Universidade Federal do ABC
Centro de Engenharia, Modelagem e Ciências Sociais Aplicadas
Trabalho de Graduação em Engenharia de Informação

ARTHUR RIGOLON LANÇA

IMPACTO DA PANDEMIA NA CONJUNTURA ECONÔMICA BRASILEIRA: PLN COMO FERRAMENTA

Santo André

2022

Arthur Rigolon Lança

IMPACTO DA PANDEMIA NA CONJUNTURA ECONÔMICA BRASILEIRA: PLN COMO FERRAMENTA

Trabalho de Graduação apresentado para conclusão da Graduação em Engenharia de Informação, como parte dos requisitos necessários para a obtenção do Título Bacharel em Engenharia de Informação

Orientação: Prof Dra. **MARGARETHE STEINBERGER-ELIAS**

Universidade Federal do ABC

Santo André

2022

RESUMO

A economia de uma sociedade é moldada e impactada por uma série de fatores, como desenvolvimento tecnológico, clima, faixa etária da população, condições geográficas, entre outras. Além disso, acontecimentos esporádicos podem causar impactos intensos e agudos, e que podem trazer consequências extremamente graves e duradouras. Em 2020, o mundo se deparou com um acontecimento mundial que abalou as estruturas sociais, econômicas e humanas de sociedades inteiras e sem distinção: a COVID-19. Nesse sentido, este trabalho buscou identificar e analisar o impacto da pandemia de COVID-19 na economia brasileira por meio de técnicas de Processamento de Linguagem Natural (PLN). A metodologia de pesquisa utilizada pautou-se em levantamento de dados bibliográficos de forma exploratória coletando informações em artigos científicos que versam sobre o tema da análise linguística, bem como relatórios mensais de macroeconomia do Instituto Brasileiro de Economia, da Fundação Getúlio Vargas, publicados ao longo de um período de 2 anos, desde o mês em que foi declarada oficialmente a pandemia no Brasil, Março de 2020. A partir deste conjunto de informações, construiu-se um corpus que foi utilizado para análise estatística e quantitativa buscando apoiar o objetivo de pesquisa. Como resultado do estudo, concluiu-se que o método foi efetivo ao identificar pontos onde o impacto da pandemia teve reflexo mais significativo na linguagem, indicando que o impacto real sofrido na economia pode ser percebido através da análise linguística do corpus.

PALAVRAS-CHAVE: COVID-19; PANDEMIA; CORPUS; PLN

ABSTRACT

The economy of a society is shaped and impacted by a series of factors, such as technological development, climate, age of the population, geographic conditions, among others. In addition, sporadic events can cause intense and acute effects, which can have extremely serious and lasting consequences. In 2020, the world faced a global event that has shaken the social, economic and human structures of entire societies and without distinction: COVID-19. In this sense, this work sought to investigate and analyze the impact of the COVID-19 pandemic on the Brazilian economy through Natural Language Processing (NLP) techniques. The research methodology used was based on an exploratory survey of bibliographic data, collecting information in scientific articles that deal with the linguistic analysis subject as well as a macroeconomics reports by the Brazilian Institute of Economics, of the Getúlio Vargas Foundation, published over a period of 2 years, since the month in which the pandemic was officially declared in Brazil, March 2020. From this data, a corpus was built to be used for statistical and quantitative analysis seeking to support the research objective. As a result of the study, it was concluded that the method was effective in identifying points where the effect of the pandemic had a more significant reflex on language, indicating that the real impact suffered on the economy can be perceived through the linguistic analysis of the corpus.

KEY-WORDS: COVID-19; PANDEMIC; CORPUS; NLP

LISTA DE ILUSTRAÇÕES

LISTA DE ABREVIATURAS E SIGLAS

COVID-19	(co)rona (vi)rus (d)isease
PLN	Processamento de Línguas Naturais
FGV	Fundação Getúlio Vargas
TTR	Type Token Ratio
IBRE FGV	Instituto Brasileiro de Economia
NLTK	Natural Language Toolkit
POS	Parts of Speech

Sumário

1. INTRODUÇÃO

A economia de uma sociedade é moldada e impactada por uma série de fatores, como desenvolvimento tecnológico, clima, faixa etária da população, condições geográficas, entre outras. Além disso, acontecimentos esporádicos podem causar impactos intensos e agudos, e que podem trazer consequências extremamente drásticas e duradouras ao contexto econômico de uma sociedade.

Um exemplo é a forma como a economia dos EUA foi extremamente impulsionada a partir da 1ª Guerra Mundial, período em que o país se tornou o principal fornecedor de alimentos e armamentos para países como França e Inglaterra. Esse papel fez com que os Estados Unidos evoluíssem, rapidamente, de um país com uma economia predominantemente agrícola e doméstica para um grande exportador de alimentos e bens industrializados.

Há muito tempo não se percebia o mundo tão frágil frente a acontecimentos mundiais avassaladores. Em 2020, a economia mundial chegou a níveis preocupantes após a confirmação da pandemia, termo que não era utilizado há décadas e que se refere à distribuição geográfica de uma doença e não propriamente à sua gravidade.

Um exemplo de acontecimento de ocorrência rara e de difícil previsibilidade foi a pandemia do COVID-19 que assolou o mundo nos últimos anos. Só no Brasil, milhões de infectados e mais de 689 mil mortos são retratos da gravidade e do impacto da doença. O que ressalta aos olhos humanos é a velocidade com que a doença atingiu o mundo todo, praticamente ao mesmo tempo. Como medidas sanitárias, decretou-se isolamento social e redução de atividades laborativas e comerciais

Os impactos não se restringem às milhares de vidas perdidas e impactadas de forma direta pela doença. Uma crise dessa magnitude na saúde pública estende seus impactos aos mais variados setores de uma sociedade, sendo o setor econômico um dos principais afetados e que acaba por atingir, em cadeia, a sustentabilidade financeira e social da população.

1.1. Objetivos e Motivação

O presente trabalho de graduação em engenharia tem o objetivo de investigar e analisar o impacto da pandemia de COVID-19 na economia brasileira na linguagem por meio de técnicas de Processamento de Línguas Naturais (PLN). PLN é uma área de estudo que busca analisar, processar e interpretar a linguagem humana através de recursos computacionais.

O trabalho pretende cumprir este objetivo através do estudo do comportamento da linguagem na intersecção entre os domínios econômico e da saúde, identificando e analisando a evolução e comportamento de um léxico relacionado a pandemia dentro de um léxico global tradicionalmente relacionado a assuntos a respeito da conjuntura econômica.

Para tanto, um corpus será construído para servir de objeto de análise e pesquisa do tema e será formado por relatórios mensais de macroeconomia do Instituto Brasileiro de Economia, da Fundação Getúlio Vargas, uma das principais instituições de ensino do país. Serão analisados relatórios publicados ao longo de um período de 2 anos, desde o mês em que foi declarada oficialmente a pandemia no Brasil, Março de 2020. Tais relatórios apresentam, de forma periódica, percepções atualizadas sobre diversos temas econômicos pré-estabelecidos.

A metodologia utilizada no estudo utiliza-se de coleta de dados bibliográficos em bancos de teses e dissertações de universidades brasileiras bem como relatórios de economia disponíveis na Fundação Getúlio Vargas (FGV). O trabalho se propõe a estudar dados quantitativos que se traduzem em qualitativos a partir das análises realizadas com o objetivo de explorar mais sobre o tema que se tornou tão expressivo na vida de todos nos últimos anos.

2. FUNDAMENTAÇÃO TEÓRICA

Para atender ao objetivo principal desse estudo buscou-se o apoio em técnicas de Processamento de Línguas Naturais (PLN). Nesse sentido, para realizar uma análise de PLN consistente e confiável, é necessário possuir um corpus de referência que atenda aos requisitos mínimos de qualidade e metodologia em sua construção. Como a quantidade de corpus em língua portuguesa é escassa, para o estudo proposto neste trabalho será necessário, primeiramente, realizar a construção de um corpus para servir como objeto de pesquisa e análise. Para isso, foram consultadas algumas fontes e referências que explicitam as técnicas, requisitos e metodologia necessários para tal tarefa, desde a escolha dos textos até o processamento e análise dos dados.

Dado o desenvolvimento das pesquisas que utilizam corpus no cenário brasileiro, surgiu a necessidade de se padronizar e sistematizar os procedimentos de tratamento e construção de um corpus a fim de garantir sua qualidade e padronização para que possa ser utilizado de forma efetiva nas pesquisas linguísticas. Dessa forma, o artigo de Aluísio e Almeida (2006) tem a intenção de introduzir a Linguística de Corpus, numa abordagem que utiliza ferramentas computacionais para realizar o tratamento de corpus especificamente aplicada ao português brasileiro.

