullback-Liebler

Como partes das ferramentas estatísticas na análise de discrepâncias ou semelhanças entre variáveis aleatórias, foi proposto a divergência de Kullback-Liebler (K-L). É importante salientar que esta função é base na construção do critério de Akaike para

[1] http://selectorgadget.com/

escolhas de modelos - critério muito utilizado na esfera econométrica (WOOLDRIDGE, 2006; GUJARATI; PORTER, 2011).

3.2.1 A Divergência de Kullback-Liebler ou Entropia Relativa

Sejam f e g as funções de densidade de duas variáveis aleatórias unidimensionais contínuas. A informação de K-L, para o caso contínuo, pode ser dada por:

$$I(f,g) = \int f(x) \log\left(\frac{f(x)}{g(x)}\right) \quad dx, \tag{3.1}$$

onde $\log$ representa o logaritmo natural. A notação $I(f,g)$ pode ser estabelecida como " a informação perdida quando g é usado para aproximar f". Dessa forma, $I(f,g)$ é a distância de g para f. Outra interpretação da divergência de K-L é em relação a uma medida de ineficiência: dado a distribuição g quando a distribuição f é verdadeira. Para o caso de distribuições discretas, teremos:

$$I(f,g) = \sum_{i=1}^{k} p(x_i) \cdot \log\left(\frac{p(x_i)}{q(x_i)}\right), \tag{3.2}$$

onde a verdadeira probabilidade do *inésimo* termo é dada por $p(x_i)$ enquanto $q(x_i)$ constitui a distribuição de probabilidades aproximadas.

Caso distribuições iguais, teríamos $I(f,g) = 0 \iff p_i = q_i$. A medida de K-L, contextualizada como uma distância entre dois modelos - isso é, a discrepância entre eles (BURNHAM; ANDERSON, 2002, p. 51)..

3.2.2 Exemplos de K-L

Podemos ilustrar a distância de K-L ($I(f,g)$) simulando distribuições. Seja f uma distribuição Gamma com parâmetros de forma α e escala β iguais a 4; isto é $\alpha = 4$ e $\beta = 4$. Consideraremos, então, 4 distribuições g_i, cada uma com 2 parâmetros: distribuição Weibull; Lognormal; gaussiana inversa; e F de Fisher-Snedecor (Tabela 1) (BURNHAM; ANDERSON, 2002, p. 54).

Tabela 1 – Distribuições para comparação

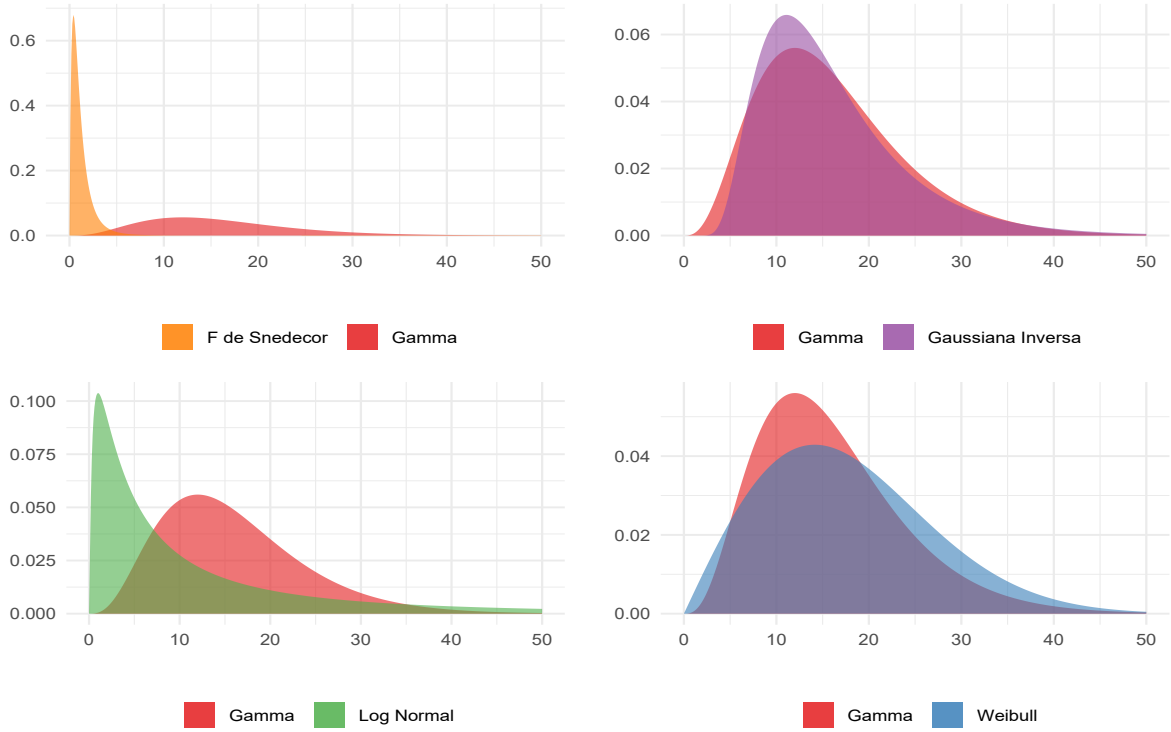
	Modelo Aproximado	$I(f,g_i)$	Ordem
g_1	Distribuição Weibull ($\alpha = 2, \beta = 20$)	0.046189	1
g_2	Distribuição Lognormal ($\theta = 2, \sigma^2 = 2$)	0.672316	2
g_3	Gaussiana Inversa ($\alpha = 16, \beta = 64$)	0.059960	3
g_4	Distribuição F de Fisher-Snedecor ($\alpha = 4, \beta = 10$)	5.745504	4

Fonte: Elaboração própria

De acordo com a Tabela 1, a distribuição simulada que mais se aproxima da Gamma – dado os parâmetros – é a distribuição de Weilbul, seguido pela gaussiana

inversa. Na Figura 3 apresentamos as distribuições de probabilidades comparadas na Tabela 1.

Figura 3 – Gráficos de gamma comparado às distribuições



Fonte: Elaboração própria

Supondo duas distribuições f e g normais $N(\Theta, \tau^2)$ e $N(\mu, \sigma^2)$ respectivamente. Se E_f é a esperança referente a f e a moda de um exemplo mais qualitativo, ou analítico, considere que f é a densidade de uma distribuição normal $\mathbb{N}(\Theta, \tau^2)$.

3.2.3 K-L para Distribuições Normais

Supondo duas distribuições f e g normais $N(\Theta, \tau^2)$ e $N(\mu, \sigma^2)$ respectivamente. Se E_f é a esperança referente a f e a moda de um exemplo mais qualitativo, ou analítico, considere que f é a densidade de uma distribuição normal $\mathbb{N}(\Theta, \tau^2)$.

Seja X uma variável aleatória governada por uma normal uniforme $N(\mu, \sigma^2)$. Isto é $X \sim N(\mu, \sigma^2)$, observamos que se denotarmos a esperança referente a f, E_f temos:

$$\begin{aligned} E_f[(X-\mu)^2] &= E_f[(X-\Theta)^2 + 2(X-\Theta)(\Theta-\mu) + (\Theta-\mu)^2] \\ &= \tau^2 + (\Theta-\mu)^2 \end{aligned}$$

então, para uma distribuição normal $g(x) = (2\pi\sigma^2)^{-\frac{1}{2}} \exp\{-(x-\mu)^2/(2\sigma^2)\}$, temos:

$$\begin{aligned} E_f[\log g(X)] &= E_f\left[\frac{1}{2}\log(2\pi\sigma^2) - \frac{(X-\mu)^2}{2\sigma^2}\right] \\ &= -\frac{1}{2}\log(2\pi\sigma^2) - \frac{\tau^2 + (\Theta-\mu)^2}{2\sigma^2} \end{aligned}$$

e, particular, se considerarmos $\mu = \Theta$ e $\sigma^2 = \tau^2$ nessa expressão, teremos:

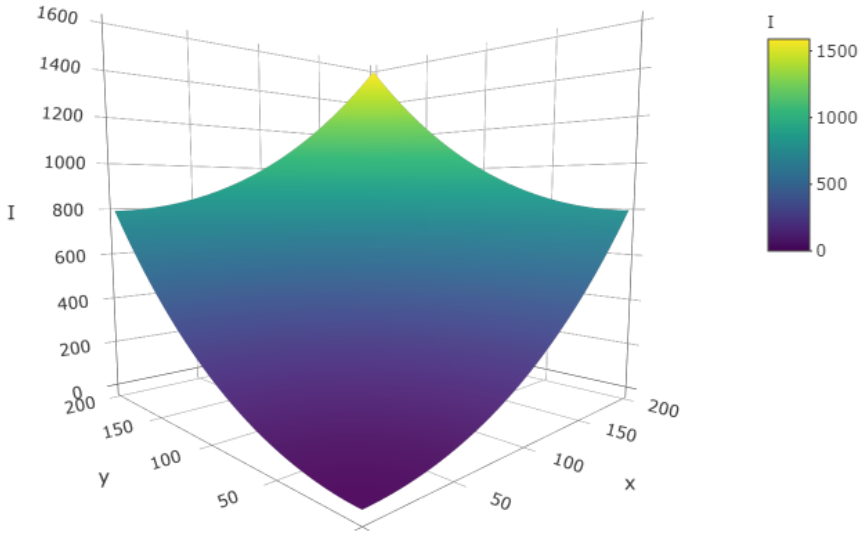
$$E_f[\log f(x)] = -\frac{1}{2}\log(2\pi\tau^2) - \frac{1}{2}$$

assim, a distância de K-L de f em relação a g é dada por (3.1):

$$\begin{aligned} I(f,g) &= E_f[\log g(X)] - E_f[\log f(X)] \\ &= \frac{1}{2}\left\{\log\frac{\sigma^2}{\tau^2} + \frac{\tau^2 + (\Theta-\mu)^2}{\sigma^2} - 1\right\} \quad , \end{aligned}$$

a Figura 4, apresenta a variação de $I(f,g)$ dado que $\sigma^2 = 1$ e $\mu = 0$. y representa τ e x representa Θ, da última equação. Na escala, os valores aproximados para a divergência de K-L.