As autoras também mostram alguns conceitos de corpus, tanto para a Linguística quanto para a Linguística de Corpus. Um dos pontos essenciais no qual o conceito de corpus para a Linguística de Corpus diferencia da Linguística é que o formato dos textos deve ser eletrônico, ou seja, um conjunto de textos físicos, em livros ou papel, não é considerado um corpus. Isso se deve justamente pelo fato de que esta abordagem está baseada no processamento dos dados através de recursos computacionais, para que se possa obter o devido tratamento e adequação do corpus.

Para McEnery e Wilson (1996 apud Aluísio e Almeida, 2006), dois estudiosos da Linguística de Corpus, um corpus deve conter, no mínimo, as seguintes características: uma amostragem suficiente da língua a ser analisada para se obter o máximo de representatividade da mesma, tamanho finito, estar em formato eletrônico e que possa ser usado futuramente como uma referência padrão daquela variedade da língua por outros estudiosos e pesquisadores.

Ainda são ressaltados pelos autores alguns pré-requisitos que devem ser considerados ao realizar o projeto de um corpus, como o fato de que os textos nele contidos devem ser autênticos, ou seja, devem ser escritos de forma natural, e não com o objetivo de serem

utilizados em uma pesquisa, além de serem escritos por falantes nativos ou aprendizes da língua em questão. Outro requisito é que o texto seja representativo da variedade linguística que se deseja estudar, buscando refletir os comportamentos linguísticos presentes em tal variedade. O corpus ainda deve ser balanceado, no sentido de que deve conter textos de tipos e gêneros distintos, evitando o enviesamento da análise devido a concentração em poucos gêneros textuais. Por último, deve ser considerado como requisito o tamanho do corpus, de forma que se adeque ao tipo de pesquisa que será realizada, não sendo mais extenso ou mais enxuto do que o necessário.

Após entender os pré-requisitos necessários para a montagem de um corpus, existem algumas etapas que compõem a metodologia para a sua construção. As etapas são divididas em seleção dos textos, compilação e manipulação do corpus, nomeação de arquivos e geração de cabeçalhos, a proteção da identidade dos participantes do corpus e pedidos de direito de uso dos textos e, por fim, anotação. A seguir, serão descritos de forma breve e resumida cada um desses passos, de forma a organizar e padronizar a construção do corpus. (ALUISIO E ALMEIDA, 2006):

2.1. Seleção dos textos

O primeiro passo na metodologia de construção de um corpus é a seleção dos textos que o compõem. Para essa etapa, deve ser considerada a modalidade do corpus, os gêneros textuais desejados para a pesquisa e para a representação da variedade linguística a ser estudada. Outro ponto a ser levado em conta é o tamanho e a quantidade dos textos selecionados, que deve sempre levar em consideração o objetivo final da pesquisa. Neste ponto, cabe um destaque para a relevância do parâmetro do tamanho quando se trata da escolha de um corpus. Segundo Sardinha (2000), não existem critérios objetivos para a determinação da representatividade de determinado corpus em relação a linguagem que o mesmo busca reproduzir, porém “pode-se tratar da questão em termos relativos. A principal maneira, ou ‘salvaguarda’ pela qual se pode garantir maior representatividade é através do aumento da extensão do corpus. Um corpus maior é em geral mais representativo do que um menor devido ao fato de conter mais instâncias de traços linguísticos raros.” (SINCLAIR, 1991).

2.2. Compilação e manipulação do corpus

Nesta etapa, a compilação do corpus consiste na coleta e armazenamento dos arquivos dos textos selecionados na etapa anterior. A coleta pode ser realizada com o auxílio de uma

ferramenta de busca da internet, como o Google, utilizando-se de palavras-chave e conceitos relevantes para o tema de pesquisa.

Já a manipulação do corpus consiste na conversão dos arquivos dos textos selecionados para um formato adequado para o processamento de dados textuais, como o *.txt*, além da limpeza e formatação dos textos, excluindo dele todos os elementos que não sejam úteis para a formação do corpus em si, como imagens, gráficos, comentários e qualquer outro elemento gráfico ou textual que não seja relevante para o problema de pesquisa a ser investigado.

2.3. Nomeação de arquivos e geração de cabeçalhos

Esta etapa destaca a necessidade de, após concluir a formatação dos textos, nomear os mesmos de forma a organizar e facilitar a recuperação posterior de cada um deles.

2.4. Proteção da identidade dos participantes de um corpus e pedidos de direito de uso dos textos

Esta etapa se refere ao cumprimento de requisitos legais e regulatórios para utilização dos textos coletados em uma pesquisa. Destaca-se que devem ser observados os direitos autorais e de uso de propriedade intelectual em todos os textos selecionados, de forma que sejam solicitadas todas as permissões necessárias, o que pode levar um tempo considerável, além de depender de inúmeras negociações e discussões com as partes envolvidas.

Outro ponto é a solicitação de consentimento para o uso de informações de indivíduos que devem ter sua privacidade preservada, ou garantir o uso de alguma forma de proteção que garanta o sigilo e privacidade de informações sensíveis àqueles que fazem parte, de alguma forma, dos textos que compõem o corpus.

Essa etapa não é uma etapa técnica referente a construção do corpus, mas é essencial por garantir a ética e a validade da pesquisa a ser realizada, bem como a possibilidade de sua divulgação de forma a não infringir nenhuma legislação e nem expor a privacidade de nenhum indivíduo, organização ou instituição por meio da realização do estudo.

2.5. Anotação

Essa etapa pode ser dividida em anotação estrutural e anotação linguística. Na anotação estrutural, estão englobados os processos de marcação de dados bibliográficos ou de

classificação, como tipo do texto, tamanho do arquivo, formatação e metadados em geral a respeito do texto. Já a anotação linguística consiste na classificação de dados semânticos, sintáticos, morfológicos ou de qualquer outro nível gramatical de análise aplicado ao texto.

Complementando este detalhamento apresentado para as etapas propostas para estruturação do corpus por Aluísio e Almeida (2006), em Leite, Takahata e Steinberger-Elias (2020) são apresentadas as referências utilizadas como fundamentação teórica para guiar a elaboração do corpus, definindo a metodologia de coleta dos textos, tratamento, classificação e processamento dos dados. É importante evidenciar que Hasan (1992, apud Leite, Takahata e Steinberger-Elias (2020)) indica que "para serem adequados, os corpora devem ser afinados com os objetivos da análise", para salientar a importância de se construir o corpus tendo em mente o objetivo da análise que se deseja realizar, e não apenas caraterísticas como o tema, gêneros textuais e linguagem utilizada.

Os autores ainda citam os três estágios fundamentais para a concepção de um corpus, definidos por Aluísio e Almeida (2006). O primeiro desses estágios é o projeto do corpus e seleção dos textos que são relevantes para o objetivo da pesquisa. O segundo estágio seria definido pela compilação de tais textos, enquanto o terceiro seria a nomeação dos textos coletados e compilados, com o objetivo de manter uma organização padronizada possibilitando assim a consulta e recuperação dos dados de forma simples.

Além de todo o processo de criação e estruturação do corpus, existem ainda alguns parâmetros e medidas que são extremamente importantes e úteis em uma análise inicial das características qualitativas do mesmo. Johansson (2008) discorre a respeito de dois desses parâmetros: Diversidade Lexical e Densidade Lexical. Ambas são medidas muito importantes para se verificar o desenvolvimento lexical do corpus, ou, de maneira mais simples, o quanto o vocabulário utilizado nesse corpus é desenvolvido, em termos de complexidade, variação e maturidade da escrita.

A autora faz uma comparação entre as duas medidas, ressaltando as principais características e evidenciando as diferenças entre elas. A Diversidade Lexical está diretamente relacionada a variedade do vocabulário empregado no texto. Segundo Johansson (2008), "para um texto ser altamente diverso lexicalmente, o orador ou escritor deve usar muitas palavras diferentes, com pouca repetição das palavras já utilizadas". A medida da Diversidade Lexical pode ser obtida pelo cálculo do TTR (Type Token Ratio), que mede a relação entre o número de palavras diferentes empregadas, chamadas de Types, e o número total de palavras utilizadas

no texto, os Tokens. A autora também alerta que essa taxa deve ser utilizada com cautela, dado que o valor de TTR é fortemente influenciado pelo tamanho do texto, sendo menor conforme o tamanho do texto aumenta. Isso ocorre pois o número de Tokens aumenta quase que invariavelmente a uma taxa muito maior do que o número de Types. De qualquer forma, a Diversidade Lexical fornece informações importantes a respeito da variedade e maturidade do vocabulário do corpus.