Figura 4 – Comparações de $I(f,g)$ quando Θ e τ variam, assumindo uma normal padrão ($N(0,1)$)



Fonte: Elaboração própria

3.2.4 K-L para Modelos Normais e de Laplace

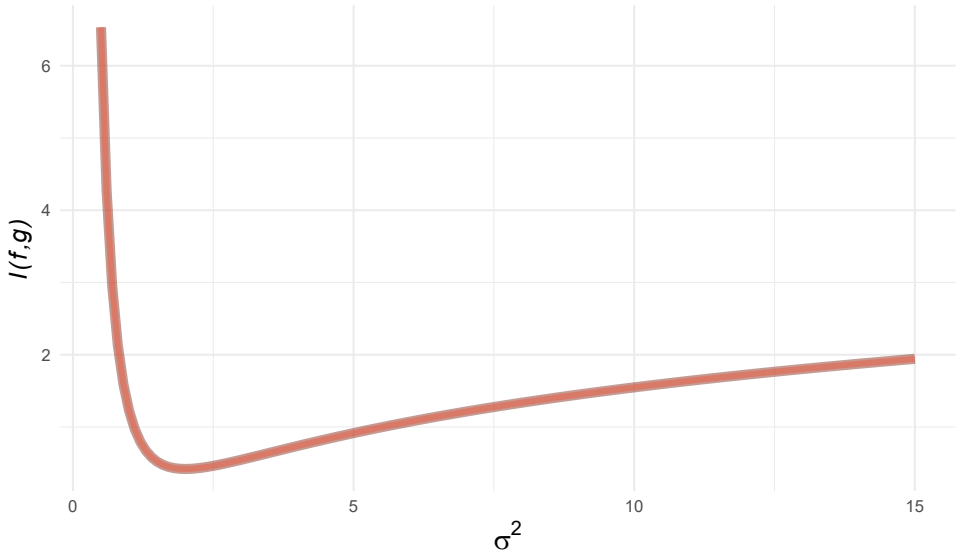
Assumimos desta vez que f segue uma distribuição de Laplace, tal que $f(x) = \frac{1}{2}\exp(-|x|)$ e $f(x) = N(\mu, \sigma^2)$. Neste caso nós teríamos:

$$\begin{aligned} E_f[\log f(X)] &= -\log 2 - \frac{1}{2}\int_{-\infty}^{\infty} |x| e^{-1|x|} dx \\ &= -\log 2 - \frac{1}{2}\int_{-\infty}^{\infty} x e^{-x} dx \\ &= -\log 2 - 1, \\ E_f[\log g(X)] &= -\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{4\sigma^2}\int_{-\infty}^{\infty} (x-\mu)^2 e^{-|x|} dx \\ &= -\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{4\sigma^2}(4 + 2\mu^2). \end{aligned}$$

então, a divergência de K-L é dada por:

$$I(f, g) = \frac{1}{2}\log(2\pi\sigma^2) + \frac{2+\mu^2}{2\sigma^2} - \log 2 - 1.$$

Figura 5 – Valores da divergência de K-L para uma distribuição normal-laplace, dado $\mu = 0$



Fonte: Elaboração própria

Na Figura (5), apresentamos a variação de $I(f, g)$, assumindo $\mu = 0$. Em ambos os casos, podemos variar os parâmetros (nas distribuições normais, fixar σ e permitir a variação de τ e Θ, e em relação ao normal laplace, permitir a variação de σ e μ)

3.3 Índice de *Otimismo*

Analisar o que um texto expressa não necessariamente é algo simples. Automatizar este processo, além disso, pode ser algo bem custoso. A questão fundamental

de uma analise de sentimentos é saber o que uma frase; texto; quer – essencialmente – expressar. Esta técnica nos permite identificar, por exemplo, a polaridade de um texto, se esse se manifesta de forma mais positiva ou negativa. É possível, além disso, contabilizar as palavras mais ditas num *corpus* - ou num conjunto de textos, bem como suas distribuições de frequências.

A partir do momento que possuímos um texto, precisamos tratá-lo. Existem técnicas diferentes para esse tratamento. Neste trabalho, utilizamos dicionários de *stopwords*. Isso é, dado um texto qualquer, removemos **todas as palavras** contidas num dicionário de *stopwords*.

É comum que em qualquer que seja o *corpus*, possuamos palavras que *de forma geral* não contribuiriam de maneira nenhuma para entendermos o que o texto está expressando. De forma geral, dicionários de *stopwords* estão em inglês - bem como os dicionários léxicos - e então somos forçados a trabalhar com textos em inglês[2]. Geralmente, palavras de um dicionário de *stopwords* concentram-se em palavras como "*the*", "*and*", "*these*","*of*" (Tabela 2). Esta tabela contém exemplos de palavras de um dicionário de *stopwords* disponível no pacote **tidytext** (SILGE; ROBINSON, 2016) disponível para a linguagem R. Este dicionário apresenta três dicionários léxicos de *stopwords*, o **onix**; o **SMART** e o **snowball**[3].

Tabela 2 – Exemplos de palavras de um dicionário de *stopwords*

a	across	all	also	and	anyway	are
a's	actually	allow	although	another	anyways	aren't
able	after	allows	always	any	anywhere	around
about	afterwards	almost	am	anybody	apart	as
above	again	alone	among	anyhow	appear	aside
according	against	along	amongst	anyone	appreciate	ask
accordingly	ain't	already	an	anything	appropriate	asking

Fonte: Elaboração própria

Neste trabalho, utilizaremos a análise de sentimentos para montarmos um índice simples de otimismo, configurado da seguinte forma (COSTA, 2016):

$$I_O = \frac{N_P}{N_P + N_N} \quad , \tag{3.3}$$

isso é, em 3.3 representamos que o valor do índice é dado pelo número total de palavras positivas (N_P) dividido pela soma do número total de palavras positivas mais o número total de palavras negativas (N_N). Ainda, de tal forma que $0 \leq I_O \leq 1$.

[2] O COPOM disponibiliza suas atas em inglês (minutes)

[3] http://www.lextek.com/manuals/onix/stopwords1.html, http://www.jmlr.org/papers/volume5/lewis04a/lewis04a.pdf, http://snowball.tartarus.org/algorithms/english/stop.txt

A classificação das palavras, é feita, entretanto, por meio de um dicionário léxico. Dicionários léxicos, na maioria das vezes, são dicionários criados por instituições de pesquisa que visam classificar palavras com valores de forma - ou não - categórica. Dessa maneira, a cada palavra é atribuído um valor ou uma característica.

A Tabela 3 apresenta palavras de um dicionário léxico proposto por Hu e Liu (2004) de polaridade contido no pacote **qdapDictionaries** (RINKER, 2013), parte do *software R*. Neste exemplo, o dicionário visa classificar uma palavra de forma positiva ou negativa, tal que $palavra = -1$ se negativa ou $palavra = 1$ se positiva.

Tabela 3 – Exemplos de palavras de um dicionário léxico

	Palavra	**Valor**	**Palavra**	**Valor**	**Palavra**	**Valor**
1	a plus	1.00	abomination	-1.00	abrade	-1.00
2	abnormal	-1.00	abort	-1.00	abrasive	-1.00
3	abolish	-1.00	aborted	-1.00	abrupt	-1.00
4	abominable	-1.00	aborts	-1.00	abruptly	-1.00
5	abominably	-1.00	abound	1.00	abscond	-1.00
6	abominate	-1.00	abounds	1.00	absence	-1.00

Fonte: Elaboração própria

3.4 Vetor Auto-regressivo

O surgimento dos modelos auto-regressivos são oriundos da década de 80, como alternativa às críticas dos grandes números de restrições impostas às estimações pelos modelos estruturais Banco Central do Brasil (2004).

> "A ideia era desenvolver modelos dinâmicos com o mínimo de restrições, nos quais todas as variáveis econômicas fossem tratadas como endógenas. Sendo assim, os modelos VAR examinam relações lineares entre cada variável e os valores defasados dela própria e de todas as demais variáveis, impondo como restrições à estrutura da economia somente: a escolha do conjunto relevante de variáveis e do número máximo de defasagens envolvidas nas relações entre elas." (Banco Central do Brasil, 2004, p.1)

matematicamente, é possível representar um modelo VAR de primeira ordem (VAR(1)) da seguinte forma:

$$\begin{aligned} y_{1t} &= \delta_1 + \phi_{11} y_{1,t-1} + \phi_{12} y_{2,t-1} + u_{1t} \\ y_{2t} &= \delta_2 + \phi_{21} y_{1,t-1} + \phi_{22} y_{2,t-1} + u_{2t} \end{aligned} \tag{3.4}$$

onde y_1 e y_2 são variáveis endógenas e u_1 e u_2 são os erros para cada equação. ϕ_{12} por sua vez representa a dependência linear de y_{11} em $y_{2,t-1}$ na presença de $y_{1,t-1}$ - isso é, representa o efeito condicional de $y_{2,t-1}$ sobre r_{1t}, dado $y_{1,t-1}$. Dessa forma,

se $\phi_{12} = 0$, então y_{1t} não depende de $y_{2,t-1}$. De forma análoga, se $\phi_{21} = 0$, então a segunda equação mostra que y_{2t} não depende de $y_{1,t-1}$ quando $y_{2,t-1}$ é dado. Ainda, podemos representar o sistema 3.4, da seguinte forma:

$$\begin{bmatrix} y_{1t} \\ y_{2t} \end{bmatrix} = \begin{bmatrix} \delta_1 \\ \delta_2 \end{bmatrix} + \begin{bmatrix} \phi_{11} & \phi_{21} \\ \phi_{12} & \phi_{22} \end{bmatrix} \begin{bmatrix} y_{1,t-1} \\ y_{2,t-1} \end{bmatrix} + \begin{bmatrix} u_{1t} \\ u_{2t} \end{bmatrix}$$

ou, de acordo com as definições apropriadas (VERBEEK, 2008, p.322):

$$\vec{Y_t} = \phi + \Theta\vec{Y}_{t-1} + \vec{\epsilon_t}$$

onde $\vec{Y_t} = [y_{1t}, y_{2t}]'$ e $\vec{\epsilon_t} = [u_{1t}, u_{2t}]'$. Isso estende um modelo auto-regressivo de primeira ordem para um caso de *mais dimensões*. De forma geral, um VAR(p) para um vetor k-dimensional pode ser dado por:

$$\vec{Y_t} = \phi + \Theta_1\vec{Y}_{t-1} + \cdots + \Theta_p\vec{Y}_{t-p} + \vec{\epsilon_t}$$

onde cada Θ_j é uma matriz $k \times k$ e $\vec{\epsilon_t}$ é um vetor de comprimento k de ruídos brancos (*white noises*), com matriz de covariância $\sum$. No caso *univariado*, poderíamos definir a matriz polinomial a partir do operador de defasagem:

$$\Theta(L) = I_k - \Theta_1 L - \cdots - \Theta_p L^p$$

sendo I_k uma matriz identidade k. Logo, poderíamos escrever o VAR como:

$$\Theta(L)\vec{Y_t} = \phi + \vec{\epsilon_t}.$$

a matriz *lag* polinomial é uma matriz $k \times k$, onde cada elemento corresponde a ordem p num polinômio L