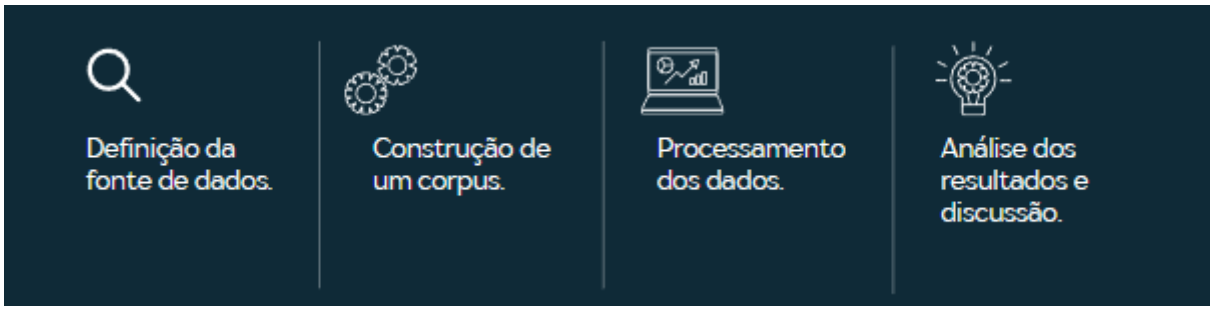
Já a Densidade Lexical tem o objetivo de medir a quantidade de informação contida no texto analisado. Esse parâmetro mede a proporção de palavras de conteúdo, como substantivos, adjetivos, verbos, entre outros, em relação ao total de palavras. As palavras de conteúdo contêm muito mais informação relevante do que as demais palavras, como preposições, conjunções, pronomes e demais, que possuem uma função predominantemente gramatical na construção da linguagem, dessa forma, espera-se que a quantidade de informação também seja maior em um texto com grande proporção de palavras de conteúdo frente ao total de palavras utilizadas.

Vale ressaltar que existe um grau de flexibilidade a respeito da forma de se quantificar a Densidade Lexical de um texto, a depender de fatores como o gênero textual utilizado ou o objetivo da pesquisa, podendo, dessa forma, o autor do estudo definir quais são as classes gramaticais que devem ser consideradas como mais relevantes em termos de conteúdo, sendo a medida de Diversidade Lexical calculada pela proporção de tais palavras em relação ao total de palavras contidas no corpus.

3. MATERIAIS E MÉTODOS

Com a fundamentação teórica apresentada anteriormente norteando a metodologia utilizada neste trabalho, o presente estudo se divide em quatro etapas principais, explicitadas na imagem abaixo.

Figura 1: Principais etapas do estudo



Fonte: Autor (2022)

3.1. Seleção dos Textos para o Corpus

Conforme mencionado na etapa de fundamentação teórica deste trabalho, a etapa de seleção de textos deve ter como norte o objetivo da análise a ser realizada. No caso deste trabalho, o objetivo é, por meio do emprego de técnicas de Processamento de Línguas Naturais, investigar o impacto da pandemia de COVID-19 na economia brasileira. Para tanto, partiu-se do princípio de que os textos a serem utilizados como fontes para a montagem do corpus devem ter como foco o cenário macroeconômico do Brasil, onde seja possível extrair um panorama geral da economia brasileira no momento em que determinado texto foi redigido.

De início, foram levados em consideração duas classes de fontes distintas: a primeira consistia em reportagens que abordassem temas relevantes da economia brasileira, enquanto a segunda englobava relatórios a respeito do cenário macroeconômico elaborados por instituições de alguma forma ligadas a economia e ao mercado financeiro.

A primeira alternativa foi descartada logo na fase inicial. Entre os motivos, pesava a dificuldade de manter a análise mais focada em temas pré-definidos, dado que uma coleta de notícias sobre economia engloba uma gama muito grande de temas abordados, deixando, dessa forma, a pesquisa extremamente ampla e complexa de analisar. Já a segunda alternativa, dos relatórios macroeconômicos, se mostrou mais completa, organizada e controlada, possibilitando estabelecer diretrizes bem definidas e facilmente replicáveis para a coleta.

Geralmente, tais relatórios são publicados com periodicidades muito bem definidas, abordando uma gama mais restrita de temas e, mais importante, adotando uma padronização entre suas diferentes edições ao longo do tempo.

Foram escolhidos para a análise, inicialmente, quatro relatórios periódicos diferentes que abordam o cenário macroeconômico brasileiro. São eles: Relatório Macro Brasil, de autoria do Itaú-Unibanco, Semana em Revista, também do Itaú-Unibanco, Relatório Macro Mensal, do banco BTG Pactual e o Boletim Macro, elaborado pela Fundação Getúlio Vargas. A seguir, uma análise do perfil de cada uma dessas publicações, destacando os principais pontos que justificam ou não a escolha para servirem de fonte para o corpus desta pesquisa.

O relatório Macro Brasil é um relatório semanal publicado pelo Itaú BBA, braço de investimentos do banco Itaú. O banco Itaú é o maior banco comercial do hemisfério sul do mundo. Com aproximadamente 40 milhões de clientes e 608 bilhões de reais em ativos, atua tanto como um banco de varejo, oferecendo serviços financeiros e bancários a clientes pessoa física, quanto como banco de atacado e investimento, atendendo grandes empresas, investidores, governos e demais instituições.

O relatório Macro Brasil é publicado semanalmente no site do Itaú BBA, e tem como característica ser um texto extremamente curto e objetivo, focado em um tema específico, relevante para o cenário macroeconômico brasileiro. São abordados temas como o status da balança comercial brasileira, análise de indicadores relevantes, como inflação, crédito, índices de inadimplência. Cada relatório foca em um tema exclusivo, quase como uma notícia de jornal, porém utilizando de uma linguagem mais especializada e com o forte auxílio de dados quantitativos, gráficos e tabelas.

O relatório é de livre acesso, podendo ser encontrado pelo público em geral. É destinado à parcela da população que possui interesse nos assuntos econômicos que afetam a conjuntura nacional, porém que tem pelo menos um pouco de familiaridade com o assunto, dado que a linguagem não é das mais simples, fazendo uso de siglas e termos específicos e exigindo um pouco de conhecimento prévio para que o leitor compreenda o contexto dos temas abordados.

Este relatório não foi escolhido para ser utilizado como fonte do corpus pois algumas de suas características dificultariam o processo de análise. Entre elas, podemos destacar a curta periodicidade e o fato de cada relatório abordar apenas um assunto específico, diferente a cada semana, o que resulta em uma quantidade de informações mais escassa e na dificuldade de se

manter um padrão de temas a serem analisados. Além disso, a linguagem empregada nos relatórios é de uma complexidade acima da média devido ao uso de termos e jargões específicos, o que em um texto curto como este, causa um impacto ainda maior na leitura.

Assim como o relatório anterior, o Semana em Revista também é um relatório semanal produzido e disponibilizado pelo Itaú. Nesse relatório, ao invés de um tema específico, são abordados uma série de assuntos relevantes relacionados ao cenário macroeconômico brasileiro. Esses assuntos são abordados em formas de tópicos, onde é mostrado o título seguido de um texto curto, de no máximo dois parágrafos sobre o tema, que por vezes é ilustrado com um pequeno gráfico ou tabela.

Nas três páginas do documento são abordados aproximadamente 6 a 8 tópicos, de forma bem resumida e objetiva, que relembram uma notícia de telejornal, no que se refere a quantidade de informação e objetividade com a qual é apresentada. A linguagem é muito próxima da utilizada no relatório anterior, porém com menor uso de siglas e expressões específicas, se destinando ao público que tem interesse nos assuntos da conjuntura econômica brasileira e que quer se informar de maneira rápida, sem grandes aprofundamentos.

O relatório Semana em Revista também não foi escolhido para ser uma das fontes do corpus. Os principais motivos, neste caso, são a fragmentação muito grande da informação, que aborda tópicos relevantes, porém em textos extremamente curtos e resumidos e a falta de constância nos temas entre as edições, dado que apesar de serem analisados alguns tópicos a cada edição, estes tópicos são diferentes de uma edição para outra, dificultando uma análise da evolução de cada um dos temas.

Já o relatório Macro Mensal, elaborado pelo banco BTG Pactual, um dos maiores bancos de investimento do Brasil, aborda temas relevantes para o cenário econômico, mostrando o panorama atual, porém reforçando com projeções e previsões futuras. É dividido em tópicos, e cada um é explorado por aproximadamente uma página. A linguagem utilizada é mais técnica, e o uso de dados em tabelas e gráficos é amplamente explorado.

Apesar de já ser um relatório mais robusto, com uma quantidade relevante de informações a serem extraídas, a linguagem mais especializada e o grande número de recursos não textuais, que dificultariam o processamento, se mostram como obstáculos para a análise proposta neste trabalho. Outro ponto é a forte presença de projeções e previsões, que não são interessantes para objetivo dessa pesquisa, que se baseia na análise do cenário macroeconômico no momento de publicação de cada texto, e não em instantes futuros e hipotéticos. Além disso,

assim como no relatório anterior, também pesa a dificuldade se analisar os temas de forma contínua, dado que os tópicos explorados são diferentes a cada edição. Esses fatores contribuíram para a decisão de também não utilizar o relatório Macro Mensal como fonte do corpus dessa pesquisa.

Por último, foi analisado o Boletim Macro, que é um relatório de periodicidade mensal elaborado pelo Instituto Brasileiro de Economia da FGV (IBRE FGV), abordando os principais temas da economia no último mês. A FGV (Fundação Getúlio Vargas) é uma das mais tradicionais instituições de ensino e pesquisa nos ramos de negócios, gestão, economia e afins do Brasil. Reconhecida internacionalmente pela sua excelência e relevância acadêmica no país. Possui cursos desde o ensino médio até Mestrados, Doutorados e MBAs de formação executiva, atuando também em inúmeras pesquisas e elaborando uma série de jornais e revistas acadêmicas relacionados aos mais diversos temas da sociedade.

O relatório Boletim Macro, elaborado pela instituição, já segue um modelo completamente diferente dos anteriores. Ele é um texto muito mais longo, de 30 a 40 páginas, que aborda uma série de temas relevantes ao cenário macroeconômico brasileiro, porém com um nível maior de aprofundamento em cada um dos temas. O relatório segue quase que o modelo de uma revista, com cada tema sendo redigido por um autor ou grupo de autores, como se fossem matérias da revista. O Boletim Macro se inicia com um texto de introdução, oferecendo um panorama resumido de tudo que está por vir no restante do relatório, inclusive realizando uma breve introdução de cada tópico que será posteriormente explorado no boletim. Após isso, cada tópico é apresentado em um texto de aproximadamente duas páginas, com um maior nível de profundidade e detalhes, onde são apresentados dados e análises dos autores a respeito de cada tema.

A linguagem utilizada pode ser considerada mais simples do que as utilizadas nos relatórios anteriores, o que é esperado dado que o caráter do boletim é predominantemente informativo e educacional. Dessa forma, são menores as ocorrências de expressões específicas e jargões econômicos de difícil compreensão para o público em geral, o que faz com que o texto possa ser compreendido de forma satisfatória pela maioria da população. De qualquer forma, é um boletim destinado àqueles que tem qualquer tipo de interesse no assunto de macroeconomia brasileira, seja por interesses acadêmicos, profissionais ou pessoais.

O resultado da análise das possíveis fontes de dados do estudo pode ser organizado de acordo com a tabela abaixo, permitindo uma visualização mais objetiva das principais características de cada uma, justificando assim, a escolha pela fonte utilizada no trabalho.

Tabela 1: Análise das Possíveis Fontes de Dados

Relatório	Instituição	Periodicidade	Observações
Relatório Macro Brasil	Itaú Unibanco	Semanal	- Texto curto e objetivo - Focado em um único tema, diferente a cada edição - Linguagem especializada
Semana em Revista	Itaú Unibanco	Semanal	- Texto em tópicos - Aborda série de temas econômicos - variam a cada edição
Relatório Macro Mensal	BTG Pactual	Mensal	- Aborda série de temas econômicos - Variam a cada edição - Linguagem muito especializada - Forte presença de previsões e projeções.
Boletim Macro	IBRE FGV	Mensal	- Texto em tópicos maiores - Aborda série de temas econômicos - Fixos - Maior aprofundamento de cada tema - Linguagem menos especializada (jargões)

Fonte: Autor (2022)

Ao final da análise, o Boletim Macro foi a fonte escolhida para o corpus a ser construído neste trabalho. Entre os principais motivos que justificam essa escolha, pode-se destacar, além da linguagem mais acessível e a organização do texto em tópicos tais como inflação, mercado de trabalho, política monetária, etc. que são explorados de forma mais detalhada do que nas fontes anteriores, o fato que de estes tópicos são constantes em todas as edições. Essa característica permite que seja realizada uma análise contínua ao longo do tempo de cada um desses tópicos, que são extremamente relevantes para se compreender o contexto macroeconômico brasileiro.

Definido o Boletim Macro do IBRE FGV como fonte para a construção do corpus, é relevante realizar uma análise mais detalhada a respeito do relatório, seus temas, seu histórico e da instituição que o publica. Como citado anteriormente, o Boletim Macro é publicado mensalmente pelo IBRE, que é o Instituto Brasileiro de Economia da Fundação Getúlio Vargas. O IBRE foi fundado em 1951 com a missão de ser uma referência em pesquisa, análise, produção e disseminação de conteúdos relacionados a macroeconomia, pesquisas econômicas aplicadas e na produção de estatísticas através de indicadores e relatórios. O IBRE também conta com publicações periódicas, como a Revista Conjuntura Econômica, editada pela

instituição desde 1947 e tida como uma das mais influentes revistas econômicas do país, e como o Boletim Macro, fonte de pesquisa deste trabalho. (IBRE FGV, 2022a)

O Boletim Macro é um relatório de análise econômica publicado pelo IBRE desde 2011. Ele é composto por análises, estatísticas e projeções a respeito de uma variedade de tópicos relevantes da economia brasileira. Estes tópicos permitem que o relatório seja dividido em seções fixas, que são abordadas a cada edição, sendo elas: Atividade Econômica, Expectativas de Empresários e Consumidores, Mercado de Trabalho, Inflação, Política Monetária, Política Fiscal, Setor Externo e Panorama Internacional, além do Observatório Político, presente em uma a cada duas edições. Complementando essas seções, o relatório ainda é composto de uma introdução, que aborda de forma resumida os assuntos a serem tratados em cada seção do relatório, e de uma última seção, chamada Em Foco, que trata sobre um tema relevante ao cenário econômico no momento, diferente a cada edição. Cada uma das seções supramencionadas é redigida por um autor ou uma equipe de autores diferentes, que fazem parte do quadro fixo de pesquisadores do IBRE, sendo cada um deles especialista no assunto que lhes cabe tratar no relatório. (IBRE FGV, 2022b)

Após a definição da fonte a ser utilizada, foi necessário definir também o período dos relatórios a serem coletados. Novamente alinhado com o objetivo de se analisar o impacto da pandemia do novo coronavírus na economia brasileira, foi decidido que seriam coletados os relatórios de um período total de dois anos, com início no mês de março de 2020, mês em que a pandemia foi decretada pelas autoridades sanitárias brasileiras. Como o boletim Macro é publicado mensalmente pela FGV, o corpus foi composto de um total de 25 edições do relatório, de março de 2020 a março de 2022.

Figura 2: Capa do Boletim Macro de Abril de 2020

Fonte: FGV IBRE (2020)

3.2. Pré-processamento

Com a fonte do corpus definida, foi iniciada a etapa de pré-processamento, que vai desde a coleta dos textos, sua compilação, consolidação e formatação, até todos os ajustes e análises iniciais necessários para dar início ao processamento do corpus pelas ferramentas de PLN.

Como o objetivo do trabalho é utilizar técnicas de PLN para investigar os impactos econômicos da pandemia de COVID-19, foi definido que as análises iniciais englobariam todos os textos publicados desde o início da pandemia, período já mencionado anteriormente neste estudo. Os Boletins Macro são publicados mensalmente e ficam disponíveis no site do Instituto Brasileiro de Economia, da Fundação Getúlio Vargas, no link. Como mencionado anteriormente, essa linha do tempo de aproximadamente dois anos abrange um total de 25 boletins publicados no período.

Inicialmente, todos esses 25 relatórios foram extraídos do site, consolidados e armazenados na extensão *.pdf*. O nome dos arquivos, quando baixados da internet para o computador, não se apresentava de forma padronizada, assumindo valores aleatórios. Portanto, para melhor organização, identificação e facilidade no tratamento das informações, foi efetuado um trabalho manual para renomear o arquivo de cada boletim de forma padronizada, adotando o formato *ano-mês-boletim-macro*. Dessa forma, foi possível organizar os arquivos seguindo uma ordem temporal, o que possibilitou identificar de maneira muito mais rápida cada arquivo ao longo das etapas seguintes deste trabalho.

Posteriormente, foi necessário realizar a indexação de cada um dos arquivos extraídos, como parte da etapa de anotação estrutural. Para isso, os textos foram organizados em uma planilha de metadados, seguindo o racional exemplificado na tabela 1, abaixo. Como os Boletins Macro são organizados em seções fixas a respeito de temas relevantes da economia, e julgando essencial realizar uma análise apartada de cada um destes temas, cada seção dos relatórios foi indexada como se fosse um texto único, de forma que na planilha de metadados apresentasse uma linha para cada seção. Desta forma, ao final do processo de indexação, a tabela continha os metadados de 262 textos distintos, organizados por data e tema, para potencializar as futuras análises.

Tabela 2 – Descrição dos campos da planilha de Metadados

Coluna	Nome	Conteúdo
A	ID	Número para identificação unívoca de cada seção.
B	Ano	Ano de publicação do Boletim.
C	Mês	Mês de publicação do Boletim.
D	Código	Código para rápida identificação da seção e ano do texto.
E	Seção	Tema fixo abordado pela seção.
F	Título	Título da seção na respectiva edição do boletim.
G	Autores	Nome dos autores da seção.