O modelo VAR implica um modelo ARMA para cada um de seus componentes. As vantagens em se considerar os componentes de forma simultânea é que o modelo acaba por ser mais parcimonioso, inclui menos defasagens, e possibilita previsões mais precisas, visto que o conjunto de informações é estendido para incluir também o histórico das outras variáveis (VERBEEK, 2008, p.322). Por uma outra perspectiva, Sims (1980) afirma que o uso de modelos VAR ao invés de equações simultâneas estruturais é vantajoso, isso porquê a distinção entre variáveis exógenas e endógenas não tem que ser feita *a priori* e de restrições 'arbitrária' não são necessárias.

3.5 Função de impulso resposta

Da mesma forma que é possível representar um auto-regressivo a partir do seu componente de média móvel, podemos escrever um VAR como um vetor de média móvel (VMA). A representação de um VMA tem funcionalidade essencial na

metodologia de Sims (1980), a qual permite traçarmos as diferentes projeções dado choques nas variáveis contidas no VAR. Considerando a representação matricial de um VAR(1), temos:

$$\begin{bmatrix} y_{1t} \\ y_{2t} \end{bmatrix} = \begin{bmatrix} \delta_1 \\ \delta_2 \end{bmatrix} + \begin{bmatrix} \phi_{11} y_{1,t-1} \\ \phi_{12} y_{1,t-1} \end{bmatrix} + \begin{bmatrix} \phi_{21} y_{2,t-1} \\ \phi_{22} y_{2,t-1} \end{bmatrix} + \begin{bmatrix} u_{1t} \\ u_{2t} \end{bmatrix} \tag{3.5}$$

podemos representar da seguinte forma:

$$\begin{bmatrix} y_{1t} \\ y_{2t} \end{bmatrix} = \begin{bmatrix} \overline{y_1} \\ \overline{y_2} \end{bmatrix} + \sum_{i=0}^{\infty} \begin{bmatrix} \phi_{11} & \phi_{12} \\ \phi_{21} & \phi_{22} \end{bmatrix}^i \begin{bmatrix} u_{1t-i} \\ u_{2t-i} \end{bmatrix} \tag{3.6}$$

a equação (3.6) expressa y_{1t} e y_{2t} em termos de $\{u_{1t}\}$ e $\{u_{1t}\}$ respectivamente. Podemos, entretanto – e ainda – reescrever (3.6) em termos de $\{u_{y_1 t}\}$ e $\{u_{y_2 t}\}$. Assim, o vetor de erros pode ser escrito como (ENDERS, 2008):

$$\begin{bmatrix} u_{1t} \\ u_{2t} \end{bmatrix} = \frac{1}{1 - b_{12} b_{21}} \begin{bmatrix} 1 & -b_{12} \\ -b_{21} & 1 \end{bmatrix} \begin{bmatrix} u_{y_1 t} \\ u_{y_2 t} \end{bmatrix} \tag{3.7}$$

combinando (3.6) e (3.7), temos:

$$\begin{bmatrix} y_{1t} \\ y_{2t} \end{bmatrix} = \begin{bmatrix} \overline{y_1} \\ \overline{y_2} \end{bmatrix} + \frac{1}{1 - b_{12} b_{21}} \sum_{i=0}^{\infty} \begin{bmatrix} \phi_{11} & \phi_{12} \\ \phi_{21} & \phi_{22} \end{bmatrix}^i \begin{bmatrix} 1 & -b_{12} \\ -b_{21} & 1 \end{bmatrix} \begin{bmatrix} u_{y_1 t} \\ u_{y_2 t} \end{bmatrix}$$

ainda, para simplificar a notação, substituiremos a matriz 2×2, tal que:

$$\Gamma_i = \frac{A_1^i}{1 - b_{12} b_{21}} \begin{bmatrix} 1 & -b_{12} \\ -b_{21} & 1 \end{bmatrix}$$

consequentemente, a representação de média móvel apresentada em (3.6) e (3.7) pode ser escrita em termos de $u_{y_1 t}$ e $u_{y_2 t}$:

$$\begin{bmatrix} y_{1t} \\ y_{2t} \end{bmatrix} = \begin{bmatrix} \overline{y_1} \\ \overline{y_2} \end{bmatrix} + \sum_{i=0}^{\infty} \begin{bmatrix} \Gamma_{11}(i) & \Gamma_{12}(i) \\ \Gamma_{21}(i) & \Gamma_{22}(i) \end{bmatrix}^i \begin{bmatrix} u_{y_1 t} \\ u_{y_2 t} \end{bmatrix} \tag{3.8}$$

ou de forma mais compacta:

$$x_t = \epsilon + \sum_{i=0}^{\infty} \Gamma_i u_{t-1} \tag{3.9}$$

A representação apresentada em (3.9) é uma ferramenta extremamente útil para examinar interações entre as variáveis y_1 e y_2. os coeficientes de Γ_i podem ser utilizados para gerar efeitos de $u_{y_1 t}$ em choques de $u_{y_1 t}$ para períodos de tempo de y_t e y_2. Dessa

forma, é possível visualizar que os quatro elementos $\Gamma_{jk}(0)$ são **multiplicadores de impacto** (ENDERS, 2008, p.295).

Os efeitos acumulados de um impulso em u_{y_1t} e/ou u_{y_2t} podem ser obtidos pela soma apropriada dos coeficientes das *funções de resposta impulso*. Por exemplo,podemos perceber que, depois de n períodos o efeito de u_{y_2t} no valor de y_{1+n} é $\Gamma_{12}(n)$. Assim, depois de n períodos a soma acumulada dos efeitos de u_{y_2t} em y_1 é

$$\sum_{i=0}^{n} \Gamma_{12}(i)$$

se y_1 e y_2 são estacionárias, os valores de $\Gamma_{jk}(i)$ converge para zero, quanto maior i for. Assim, *os choques não podem ter efeitos permanentes em séries estacionárias*. Disso segue que:

$$\sum_{i=0}^{\infty} \Gamma_{jk}^2(i) \text{ é finito}$$

os quatro conjuntos de coeficientes $\Gamma_{11}(i)$, $\Gamma_{12}(i)$, $\Gamma_{21}(i)$ e $\Gamma_{22}(i)$ são chamados de **função de resposta impulso**.

4 RESULTADOS OBTIDOS

O objetivo deste capítulo é apresentar os resultados obtidos a partir da metodologia desenvolvida no capítulo 3. É realizado um exercício prático que visa melhor entender as expressões que as reniões do banco central desejam transmitir por meio das atas. Utilizaremos, entretanto, suas versões em inglês (*minutes*), devido ao fato de não haver dicionários léxicos em português que permita um tratamento às atas.

O trabalho feito aqui é referente ao período de 2003-2018. Que por sua vez compete da gestão Meirelles à gestão Ian Goldfajn frente ao BCB – do governo Lula ao governo Temer. Vale salientar, além disto, o *software*/linguagem de programação utilizado: este trabalho foi feito exclusivamente por meio da linguagem R (R Core Team, 2019). Sua escolha foi devido a esta linguagem ser considerada moderna e de alto nível, possibilitando de forma simples uma análise robusta dos dados. Ainda, por ser uma linguagem de programação voltada para estatística, como no próprio manual "feita por estatísticos, para estatísticos", o R possibilita uma vasta gama de opções para análise textual e de dados. Os códigos deste capítulo, bem como deste trabalho estão disponíveis no repositório online da monografia <https://github.com/gustavovital/Monografia>.

4.1 Base de dados

De acordo com o período analisado, de 2003 à 2018, foram disponibilizadas 140 atas do BCB das respectivas reuniões. Dessa forma, totaliza-se 140 reuniões do COPOM. A partir de um algoritmo de *web scraping*, acessamos essas atas e fizemos download, das suas correspondentes em inglês (*minutes*).

O período escolhido é relacionado com as mudanças políticas e econômicas no Brasil, levando em consideração, também, o tamanho amostral, afim de compreendermos melhor o que se passou nesses anos.

Inicialmente, por meio do pacote *rvest* (WICKHAM, 2019) descobrimos as URL's referentes as atas do COPOM - isso é, de todas. Feito isto, obtemos as URL's por meio do pacote "**cronoAno a**"[1], bem como obtemos os links das atas.

Por meio de uma estrutura de repetição simples armazenamos estas atas em listas e realizamos os *downloads* das atas, já do período selecionado, 2003-2018 . Armazenamos estas num diretório.

Utilizando o pacote *tm* (FEINERER; HORNIK; MEYER, 2008), fazemos a leitura

[1] http://selectorgadget.com/

dos PDF's e criamos um *dataframe* que contém os *corpus*[2] das atas. Nosso *dataframe* fica com três colunas (numero da ata, nome do PDF, e *corpus*) e 140 linhas (cada uma referente a uma ata).

Para cada *corpus* contido no documento aplicamos uma técnica de *stop words*, por meio do dicionário *stopwords* (SILGE; ROBINSON, 2016), e adicionamos palavras que julgamos ser desnecessárias, assumindo que tal palavra aparece *pelo menos* 400 vezes no total dos corpus. Dentre as palavras que desconsideramos, podemos citar como exemplo numerais ou nomes próprios sem relevância para uma análise econômica. Além disso, removemos palavras com erro de leitura devido a quebra de linhas (por exemplo `ninflation`)[3].