Fonte: Autor (2022)

Ainda na etapa de anotação estrutural, foi necessário coletar alguns dados quantitativos essenciais em uma análise de PLN: a quantidade de Types e Tokens do corpus. Dado que o corpus construído para este estudo é composto de uma série de boletins mensais, e cada um destes boletins é dividido em seções, nosso corpus total também pode ser dividido em porções, chamadas de subcorpora, compostas pelas seções de cada boletim. Dessa forma, para uma análise completa, ao coletar a quantidade de Types e Tokens deve-se considerar não só a ocorrência de tais parâmetros no texto total, mas também os valores para cada subcorpora. Essa informação será útil para auxiliar a determinar a quantidade e qualidade de informação composta em cada subcorpora, possibilitando o direcionamento do foco das análises posteriores, que serão mais detalhadas e complexas.

Para coletar a informação de Types e Tokens no corpus, foi utilizado o pacote NLTK, na linguagem Python. NLTK é a sigla para Natural Language Toolkit, que é um conjunto de programas e bibliotecas utilizadas para processamento estatístico de linguagem natural. Com o uso do pacote, é possível coletar informações importantes, sendo o número de Types e Tokens uma delas. Porém, para possibilitar o processamento do corpus, foi necessário aplicar um tratamento prévio em questão de formatação e identificação das seções dentro do mesmo. Primeiramente, foi necessário converter e consolidar todos os boletins que compõem o corpus em um único arquivo do formato .txt, dado que o NLTK não realiza o processamento de arquivos no formato .pdf. Além disso, foi necessário realizar uma demarcação no arquivo de forma que a ferramenta fosse capaz de identificar cada seção dentro do corpus, coletando as informações de cada uma individualmente. Para isso, ao final de cada seção, foram acrescentados os caracteres “\n”, como forma de demarcação do fim de uma seção e início de outra.

Com o arquivo em um formato adequado e com as seções devidamente identificadas, foi possível utilizar o NLTK para coletar as quantidades de Types e Tokens em cada uma das seções, e consequentemente, para cada boletim e para o corpus como um todo. Esses dados foram compilados na planilha de metadados onde foram catalogadas todas as seções do corpus, sendo alocados em colunas que continham as informações de Types e Tokens para cada uma das seções, além de uma terceira coluna com a informação do total de stopwords por texto. Stopwords são palavras que não carregam conteúdo semântico relevante para o texto, mas possuem apenas uma função estrutural e gramatical, como conjunções, pronomes, preposições, entre outros.

Tendo coletado tais dados, foi possível ainda adicionar uma quarta coluna, que continha o cálculo do Type Token Ratio (TTR), relação entre o número de Types e Tokens, de cada seção, utilizado como medida da diversidade lexical de cada um dos textos.

Como mencionado anteriormente, o TTR deve ser usado com cautela ao comparar a diversidade lexical de diferentes textos, sempre levando em conta os tamanhos de cada texto, dado que quanto maior o texto, maior o número de tokens, ao passo que o número de Types aumenta em uma proporção consideravelmente menor, fazendo com que, no final, o valor de TTR seja menor.

De acordo com Figueiredo Filho e Silva Júnior, 2009, “o coeficiente de correlação de Pearson (r) é uma medida de associação linear entre variáveis.”. De forma simplificada, ele mede o quanto duas variáveis distintas estão relacionadas e qual é o sentido dessa relação, inverso ou direto. Ainda de acordo com Figueiredo Filho e Silva Júnior, 2009, “O coeficiente de correlação Pearson (r) varia de -1 a 1. O sinal indica direção positiva ou negativa do relacionamento e o valor sugere a força da relação entre as variáveis. Uma correlação perfeita (-1 ou 1) indica que o escore de uma variável pode ser determinado exatamente ao se saber o escore da outra. No outro oposto, uma correlação de valor zero indica que não há relação linear entre as variáveis”.

Levando em conta a relação entre o tamanho do texto e seu TTR, fica clara a necessidade de se utilizar textos de tamanho não muito distintos para conduzir uma análise comparativa a respeito da diversidade lexical. Dessa forma, foi necessário calcular a variância entre os tamanhos de cada subcorpora que compõe o corpus para identificar se há grandes discrepâncias entre os tamanhos dos textos. A variância é um indicador estatístico que analisa o quanto cada valor de uma série de dados varia em relação ao valor médio dessa série. Nesse

sentido, quanto mais alta a variância, mais os valores estão distantes do valor médio do conjunto de dados. A variância calculada para os tamanhos dos textos que compõem o corpus foi de 0,1442, o que significa os valores estão mais próximos ao tamanho médio de todos os subcorpora. Por ser baixo, esse valor demonstra que, em média, os textos possuem tamanhos próximos, portanto o uso do TTR como medida de diversidade lexical é válido, e pode ser levado em conta neste estudo.

O próximo passo da análise do corpus é a coleta e processamento das chamadas POS (Parts of Speech), que são as classes gramaticais nas quais se encaixam os Tokens que formam o texto. Para operacionalizar essa tarefa, foi utilizada a ferramenta CoGroo. Um corretor gramatical utilizado em softwares de texto de plataforma livre, porém que permite, com o emprego de algumas técnicas de programação, processar um corpus e extrair informações consolidadas a respeito das POS de um texto.

Com o auxílio do CoGroo, foram levantados todos os tokens do corpus, ou seja, todas as palavras que compõem o corpus, e cada uma delas foi classificada de acordo com sua classe gramatical. Com os dados extraídos em uma planilha, foi possível analisar e trabalhar de forma quantitativa com os mesmos. Primeiramente foram selecionadas as classes gramaticais que seriam mais relevantes e alinhadas com o intuito deste estudo. Para realizar uma análise a respeito do cenário econômico é necessário analisar palavras que forneçam sentido semântico ao texto, muito mais do que palavras que têm uma função predominantemente gramatical e estrutural. Assim sendo, foram escolhidas as classes de substantivos, adjetivos, verbos e advérbios como classes gramaticais que carregam maior conteúdo.

Foi levantada a frequência de ocorrência de cada uma dessas classes gramaticais para cada seção que compõe o corpus. Esses dados também foram acrescentados a planilha de metadados, considerando uma coluna para cada classe gramatical.

Figura 3 :Recorte da Planilha de Metadados com número de Types, Tokens e POS.

ID	Ano	Mê	código	Seção	Types	Tokens	Stopwords	Div. Lexical	Substantivo	Adjetivo	Advérbio	Verbo
9	2020	3	Mar2020Panorama	Panorama Internacional	400	742	212	0,53908	207	85	48	81
11	2020	4	ABR2020INTRO	Introdução	809	2431	721	0,33278	675	208	161	243
12	2020	4	ABR2020AtivEcon	Atividade Econômica	347	704	203	0,49290	226	74	37	72
13	2020	4	ABR2020Expectativa	Expectativas de Empresários e Consumi	317	606	177	0,52310	172	58	34	60
14	2020	4	ABR2020Mercado	Mercado de Trabalho	264	572	172	0,46154	167	55	31	55
15	2020	4	ABR2020Inflacao	Inflação	243	527	159	0,46110	178	32	23	47
16	2020	4	ABR2020PolMon	Política Monetária	336	626	194	0,53674	162	57	49	74
17	2020	4	ABR2020PolFiscal	Política Fiscal	219	458	131	0,47817	127	27	16	57
18	2020	4	ABR2020SetorExt	Setor Externo	298	769	241	0,38752	240	56	21	70
19	2020	4	ABR2020Panorama	Panorama Internacional	411	905	256	0,45414	262	105	65	86
20	2020	4	ABR2020Observatorio	Observatório Político	593	1374	408	0,43159	317	121	85	142
21	2020	4	ABR2020IBRE	Em Foco IBRE	517	1384	402	0,37355	407	93	73	120
22	2020	5	Mai2020INTRO	Introdução	850	2153	668	0,39480	618	195	119	229
23	2020	5	Mai2020AtivEcon	Atividade Econômica	332	724	206	0,45856	245	60	25	61
24	2020	5	Mai2020Expectativa	Expectativas de Empresários e Consumi	293	581	166	0,50430	177	54	24	48
25	2020	5	Mai2020Mercado	Mercado de Trabalho	305	681	203	0,44787	192	55	32	61
26	2020	5	Mai2020Inflacao	Inflação	207	442	151	0,46833	139	27	23	48
27	2020	5	Mai2020PolMon	Política Monetária	330	596	159	0,55369	128	57	43	97
28	2020	5	Mai2020PolFiscal	Política Fiscal	272	589	169	0,46180	176	53	33	67
29	2020	5	Mai2020SetorExt	Setor Externo	545	1409	429	0,38680	414	161	79	129

Fonte: AUTOR (2022)

Tendo coletado os dados a respeito das classes gramaticais de cada token do corpus, foi possível realizar o processamento dessa informação para calcular o indicador de Densidade Lexical do corpus total e de cada uma das seções que o compõem. Como mencionado no capitulo Fundamentação Teórica, a Densidade lexical é calculada pela relação entre a quantidade de palavras de conteúdo e o total de palavras em um texto. Contudo, como também mencionado previamente, existe certa flexibilidade para se definir a metodologia de cálculo deste indicador, no que tange a escolha do que serão consideradas como palavras de conteúdo, que podem ser definidas pelo autor do estudo a depender de fatores como a finalidade da pesquisa, gêneros textuais utilizados e outras variáveis.

Como no caso deste estudo já foram definidas as classes de substantivos, adjetivos, advérbios e verbos como as classes que carregam maior conteúdo semântico relevante para a análise do cenário econômico, serão as palavras pertencentes a estas classes consideradas, então, palavras de conteúdo. Com essa definição, para obter o valor da Densidade Lexical foi necessário somar a quantidade de palavras pertencentes às classes mencionadas e obter a proporção entre elas e a quantidade total de palavras em cada um dos textos. Esta tarefa foi realizada através do Excel, na própria planilha de metadados, onde já constavam os números de tokens pertencentes a cada uma das classes gramaticais escolhidas.

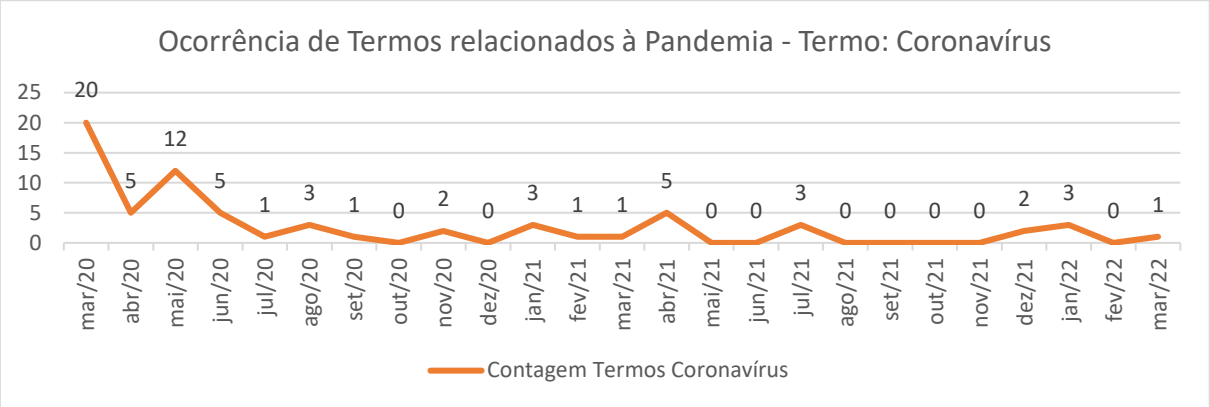
Ainda como parte da etapa de pré-processamento, uma análise inicial dos dados contidos nos textos do corpus se mostrou necessária para duas finalidades. A primeira, certificar que a fonte escolhida continha dados suficientes e relevantes para o estudo proposto neste trabalho, e a segunda, ajudar a determinar quais textos, dentre todos os meses extraídos,

deveriam fazer parte do corpus final, que seria utilizado nas análises da fase de processamento. Em momento posterior, dados sem relevância suficiente foram desconsiderados.

Para essa análise, foram coletadas, de forma manual, as frequências de ocorrência de alguns termos relacionados a pandemia de COVID-19 em cada um dos textos ao longo dos meses. Os termos escolhidos para serem acompanhados foram *COVID-19, Coronavírus, Vírus e Pandemia.* Com o auxílio da ferramenta de busca de texto presente no software leitor de PDF, a ocorrência de cada um desses termos foi contabilizada nos arquivos de cada mês, bem como a soma da ocorrência de todos os termos, que compôs o total de frequência de termos para cada mês. Os resultados desta contabilização foram consolidados e estão ilustrados nos gráficos abaixo, que mostram a frequência de ocorrência de cada termo individualmente e do total dos termos somados ao longo das edições do Boletim Macro.

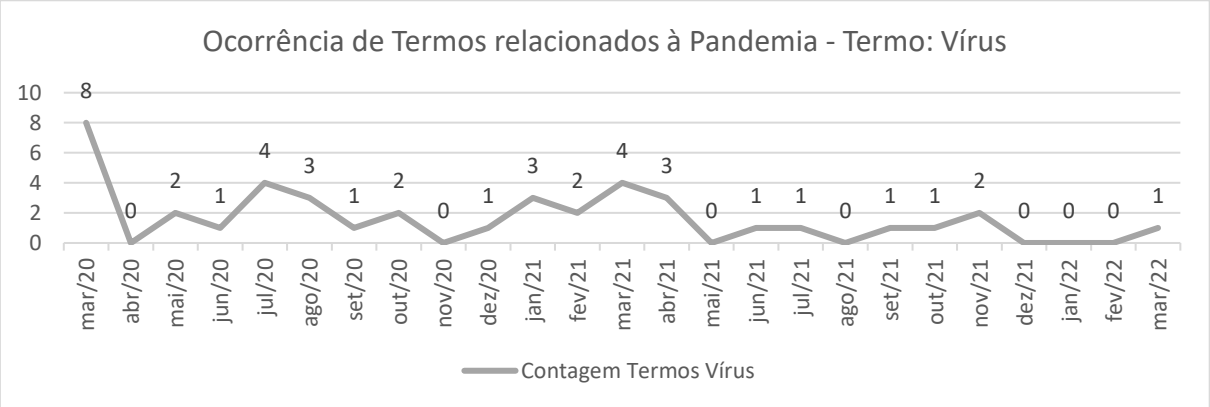
Os gráficos 2 e 3 mostram a ocorrência dos termos Coronavírus e Vírus ao longo do período do corpus. Pode-se notar que os termos foram mais observados no período inicial da pandemia, se tornando menos frequentes no decorrer do período da doença.

Gráfico 2: Frequência de ocorrência do termo *Coronavírus* nos relatórios mensais.



Fonte: AUTOR (2022)

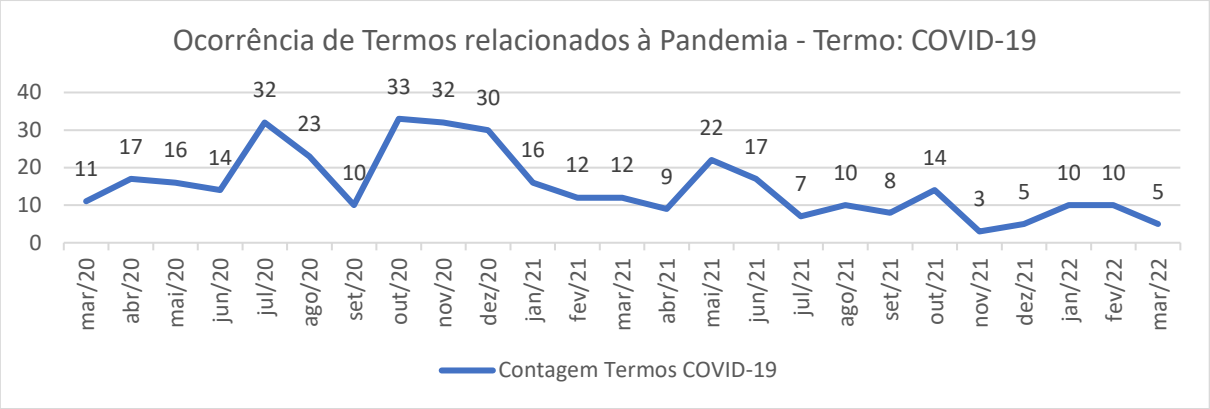
Gráfico 3: Frequência de ocorrência do termo *Vírus* nos relatórios mensais.



Fonte: AUTOR (2022)

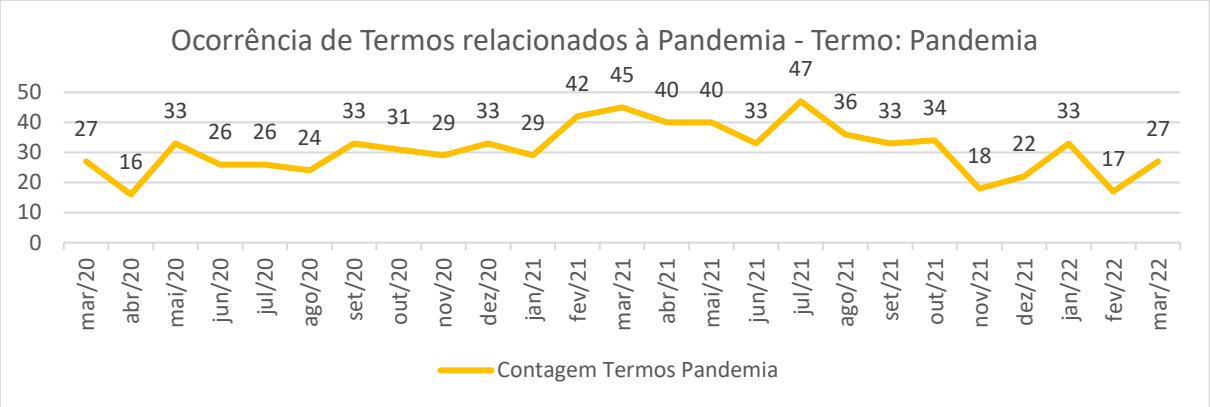
Os gráficos 4 e 5 mostram a ocorrência dos termos COVID-19 e Pandemia ao longo do mesmo período. Para estes, pode-se notar que a frequência de ocorrência foi mais constante, tendo sempre uma frequência relativamente alta e sem grandes desvios ao longo do tempo.