Figura 6 – Núvem de palavras das palavras que mais aparecem nas atas do COPOM durante todo o período analisado

Fonte: BCB. Elaboração própria

A Figura 6 apresenta em escala as palavras que mais aparecem nas atas do

[2] Efetivamente os textos que iremos trabalhar neste *corpus*, contém basicamente os caracteres das leituras dos PDF's

[3] As palavras com erro de leitura foram desconsideradas nesta parte do processo. Na contagem geral de palavras, passamos a considerar, por essas poderem interferir no resultado final.

COPOM, levando em consideração a técnica aplicada de *stopwords*. Nitidamente, nos atentamos ao fato de **inflation**, **prices**, e **market** tomarem o primeiro plano, quanto a nossa atenção às palavras mais recorrentes.

4.1.1 Estatísticas Descritivas

O primeiro *dataframe* de análise que obtemos diz respeito as frequências das palavras de cunho econômico. Feito isso, é interessante uma contagem das palavras. Na Tabela 4, podemos ver as palavras que mais aparecem nas atas do COPOM no período estudado. Ainda, devemos salientar: essa é a contagem "**líquida**" das palavras - isso é, já foi feita a correção relativa às palavras repetidas ou não contabilizadas[4].

Tabela 4 – Palavras que mais aparecem nas atas do COPOM (2003-2018)

Palavra	n	Palavra	n	Palavra	n	Palavra	n
inflation	7370	committee	1785	decreased	1579	capital	1279
prices	5079	industrial	1785	expansion	1468	ipca	1256
monetary	2882	sales	1778	average	1433	demand	1221
policy	2651	credit	1775	international	1427	sector	1187
market	2562	consumer	1701	domestic	1364	economy	1166
production	2373	activity	1698	previous	1327	observed	1159
growth	2194	expectations	1593	basis	1316	increases	1157

Dito isso, pode-se, também, apresentar a contagem das palavras referente aos períodos específicos analisados. As Tabelas 5, 6, e 7 apresentam essas contagens.

Tabela 5 – Palavras que mais aparecem nas atas do COPOM (período Meirelles)

Palavra	n	Palavra	n	Palavra	n	Palavra	n
inflation	4348	industrial	1329	consumer	1033	activity	926
prices	3212	growth	1272	expansion	983	domestic	926
market	1595	policy	1259	average	977	expectations	907
production	1559	sales	1211	ipca	935	previous	871
monetary	1395	credit	1162	basis	932	capital	864

Tabela 6 – Palavras que mais aparecem nas atas do COPOM (período Tombini)

Palavra	n	Palavra	n	Palavra	n	Palavra	n
inflation	2377	growth	903	activity	607	sector	500
prices	1781	committee	844	credit	605	bcb	492
monetary	995	production	813	sales	567	expansion	477
policy	960	decreased	754	industry	564	observed	475
market	919	consumer	663	international	526	changed	473

[4] Por exemplo, na contagem bruta, teríamos levado em consideração palavras como *nprices, price* e *nprices*

Tabela 7 – Palavras que mais aparecem nas atas do COPOM (período Goldfajn)

Palavra	n	Palavra	n	Palavra	n	Palavra	n
inflation	645	expectations	246	brazilian	156	easing	117
monetary	492	risks	211	baseline	153	projections	117
policy	432	activity	165	balance	126	governor	116
committee	429	department	165	central	119	process	116
economy	306	evolution	165	adjustments	118	risk	116

Trivialmente, percebemos que a palavra que mais aparece nas atas é *inflation*. O controle da inflação frente a meta para a inflação é um dos pontos de interesse do BC, além disso temos que levar em consideração o crescimento da inflação ao final do governo Dilma. *Price* deixa de aparecer como uma palavra importante nas atas da gestão Goldfajn e *policy* não aparece nas atas referentes ao governo Lula.

A partir da contagem total das palavras[5] devemos definir nosso objeto de estudo. Trabalharemos com as 10 palavras de cunho econômico que mais aparecem nas atas do COPOM.

Alguns pontos têm que ser considerados. Primeiramente, esse estudo prevê um período de 16 anos, um total de 140 atas. Entretanto, os períodos de gestão do BCB não foram regulares, e mesmo que fossem, não poderíamos supor que o número de atas por período seria o mesmo, bem como os tamanhos das atas.

4.1.2 Frequência das Principais Palavras

Das 140 atas que utilizamos, no período Lula (gestão Meirelles) temos um total de 76 atas; no período Dilma (gestão Tombini) temos um total de 45 atas; por fim, no período Temer (gestão Goldfajn) temos um total de 19 atas. Ainda, a partir das 10 palavras *de cunho econômico* que mais aparecem nas atas do COPOM, podemos fazer um estudo dirigido para cada período. Inicialmente, vamos considerar os valores *absolutos* das palavras - isso é, a simples contagem de quantas vezes cada palavra aparece, independentemente do tamanho da ata[6].

As palavras que mais aparecem em cada período, como já exposto anteriormente, se encontram nas Tabelas 5, 6, e 7. Podemos, agora, ilustrar, por meio de um histograma, suas distribuições; e um gráfico de linha, para em relação às ocorrências das palavras.

Inicialmente, como feito na seção anterior, faremos isso para todos os períodos (2003-2018)[7]. A distribuição das palavras acaba por não seguir um distribuição iden-

[5] O *dataframe* referente possui mais de 12.000 palavras identificadas, dentre as quais considera-se numeral como um caractere
[6] Ver repositório online
[7] Ver repositório online

tificável visualmente. Nos histogramas, podemos notar isso quando em relação aos *kernels* apresentados. As figuras estão disponíveis no repositório da Monografia.

4.1.3 Frequências Relativas

Como é possível perceber, se fossemos trabalhar com as ocorrências das palavras de forma *absoluta*, teríamos um problema notável: como os tamanhos das atas variam, não poderíamos comparar as ocorrências das palavras de forma direta. Desta forma, como metodologia utilizada, passaremos a considerar as frequências relativas das palavras em relação a cada ata.

Basicamente, a metodologia utilizada foi, dado o número de ocorrências de uma palavra, divide-se este número pelo total de palavras nas atas - já desconsiderado as palavras de dicionários de *stopwords*, tal que:

$$Fp_i = \frac{p_i}{N_i}$$

em que Fp_i é a frequência da palavra p, dada por $\frac{p_i}{N_i}$; ou seja, o número de ocorrências da palavra p, dividido pelo número total de ocorrência de todas as palavras N, na ata i.

Podemos perceber um tendência principal nas palavras *inflation* e *monetary*, possivelmente relacionada com a conjuntura econômica que o país vivia - crise econômica acentuada em 2015. Enquanto, por exemplo, *prices*, *industrial*, *credit*, *growth*, *consumer*, e *market* praticamente deixam de aparecer no período Goldfajn - por vezes nem aparecem, se referem somente ao responsável da sessão, no BC, isso é não há nenhuma interpretação econômica para essas palavras nas atas no período pós-impeachment

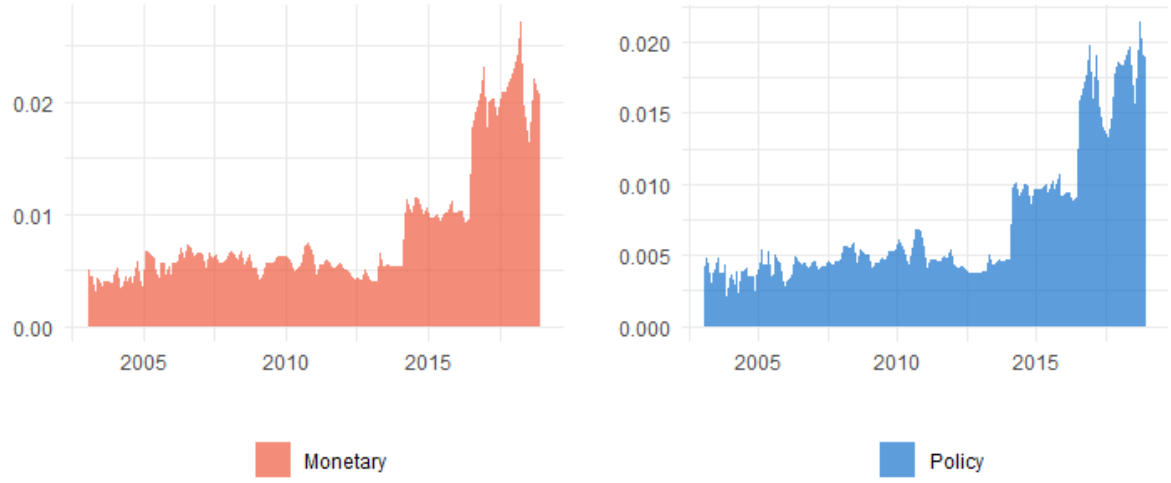
É possível ainda, sugerir uma interpretação econômica para o desaparecimento dessas palavras. Se analisarmos o caso das palavras *monetary* e *policy*, percebemos que a correlação das frequências dessas palavras é de `0.973746`, nos dando um indício de que essas são utilizadas conjuntamente, em *monetary policy* - de tal forma que isso indique que o andamento dessas palavras se refira ao tratamento com a política monetária[8].

A Figura 7 apresenta as frequências relativas às atas das palavras *monetary* e *policy* no decorrer do período analisado. Nitidamente, é possível verificar uma correlação dessas duas palavras.