Gráfico 4: Frequência de ocorrência do termo *COVID-19* nos relatórios mensais.



Fonte: AUTOR (2022)

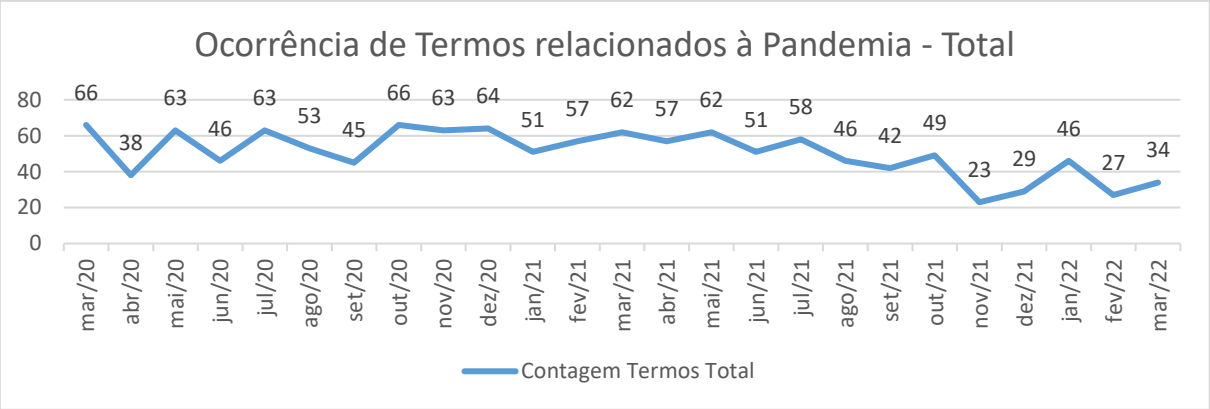
Gráfico 5: Frequência de ocorrência do termo *Pandemia* nos relatórios mensais.



Fonte: AUTOR (2022)

Já o gráfico 6 representa a consolidação da frequência de todos os termos ao longo do tempo. A informação contida nele será a base utilizada em análises posteriores.

Gráfico 6: Frequência de ocorrência total dos termos relacionados à pandemia nos relatórios mensais.



Fonte: AUTOR (2022)

Após a coleta e consolidação destes dados, foi necessário atestar e validar sua relevância para o estudo proposto neste trabalho. Dessa forma, verificou-se que a ocorrência dos termos relacionados a pandemia é um indicador que possui uma relação mínima satisfatória com os reais impactos da pandemia na economia brasileira. Para verificar essa possível relação, a metodologia utilizada foi a de coletar dados a respeito de determinado indicador socioeconômico oficial que refletisse o real impacto causado pela pandemia na economia brasileira e calcular o índice de correlação entre a frequência de termos relacionados à pandemia e o indicador escolhido.

Inicialmente, o primeiro indicador cogitado foi o IPCA – Índice Nacional de Preços ao Consumidor Amplo. Segundo o IBGE – Instituto Brasileiro de Geografia e Estatística, o IPCA (ano) “tem por objetivo medir a inflação de um conjunto de produtos e serviços comercializados no varejo, referentes ao consumo pessoal das famílias, cujo rendimento varia entre 1 e 40 salários mínimos, qualquer que seja a fonte de rendimentos.”, sendo o principal indicador de inflação do Brasil. O indicador, no entanto, não apresentou nenhuma correlação com a ocorrência dos termos relacionados à pandemia ao longo dos textos. Seu uso, portanto, foi descartado devido ao fato de que, para esta análise, é necessária a utilização de um indicador que reflita de forma imediata os impactos sofridos na economia real. O IPCA e demais indicadores de inflação não se aplicam a esta definição, pois a percepção do impacto real nesses indicadores geralmente é de médio prazo, podendo ser ainda postergadas devido a medidas de expansão da base monetária tomadas pelos governos e bancos centrais em momentos de crise, como foi o caso do auxílio-emergencial no Brasil.

Dessa forma, com a intenção de utilizar um indicador que pudesse refletir os impactos econômicos de forma imediata, a Taxa de Desocupação calculada pelo IBGE se mostrou uma escolha muito mais alinhada aos objetivos e necessidades da pesquisa. Este índice é divulgado mensalmente, evidenciando o percentual de desocupação de cada trimestre móvel (média dos últimos três meses). Na visão de Kon (2021):

> O IBGE conceitua as Pessoas Desocupadas como a parcela da População Economicamente Ativa (PEA) que engloba indivíduos sem trabalho na semana de referência, mas que estavam disponíveis para assumir um trabalho nessa semana e que tomaram alguma providência efetiva para conseguir trabalho no período de referência de 30 dias. Dessa forma, a mensuração da Taxa de Desocupação é expressa pelo percentual de pessoas desocupadas na semana de referência em relação à PEA nessa semana.

Tendo definido a Taxa de Desocupação como indicador a ser utilizado como referência para o impacto real da pandemia na economia, foi realizada a extração de sua série histórica de dados do site do IBGE e o armazenamento destes dados numa planilha de Excel, juntamente com os dados de frequência de ocorrência dos termos relacionados à pandemia nos relatórios Boletim Macro, para que posteriormente fosse calculado o coeficiente de correlação entre as mesmas.

Dessa forma, com o auxílio da ferramenta Excel, foi obtido o índice de correlação entre as duas séries de dados com o intuito de medir qual o nível de associação entre elas e, portanto, verificar se a ocorrência de termos relacionados a pandemia nos relatórios possuía relação com o real impacto causado na economia. O coeficiente obtido foi de 0,7076, o que indica uma relação positiva entre as duas variáveis de 70,76%. O resultado foi considerado satisfatório no sentido de demonstrar uma associação suficientemente relevante entre a frequência dos termos relacionados à pandemia e os níveis de desocupação no Brasil, mostrando que os impactos econômicos causados pela pandemia de COVID-19 possuem reflexos que podem ser identificados e mensurados através da análise linguística.

Os procedimentos descritos até aqui fizeram parte da etapa de pré-processamento, que em PLN é representada pelas etapas prévias ao processamento dos dados, que buscam coletar, formatar, mensurar e validar o corpus, de forma a garantir que o mesmo esteja adequado para a finalidade da pesquisa, com equilíbrio e representatividade da língua estudada, além de permitir o processamento pelas ferramentas necessárias. Já na etapa de processamento, o conjunto de dados é manipulado de forma a extrair resultados pertinentes à pesquisa, que

possibilitarão a formação de hipóteses, sua verificação através da análise das informações e obtenção de conclusões a partir dos resultados alcançados.

Dessa forma, a etapa de processamento e análise do conjunto de dados contido no corpus que é base deste estudo teve início a partir da exploração do conceito de Densidade Lexical. Para investigar o impacto da pandemia no cenário macroeconômico brasileiro através dos textos que compõem as edições do Boletim Macro, foi necessário primeiramente identificar em quais desses textos o impacto foi refletido através da linguagem ao longo do tempo.

Como forma de se estabelecer uma métrica que possa acompanhar quantitativamente este impacto ao longo do corpus, o conceito de Densidade Lexical foi explorado de forma mais criteriosa e personalizada. A Densidade Lexical é uma proporção entre palavras que carregam conteúdo relevante e a quantidade de palavras totais do texto. Porém, como mencionado anteriormente, em uma pesquisa, o autor pode definir o que serão consideradas palavras de conteúdo de forma flexível, a depender da finalidade do estudo.

Como no caso deste estudo o objetivo é investigar o impacto da pandemia na economia brasileira, foi definido que dois grupos de palavras seriam consideradas palavras de conteúdo relevante para a finalidade dessa pesquisa. O primeiro grupo, formado por palavras que carregam significado relacionado à pandemia de COVID-19. Já o segundo, composto por palavras pertencentes ao contexto econômico. Tais palavras estarão definidas no contexto deste trabalho como Pandêmicas e Econômicas, respectivamente. O intuito, ao classificar as palavras a partir desta definição, é compor indicadores de Densidade Lexical distintos, que medirão a proporção de palavras pandêmicas e econômicas ao longo dos textos do Boletim Macro. Estes indicadores serão referidos ao longo deste texto como Densidade Pandêmica e Densidade Econômica, ou pelos termos Pandemicidade e Economicidade, respectivamente.