A Figura 8 apresenta a evolução do ipca acumulado em 12 meses de acordo com os períodos de gestão do BCB, isso é: o aumento em relação a frequência do aparecimento das palavras *policy* e *monetary* pode estar relacionado com uma possível

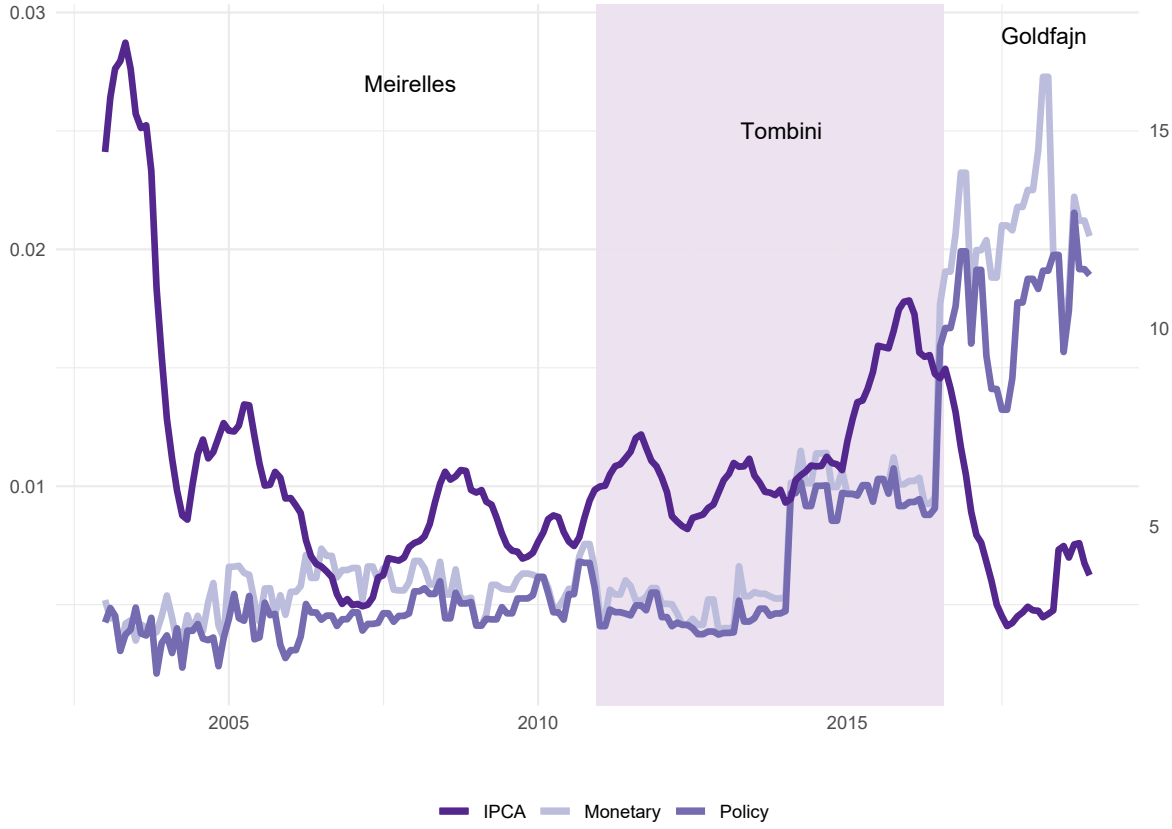
[8] Não utilizamos essa técnica de trabalho neste documento. Se fosse o objetivo trabalhar com conjuntos de palavras, trabalharíamos com *clusters*

Figura 7 – Comparação das frequências de *Monetary* e *Policy*



Fonte: BCB. Elaboração própria

Figura 8 – Comparação das frequências de *Monetary* e *Policy* com o IPCA acumulado em 12 meses



Fonte: BCB. Elaboração própria

ênfase a mudança de política monetária adotada pelo BC. A partir de 2015 há um *boom* na inflação acumulada, sendo o tratamento dessa uma prioridade, faz sentido que se

espere uma mudança em relação à política monetária, se esta não estiver surtindo efeito.

4.2 A divergência de Kullback-Liebler nas frequências das Atas do COPOM

Como primeiro exercício prático deste trabalho, é proposto o cálculo da divergência de K-L. Isso é, verificaremos se as expressões das atas nos períodos analisados possuem similaridades com base nas 148 palavras que mais aparecem no conjunto dos 140 *minutes* do COPOM. Num primeiro momento, a intenção foi o cálculo das 200 palavras que mais aparecem nas atas, entretanto, durante o período Goldfajn algumas palavras deixam de aparecer[9], já como um indício de mudança em relação a abordagem de política monetária e – assim – não seria possível o cálculo da divergência de K-L.

Como o tamanho das atas variam de período a período analisado – bem como de ata para ata – foi necessário trabalharmos com a frequência relativa dessas palavras.

É sabido que a política monetária adotada pelos períodos Meirelles e Tombini foram políticas monetárias sobre vigência do *Partido dos Trabalhadores* (PT); por um outro lado, durante o período Goldfajn, o partido vigente foi o Partido do Movimento Democrático Brasileiro (PMDB). É esperado então que haja uma similaridade maior em relação aos períodos Meirelles-Tombini do que, por exemplo, os períodos Meirelles-Goldfajn. Ainda, podemos assumir uma possível transição de política monetária no que diz respeito ao período Tombini – dessa forma, consideramos as distribuições de palavras, ou expressão das atas, levando em consideração três períodos distintos de política brasileira: Meirelles (Lula), Tombini (Dilma), e Goldfajn (Temer).

4.2.1 Análise para os Períodos de Gestão do Banco Central

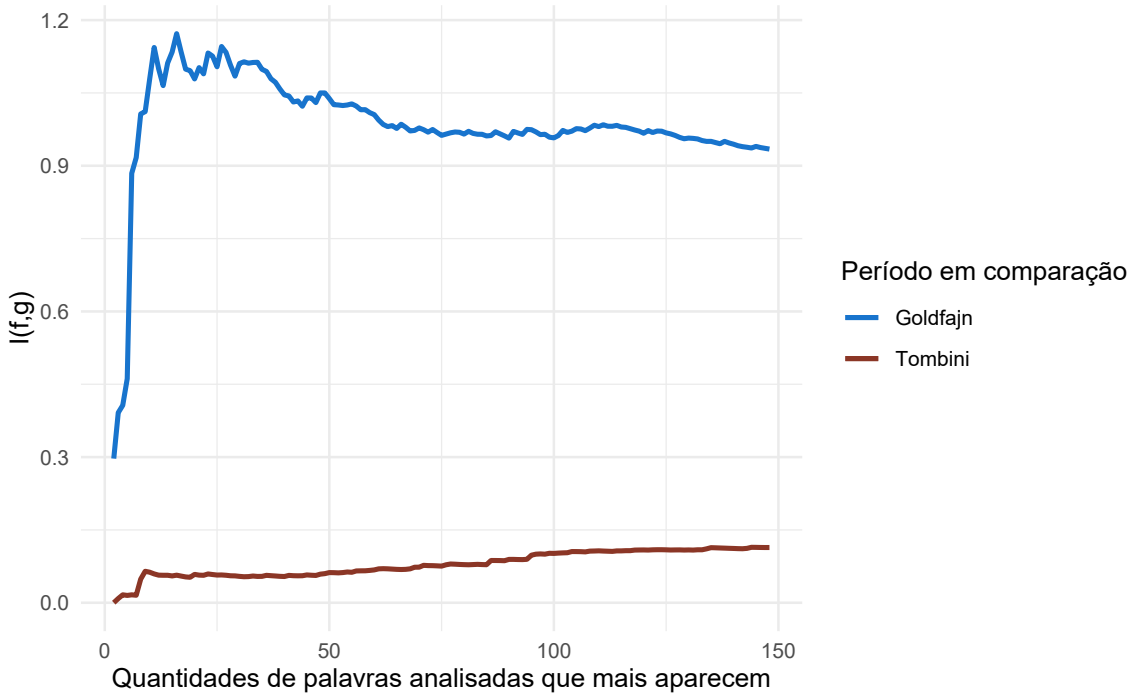
Essa subseção é dividida em três partes. Na primeira faremos uma análise levando em consideração como distribuição das palavras tomando como referência o período Meirelles, na segunda o período Tombini, e por fim o período Goldfajn. Isso é necessário pois a medida de informação de K-L não é uma medida simétrica, logo quando analisada tomando como referência o período Meirelles contra, por exemplo o período Tombini, não teríamos o mesmo resultado quando tomando o período Tombini como referência. Ao final é apresentado uma tabela geral com os valores da divergência de K-L para as 10, 20, 50, 100, e 148 palavras.

[9] As palavras que deixam de aparecer nas atas são: sales, basis, capital, industry, comparison, ibge, exports, durable, manufacturing, expanded, grew, series, construction, totaled, rose, oil, registered, agricultural, imports, products, vehicles, igp, surplus, jobs, recorded, record, output, ipa, fgv, maturing, intermediate, household, smoothed, totaling, calculated, bears, million, trimmed, tradable, liquidity, driven, trailing, br, categories, evaluates, earmarked, equipment, individuals, net, delinquency, commodities, fixed

Período Meirelles

A primeira comparação feita é em relação ao período Meirelles. Tomando a distribuição de palavras desse período como referência, é calculado o valor da divergência de K-L em comparação aos outros dois períodos: Tombini e Goldfajn.

Figura 9 – Divergência de K-L conforme o acréscimo de palavras para o cálculo (período Meirelles)



Fonte: Elaboração própria.

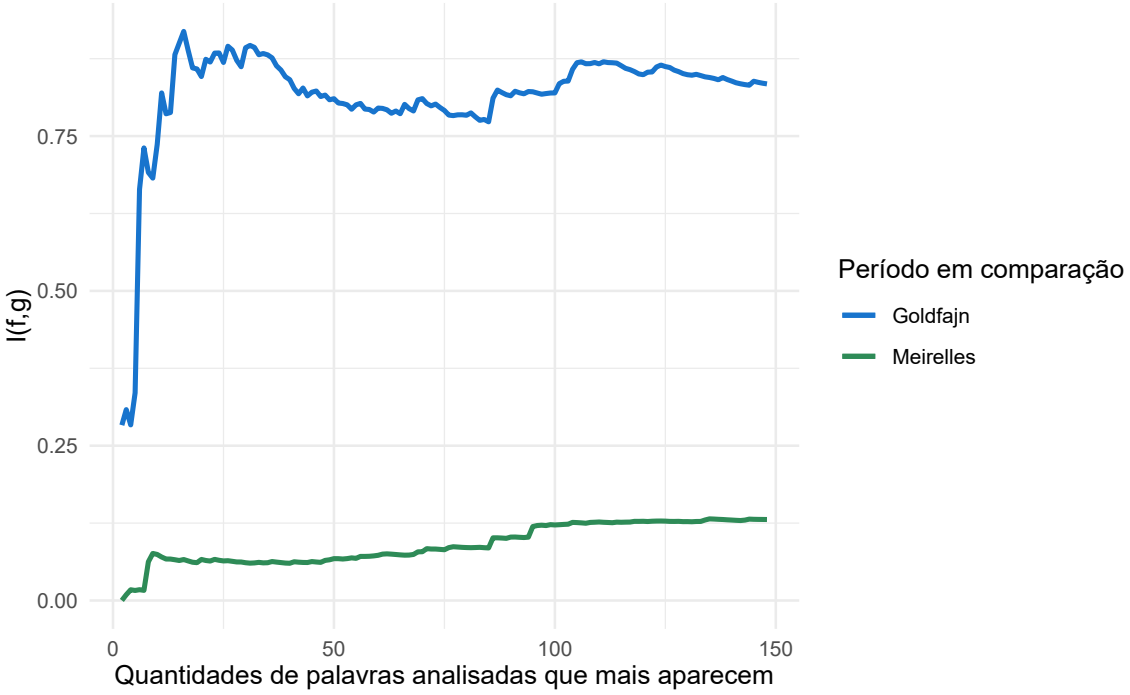
A Figura 9 apresenta a evolução da divergência de K-L conforme o acréscimo das palavras que mais aparecem nas atas. Nitidamente, as distribuições de palavras do período Meirelles se aproxima da distribuições de palavras do período Tombini de forma mais expressiva do que do período Goldfajn, com uma diferença mínima próxima a 0.3 no valor do critério de K-L.

Período Tombini

Tomando a distribuição de palavras do período Tombini como referência, é calculado o valor da divergência de K-L em comparação aos outros dois períodos: Meirelles e Goldfajn.

A Figura 10 apresenta, então, a evolução da divergência de K-L conforme o acréscimo das palavras que mais aparecem nas atas quando tomamos por referência o período Tombini. Nitidamente, as distribuições de palavras desse período analisado se aproxima da distribuições de palavras do período Meirelles de forma mais expressiva –

Figura 10 – Divergência de K-L conforme o acréscimo de palavras para o cálculo (período Tombini)



Fonte: Elaboração própria.

novamente – do que do período Goldfajn, com uma diferença mínima próxima a 0.25 no valor do critério de K-L.

É enfatizado, então, uma semelhança quando as distribuições de palavras estão contidas num período de mesma gestão política vigente (PT); enquanto quando comparado ao período político PMDB as distribuições se tornam mais distantes.

Período Goldfajn

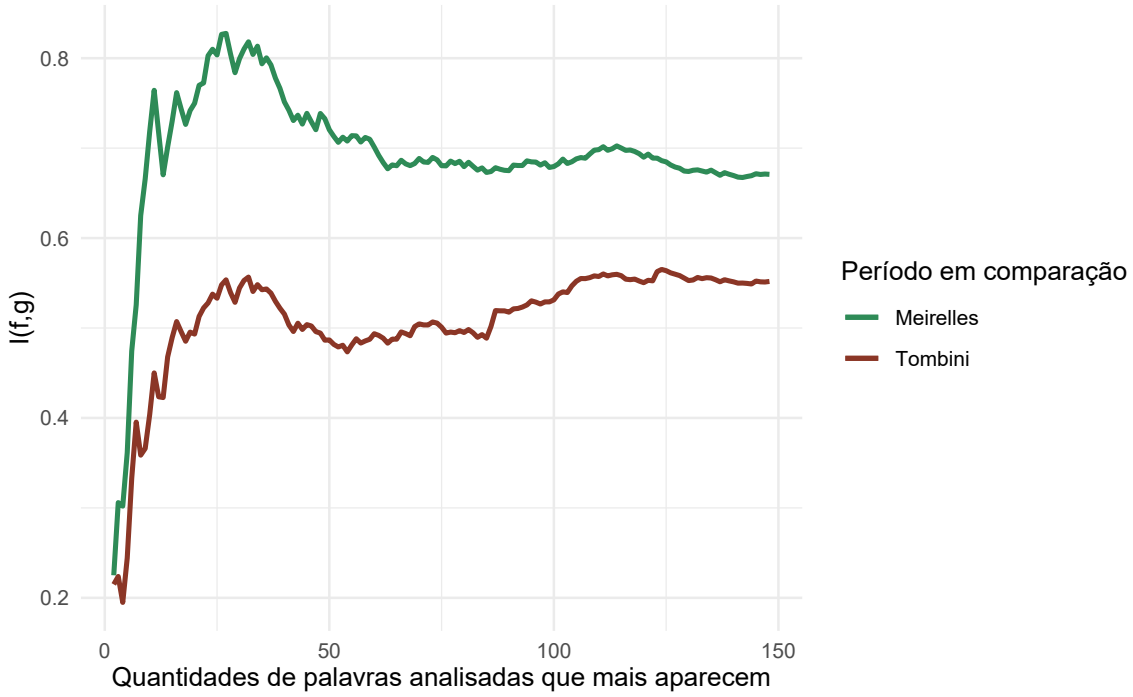
Por fim, analisa-se como período de referência a gestão Goldfajn. Então, é feita uma comparação frente aos outros dois períodos: período Meirelles e período Tombini.

Como já apresentado, é esperado que essa distribuição de palavras seja mais distinta à distribuição de palavras da gestão Meirelles – consequentemente mais próxima a gestão Tombini.

De fato, a Figura 11 demonstra que ainda que as distâncias das distribuições tenham sido reduzidas (quando o período Goldfajn é referência), elas ainda indicam uma proximidade maior ao período Tombini do que ao período Meirelles.

De forma geral, podemos apresentar os resultados das Figuras 9, 10 e 11 como um indicador do que acontece frente as expressões das atas no que diz repeito ao âmbito de política monetária. Isso é, no período Meirelles a distribuição de palavras

Figura 11 – Divergência de K-L conforme o acréscimo de palavras para o cálculo (período Tombini)



Fonte: Elaboração própria.

realmente se assemelha mais ao período Tombini e menos ao período Goldfajn. Seja pela própria política monetária ou mesmo em relação ao âmbito de um outro cenário macroeconômico.

Uma outra abordagem que pode ser levada em consideração é em relação aos termos econômicos. Conforme o número de palavras nas distribuições aumentam a tendência é que menos palavras sejam relacionadas a questão econômica/conjuntural vigente em um determinado período.

Se nos atentarmos as dez palavras mais recorrentes, teremos – em ordem: *inflation, prices, market, production, monetary, industrial, growth, policy, sales,* e *credit*. Entretanto, e obviamente, de **todas** as palavras analisadas, nem todas são referentes às questões econômicas de forma *direta*. Por exemplo, ainda entre as palavras mais recorrentes em todo o período analisado podemos perceber palavras como *previous* e *observed* – que obviamente estão relacionadas a termos econômicos, mas não são termos econômicos em si.

Tabela 8 – Valores de $I(f, g)$ para diferentes números de palavras utilizadas para o cálculo

Período Referente: Meirelles					
	N^o de palavras mais recorrentes utilizadas para o cálculo				
Período comparado	10	20	50	100	148
Tombini	0.063	0.058	0.062	0.1	0.11
Goldfajn	1.1	1.1	1	0.96	0.94
Período Referente: Tombini					
	N^o de palavras mais recorrentes utilizadas para o cálculo				
Período comparado	10	20	50	100	148
Meirelles	0.074	0.066	0.068	0.12	0.13
Goldfajn	0.74	0.85	0.81	0.82	0.84
Período Referente: Goldfajn					
	N^o de palavras mais recorrentes utilizadas para o cálculo				
Período comparado	10	20	50	100	148
Meirelles	0.72	0.75	0.72	0.68	0.67
Tombini	0.4	0.49	0.49	0.53	0.55

Fonte: Elaboração própria

A Tabela 8 apresenta os diferentes valores das distâncias de K-L para diferentes números de palavras utilizados para o cálculo da divergência de K-L. De forma quase genérica, quanto maior o número de palavras utilizadas para o cálculo, maior o valor da divergência.

4.3 Índices de Otimismo e Expressões das Atas

O objetivo, agora, deste trabalho é apresentar um possível índice de otimismo. Como apresentado no capítulo 3, o índice de otimismo consiste *basicamente* em considerar o numero de palavras positivas identificadas nas atas dividido pelo número total de palavras positivas e negativas contidas nestas, como segue na Equação (2.1).

Para criarmos este índice é necessário algumas mudanças na série das atas. Como se sabe, as reuniões do BC não são/foram sempre regulares em relação as datas. Isto é, não temos um série regular de tempo. A metodologia utilizada foi transformar esta série em mensal - normalmente o COPOM se reúne oito vezes por ano, desta forma o que foi feito foi considerar o último valor observado do índice referente à ata da reunião. Para a contagem das palavras, foi utilizado o dicionário do pacote *qdap* do R (RINKER, 2013), e a partir dele realizamos a soma das palavras positivas e negativas em cada ata - por fim, de todas as atas.

Como dito anteriormente, o *score* do índice varia de 0 a 1, dessa forma é possível verificarmos uma possível correlação entre o índice e a situação conjuntural brasileira.

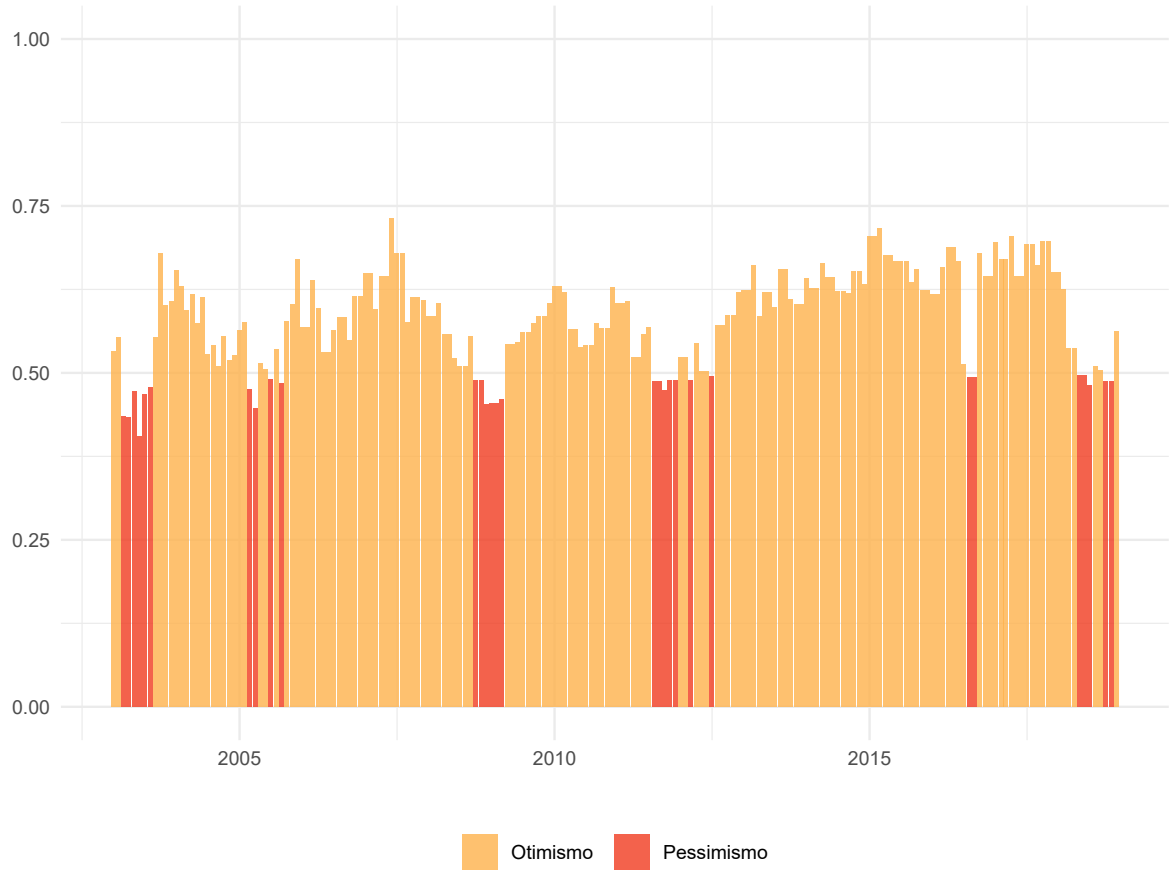
Tabela 9 – Exemplo de *scores* e contagem de palavras positivas e negativas

	Palavras		Score
	Positivas	Negativas	
ATA 80	66	58	0.532
ATA 81	73	59	0.553
ATA 82	53	69	0.434
ATA 83	39	51	0.472

Fonte: Elaboração própria

A Tabela 9 apresenta um exemplo de *score* para as atas 80 à 83. A partir dos *scores* podemos definir momentos de otimismo e pessimismo para as atas dos COPOM. Isso é, dado que o índice varia de 0 a 1, quando o valor do *score* é menor que 0.5, podemos determinar um período de pessimismo na ata; caso contrário, determina-se otimismo.

Figura 12 – Índice de otimismo ao longo do período analisado



Fonte: Elaboração própria

A Figura 12 apresenta a variação do índice proposto. Na barra abaixo da série é apresentado momentos de “otimismo” e “pessimismo” de acordo com o *score*. De forma geral, percebe-se que o índice se mantem na maior parte do tempo de forma

positiva - mesmo em momentos de crise. Argumenta-se que a abordagem do COPOM aos momentos críticos na conjuntura brasileira é dado de forma otimista.

O índice de otimismo proposto pode ser utilizado como uma ferramenta de apoio à analise conjuntural brasileira, de tal forma que seria possível determinar a expressão do BCB para o período atual. Assim, este poderia servir como um indicador - *proxy* - para o tipo de atuação do BCB frente a uma possível situação conjuntural mais crítica.

4.4 Exemplo de Aplicação

Como proposto em Shapiro, Sudhof e Wilson (2018) e Shapiro e Wilson (2019), a partir de um índice de *positividade* ou *negatividade* é possível tentarmos entender o que aconteceria com a economia ou com esse índice, dados choques de algumas variáveis macroeconômicas.

Para entendermos o que poderia acontecer em alguns cenários, iremos nos basear em Shapiro e Wilson (2019) e iremos reproduzir o exercício de função impulso resposta para algumas variáveis macroeconômicas brasileiras. Isso é, assumindo endogeneidade nas variáveis, estima-se um VAR para melhor compreender o que aconteceria se, por exemplo, num cenário de choque positivo no índice, como o Índice de Preços ao Consumidor Amplo (IPCA) acumulado em 12 meses reagiria.

4.4.1 Base de Dados e Período de Estimação

Para a realização deste exercício, utilizaremos o IPCA acumulado em 12 meses, como uma *proxy* da inflação acumulada; e o Índice de Atividade Econômica do Banco Central (IBC-Br) - com ajuste sazonal. respectivamente, representam as séries de número 13522 e 24364, do SGS - Sistema Gerenciador de Séries Temporais (BCB). Além dessas séries, utilizaremos, também, o índice proposto em (3.3).

Uma primeira questão que nos toma, é se as séries são estacionárias. Para nos certificarmos da presença ou não de estacionariedade, utilizaremos o teste de Dickey-Fuller aumentado (ADF). Esse teste tem como hipótese nula (H_0) a presença de raiz unitária; usualmente, sua hipótese alternativa (H_A) é aceita como presença de estacionariedade - condição necessária para a estimação do VAR.

A Tabela 10 apresenta os resultados dos testes de raiz unitária realizados. De acordo com os testes realizados, podemos rejeitar a hipótese de presença de raiz unitária para todas as séries em diferença. Trabalhando em nível, entretanto, só podemos rejeitar esta hipótese para as séries IPCA e índice - mesmo assim, para as últimas duas séries, somente com presença de intercepto e tendência.

É necessário, desta forma, trabalharmos com a série `D(IbcBr)` - a série em diferênça do Índice de Atividade Econômica do Banco Central.

Tabela 10 – Testes de raiz unitárias para as variáveis utilizadas no exercício

Variáveis	ADF (-)	ADF (c)	ADF (ct)	Integração
IbcBr	1.9531	-2.1196	-0.7971	I(1)
IPCA***	-2.7598	-4.3018***	-4.1054***	I(0)
Índice	-0.4107	-3.7810***	-3.9939***	I(0)
D(IbcBr)	-7.4123***	-7.6612***	-7.8751***	I(0)
D(IPCA)	-6.4765***	-6.5740***	-6.6907***	I(0)
D(Índice)	-11.4768***	-11.4444***	-11.4429***	I(0)
Est. de Teste (Tau) (10%)	-1.6200	-2.5700	-3.1300	-
Est. de Teste (Tau) (5%)	-1.9500	-2.8800	-3.4300	-
Est. de Teste (Tau) (1%)	-2.5800	-3.4600	-3.9900	-

Fonte: Elaboração própria.

Nota: * Rejeita-se a hipótese nula a nível de 10%, ** Rejeita-se a hipótese nula a nível de 5%, *** Rejeita-se a hipótese nula a nível de 1%

4.4.2 O Modelo Estimado

A partir do pacote *vars* pode-se estipular a ordem do VAR, dado critérios de informação de akaike, Hannan-Quinn, e Bayesian Information Criterion; respectivamente AIC, HQ, e BIC (PFAFF et al., 2008). Foi verificado que para todos os critérios de informação apontaram para um VAR de segunda ordem. Isso é: foi regredido nas três variáveis utilizadas (D(IbcBr), IPCA, e índice) até a segunda ordem de defasagem de cada uma, de tal forma que o sistema de cada apresente 6 regressores. A equação estimada do VAR(2) para o indice, ficaria, então:

$$indice_t = \Gamma_1 indice_{t-1} + \Gamma_2 indice_{t-2} + \Gamma_3 IPCA_{t-1} + \Gamma_4 IPCA_{t-2} \quad + \\ + \Gamma_5 \Delta IbcBr_{t-1} + \Gamma_6 \Delta IbcBr_{t-6} + const + trend + \epsilon$$

feito isso, e realizadas das estimações, pode-se estimar a função resposta ao impulso.

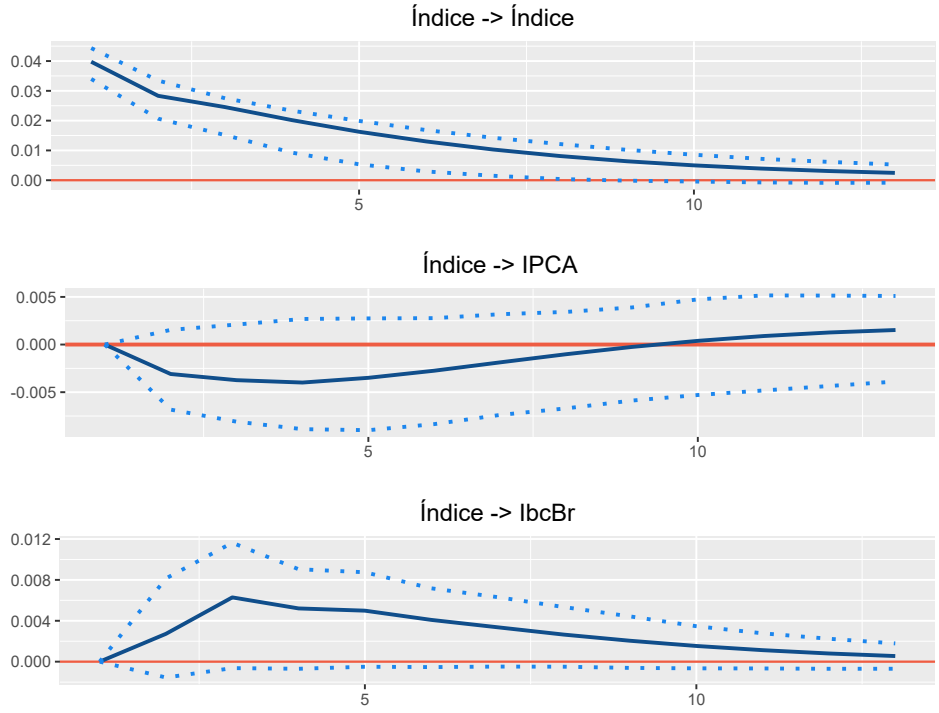
4.4.3 Impulso Resposta ao Índice

Como apresentado em (3.5) em Shapiro e Wilson (2019) e em Jordà (2005), realizamos o procedimento de estimação do VAR(2) e obtivemos suas funções de resposta ao impulso do índice.

A Figura 13 apresenta a projeção para um choque em índice para um período posterior de 12 meses. Isto é, de acordo com a estimação, obtida a partir da *identificação de cholesky* (ver Enders (2008)), dado um choque positivo no índice de otimismo obtido, a tendência de projeção seria que, para as variáveis obtidas:

1. A varição do Índice de Atividade Econômica do Banco Central tenderia a ser positiva por um período de 12 meses a frente. Ainda, em cerca de 15 períodos

Figura 13 – Resposta das variáveis utilizadas a um choque em *índice*



Fonte: Elaboração própria

esta voltaria ao seu estado original (não representado na figura);

2. Em relação ao IPCA, o Índice de Preços ao Consumidor amplo sofreria uma queda inicial, em relação ao acumulado em 12 meses. Posteriormente, essa queda seria convertida em acréscimo, entretanto, com valor inferior a queda inicial. Após cerca de 17 períodos o valor do IPCA acumulado em 12 meses tenderia ao seu estado inicial (não representado na figura).

É válido salientar que há uma vasta opção de variáveis macroeconômicas brasileiras que poderiam ser utilizadas como referência para uma possível correlação com o índice de otimismo. Utilizamos o IPCA e o Ibc-Br como forma de aproximar o exercício da proposta inicial de Shapiro e Wilson (2019), visto que foram variáveis de escolhas próximas. Nada impede, entretanto, de utilizarmos outras variáveis como componentes do VAR - bem como nada impede a utilização de outras técnicas de estimação, ou mesmo testes para verificação de causalidade, por exemplo.

5 CONSIDERAÇÕES FINAIS

Nesta monografia apresentamos as técnicas de *web scraping* conjuntamente com as ferramentas de análise de sentimentos aplicados à métodos quantitativos para melhor compreender as expressões das atas do COPOM.

A partir da análise realizada foi possível encontrar semelhanças e diferenças quanto às expressões das atas em relação a diferentes períodos de gestão de política monetária do Banco Central do Brasil.

Levando em consideração estatísticas descritivas e medidas estatísticas mais robustas, podemos considerar que houve – de fato – uma maior semelhança em relação às políticas monetárias do período PT (Meirelles e Tombini) do que do período PMDB (Goldfajn).

Através de algorítimos e técnicas de *web scraping* e *text mining* foram feitos os *downloads* das atas do COPOM, e o *corpus* foi tratado mediante dicionários e algorítimos de *stopwords*.

Por meio de nuvens de palavras (*wordclouds*) e tabelas descritivas, foram apresentados os termos mais utilizados em cada período, de 2003 à 2018 bem como a variação da divergência de Kullback-Liebler quando comparamos esses principais termos – até finalmente chegarmos a conclusão de que para **todos** os períodos a divergência de K-L foi menor em relação a Meirelles-Goldfajn (qualquer que seja o período referência) do que para qualquer outro período analisado.

Ainda, foi proposto um índice de otimismo, baseado em Costa (2016) para uma melhor compreensão de como as reuniões do Comitê de Política Monetária se comportavam frente às variações do cenário macroeconômico brasileiro. Dessa forma, foi possível apresentar períodos de otimismo e de pessimismo para as reuniões do COPOM de 2003 à 2018, de acordo com os termos mais utilizados nas atas das reuniões periódicas do BCB.

Indo além, e com base na metodologia adotada por Shapiro e Wilson (2019), foi apresentado e estimado um VAR(2) bem como uma função de resposta ao impulso, apresentando uma relação de endogeneidade para o índice proposto, o Índice de Atividade Econômica do Banco Central e o IPCA acumulado em 12 meses.

A função de resposta ao impulso indicou que mediante um choque positivo no índice de otimismo, o Índice de Atividade Econômica do Banco Central sofreria uma variação positiva; bem como dado esse mesmo choque, o IPCA acumulado em 12 meses sofreria uma variação negativa – assim, afirmando um sentido econômico.

Dessa forma, foi possível concluir que quantificar dados textuais pode ser interessante no âmbito de entendermos como o Banco Central do Brasil se expressa, bem como compreender que essas expressões reafirmam uma tendência em relação à política monetária adotada em diferentes períodos – seja quando essas reafirmam posições frente à variáveis macroeconômicas ou mesmo em relação a linha política adotada.

6 REFERÊNCIAS

Banco Central do Brasil. Modelos auto-regressivos. Relatório de inflação, junho 2004. Citado na página 30.

BARRO, R. J.; GORDON, D. B. Rules, discretion and reputation in a model of monetary policy. *Journal of monetary economics*, Elsevier, v. 12, n. 1, p. 101–121, 1983. Citado na página 21.

BARSKY, R. B.; SIMS, E. R. Information, animal spirits, and the meaning of innovations in consumer confidence. *American Economic Review*, v. 102, n. 4, p. 1343–77, 2012. Citado na página 19.

BHOLAT, D. et al. Text mining for central banks. *Available at SSRN 2624811*, 2015. Citado 5 vezes nas páginas 11, 14, 15, 16 e 17.

BLINDER, A. S. Distinguished lecture on economics in government: What central bankers could learn from academics–and vice versa. *Journal of Economic perspectives*, v. 11, n. 2, p. 3–19, 1997. Citado na página 20.

BURNHAM, K. P.; ANDERSON, D. R. A practical information-theoretic approach. *Model selection and multimodel inference, 2nd ed. Springer, New York*, 2002. Citado na página 25.

COSTA, H. C. Big data, machine learning e text mining em economia: Estudos recentes e análise de sentimento do bacen. 2016. Citado 5 vezes nas páginas 17, 18, 24, 29 e 49.

CURSO-R. *O fluxo do web scraping*. 2018. <https://www.curso-r.com/blog/2018-02-18-fluxo-scraping/>. Acesso: 01/09/2019. Citado na página 24.

ENDERS, W. *Applied econometric time series*. [S.l.]: John Wiley & Sons, 2008. Citado 3 vezes nas páginas 32, 33 e 47.

FEINERER, I.; HORNIK, K.; MEYER, D. Text mining infrastructure in r. *Journal of Statistical Software*, v. 25, n. 5, p. 1–54, March 2008. Disponível em: <http://www.jstatsoft.org/v25/i05/>. Citado na página 34.

GUJARATI, D. N.; PORTER, D. C. *Econometria Básica*. [S.l.]: Amgh Editora, 2011. Citado na página 25.

HU, M.; LIU, B. Mining opinion features in customer reviews. In: *AAAI*. [S.l.: s.n.], 2004. v. 4, n. 4, p. 755–760. Citado na página 30.

JORDÀ, Ò. Estimation and inference of impulse responses by local projections. *American economic review*, v. 95, n. 1, p. 161–182, 2005. Citado 2 vezes nas páginas 18 e 47.

KULLBACK, S.; LEIBLER, R. A. On information and sufficiency. *The annals of mathematical statistics*, JSTOR, v. 22, n. 1, p. 79–86, 1951. Citado na página 12.

LOUGHRAN, T.; MCDONALD, B. When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of Finance*, Wiley Online Library, v. 66, n. 1, p. 35–65, 2011. Citado na página 20.

MCLAREN, N.; SHANBHOGUE, R. Using internet search data as economic indicators. *Bank of England Quarterly Bulletin*, n. 2011, p. Q2, 2011. Citado 2 vezes nas páginas 15 e 16.

NYMAN, R. et al. News and narratives in financial systems: exploiting big data for systemic risk assessment. Bank of England Working Paper, 2018. Citado na página 16.

PFAFF, B. et al. Var, svar and svec models: Implementation within r package vars. *Journal of Statistical Software*, v. 27, n. 4, p. 1–32, 2008. Citado na página 47.

R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2019. Disponível em: <http://www.R-project.org/>. Citado na página 34.

RINKER, T. W. *qdapDictionaries: Dictionaries to Accompany the qdap Package*. Buffalo, New York, 2013. 1.0.7. Disponível em: <http://github.com/trinker/qdapDictionaries>. Citado 3 vezes nas páginas 17, 30 e 44.

SEMSEO. *¿Qué utilidad tiene el Web Scraping?* 2017. <https://blog.semseoymas.com/que-utilidad-tiene-el-web-scraping-175.htm>. Acesso: 15/09/2019. Citado na página 24.

SHAPIRO, A. H.; SUDHOF, M.; WILSON, D. Measuring news sentiment. In: FEDERAL RESERVE BANK OF SAN FRANCISCO. [S.l.], 2018. Citado 3 vezes nas páginas 18, 19 e 46.

SHAPIRO, A. H.; WILSON, D. Taking the fed at its word: Direct estimation of central bank objectives using text analytics. In: FEDERAL RESERVE BANK OF SAN FRANCISCO. [S.l.], 2019. Citado 7 vezes nas páginas 19, 20, 21, 46, 47, 48 e 49.

SILGE, J.; ROBINSON, D. tidytext: Text mining and analysis using tidy data principles in r. *JOSS*, The Open Journal, v. 1, n. 3, 2016. Disponível em: <http://dx.doi.org/10.21105/joss.00037>. Citado 2 vezes nas páginas 29 e 35.

SIMS, C. A. Macroeconomics and reality. *Econometrica: journal of the Econometric Society*, JSTOR, p. 1–48, 1980. Citado 2 vezes nas páginas 31 e 32.

THORNTON, D. L. What does the change in the fomc's statement of objectives mean? *Economic Synopses*, Federal Reserve Bank of St. Louis, v. 2011, n. 2011-01-01, 2011. Citado na página 21.

VERBEEK, M. *A guide to modern econometrics*. [S.l.]: John Wiley & Sons, 2008. Citado na página 31.

WALSH, C. E. *Monetary theory and policy*. [S.l.]: MIT press, 2017. Citado 2 vezes nas páginas 19 e 21.

WICKHAM, H. *rvest: Easily Harvest (Scrape) Web Pages*. [S.l.], 2019. R package version 0.3.4. Disponível em: <https://CRAN.R-project.org/package=rvest>. Citado na página 34.

WOOLDRIDGE, J. M. *Introdução à econometria: uma abordagem moderna*. [S.l.]: Pioneira Thomson Learning, 2006. Citado na página 25.