Como próximo passo, foi necessário classificar as palavras do corpus dentre aquelas que possuíam conteúdo pandêmico, conteúdo econômico, e aquelas cujo significado não se encaixa em nenhuma das duas definições. Esse processo foi realizado de forma manual, dado a dificuldade de acesso a uma ferramenta computacional que possa realizar diretamente uma análise tão específica quanto a necessária neste caso. O procedimento foi realizado com a ajuda da ferramenta Excel, onde já estava a lista de tokens levantada anteriormente com a ajuda do software Cogroo. Para diminuir o esforço operacional em analisar cada token individualmente, o primeiro passo foi transformar a lista de tokens em uma lista de types, ou seja, remover todas a palavras duplicadas da lista, fazendo com que cada palavra tivesse que ser classificada apenas

uma vez, e o resultado desta classificação replicado para todas as outras repetições ao longo do corpus. Vale ressaltar que tal procedimento só pôde ser realizado devido ao fato de que se optou por uma classificação das palavras que não leva em conta o contexto em que cada uma se encontra dentro do texto, considerando-se apenas o significado semântico individual de cada palavra.

Tendo retirado as ocorrências duplicadas, a lista de palavras foi reduzida de um total de aproximadamente 300 mil palavras para aproximadamente 16 mil. Com essa nova lista, foi iniciado o processo de classificação manual das palavras. Cada palavra da lista foi analisada, uma a uma, tendo recebido uma classificação dentre 3 tipos possíveis: “pandêmica”, “econômica” ou “nenhum”. Este último, caso a palavra não carregasse, por si só, nenhum significado relacionado a pandemia ou ao cenário econômico. Por ser um processo extremamente manual e, portanto, sujeito a falhas humanas, o mesmo foi repetido mais uma vez pelo autor, e outra pela professora orientadora Margarethe Steinberger-Elias, de forma a validar a classificação feita inicialmente e corrigir possíveis falhas de classificação que possam ter ocorrido no primeiro processo. Dessa forma, as diferenças encontradas entre as classificações realizadas separadamente foram confrontadas, analisadas e discutidas uma a uma entre o autor e a orientadora deste estudo, para se definir a classificação final que deveria ser dada a palavra, e assim compor uma lista definitiva.

Apesar do contexto me que cada palavra se encontra no texto não ter sido considerado para sua classificação, é necessário ressaltar algumas particularidades no tratamento de alguns casos. Primeiramente, algumas palavras que são muito usadas em contextos econômicos, porém que não expressam necessariamente, por si só, um significado relacionado a economia, como “Serviços”, “Trabalhadores” e “Negociação”, foram consideradas com a classificação “nenhum”, dado que não há como inferir qual significado carregam sem analisar o contexto. Já outras palavras, como “Isolamento”, “Distanciamento”, “Aglomeração” e “Casos”, apesar de não terem também um significado necessariamente pandêmico, isoladamente, é possível inferir com certa segurança que a esmagadora maioria das vezes foram utilizadas em contextos da pandemia, dado o espaço temporal no qual os boletins foram analisados. Dessa forma, as mesmas foram consideradas como pandêmicas.

Outros casos relevantes a serem comentados são os de expressões compostas de mais de uma palavra. Para aqueles casos em que a expressão é composta de uma ou mais palavras que, isoladamente, possuem contexto econômico ou pandêmico, apenas estas contavam com a devida classificação, sendo as demais palavras que formam a expressão atribuídas à

classificação “nenhum”. Já para os casos em que a expressão, dentro do contexto do texto, claramente possui um significado econômico, porém, individualmente, nenhuma de suas palavras carrega este significado, nenhuma das palavras que compõem a expressão foram consideradas nem como “pandêmicas” ou como “econômicas”. O mesmo ocorreu para expressões em outras línguas, dado que este estudo se trata da análise de um corpus em língua portuguesa.

Com a classificação finalizada, o resultado de cada palavra foi replicado para todas as suas repetições ao longo da lista inicial de tokens. Dessa forma, os dados estavam no formato necessário para que fosse realizada a análise dos indicadores de Densidade Pandêmica e Econômica. Inicialmente, os mesmos foram calculados para o conjunto total de tokens do corpus, buscando obter uma métrica de referência que representasse a média dos indicadores ao longo do corpus para que posteriormente os indicadores de cada um dos textos pudessem ser observados e analisados de forma relativa à média total. Os valores obtidos para os indicadores de Densidade Pandêmica e Econômica, para o corpus como um todo foram, respectivamente, 0,60% e 3,17%.

Além disso, outro indicador quantitativo que pôde ser calculado a partir da análise dos dados foi a relação entre a quantidade de palavras pandêmicas em relação a quantidade de palavras econômicas no corpus separadamente para cada um dos textos. Naturalmente, por se tratar de um relatório que trata do cenário macroeconômico brasileiro, o índice de palavras econômicas tende a ser relativamente maior do que o número de palavras pandêmicas. Ao calcular essa relação para o corpus total tivemos a confirmação dessa premissa, dado que o indicador mostrou uma relação de 0,1886, ou seja, a incidência de palavras pandêmicas representa, em média no corpus, 18,86% da incidência de palavras econômicas. O acompanhamento dessa métrica é útil para identificar distorções pontuais que sejam maiores do que o normal e que podem indicar uma maior presença de efeitos da pandemia através da linguagem.

Com os parâmetros calculados para o corpus todo e para cada texto individualmente, foi possível plotar os resultados de forma unificada em um gráfico que permitisse a visualização da evolução dos indicadores para cada texto ao longo do corpus. Esse procedimento, porém, não se mostrou muito proveitoso para a análise, dado que a grande quantidade de textos, e consequentemente de dados, prejudicou a visualização dos parâmetros e identificação de pontos de grande variação ou distorção nos mesmos. Como forma de atenuar essa questão, a abordagem adotada foi de analisar os textos separando-os pelo macro-tema tratado em cada

seção. Como explicitado anteriormente ao apresentar o Boletim Macro como fonte do corpus deste trabalho, as edições mensais do relatório são divididas em seções que abordam temas pré-definidos, como Mercado de Trabalho, Inflação, Política Monetária, entre outros. Dessa forma, a estratégia adotada teve a intenção de analisar cada um destes temas ao longo das edições, buscando identificar o impacto da pandemia refletido na linguagem para cada um dos temas ao longo do tempo.

Os dados então foram divididos de acordo com o tema de cada texto e analisados separadamente. Como resultado, os números de tokens pandêmicos e econômicos, os indicadores de Densidade Pandêmica e Econômica e a relação entre tokens Pandêmicos e Econômicos foi organizada em tabelas que consolidam os dados de cada um dos temas. Além disso, foram calculados os valores médios de cada uma dessas variáveis para cada conjunto de textos do mesmo tema, para estabelecer parâmetros de referência para análise dos textos de cada tema individualmente. Abaixo está exemplificada a tabela com os parâmetros para os textos do tema “Panorama Internacional”.

Quadro 1- Parâmetros – Seção: PANORAMA INTERNACIONAL

Seção	Qtd Tokens Pandêmicos	Pandemicidade	Qtd Tokens Econômicos	Economicidade	Rel. Pand./Econ.
Mar2020Panorama	1	0,124%	55	6,799%	1,82%
ABR2020Panorama	2	0,176%	31	2,729%	6,45%
Mai2020Panorama	9	1,163%	27	3,488%	33,33%
Jun2020Panorama	3	0,338%	55	6,194%	5,45%
Jul2020Panorama	4	0,691%	5	0,864%	80,00%
Ago2020Panorama	1	0,149%	40	5,979%	2,50%
Set2020Panorama	1	0,155%	36	5,581%	2,78%
Out2020Panorama	7	0,873%	27	3,367%	25,93%
Nov2020Panorama	8	0,681%	17	1,448%	47,06%
Dez2020Panorama	8	1,153%	23	3,314%	34,78%
Jan2021Panorama	4	0,542%	27	3,659%	14,81%
Fev2021Panorama	7	1,019%	21	3,057%	33,33%
Mar2021Panorama	2	0,197%	48	4,734%	4,17%
Abr2021Panorama	1	0,154%	21	3,226%	4,76%
Mai2021Panorama	1	0,139%	28	3,883%	3,57%
Jun2021Panorama	5	0,746%	19	2,836%	26,32%
Jul2021Panorama	19	0,455%	74	1,772%	25,68%
Ago2021Panorama	8	0,660%	33	2,723%	24,24%
Set2021Panorama	3	0,391%	36	4,688%	8,33%
Out2021Panorama	4	0,488%	20	2,439%	20,00%
Nov2021Panorama	2	0,309%	26	4,019%	7,69%
Dez2021Panorama	3	0,317%	53	5,603%	5,66%
Jan2022Panorama	2	0,128%	65	4,159%	3,08%
Fev2022Panorama	0	0,000%	24	4,131%	0,00%
Mar2022Panorama	4	0,490%	39	4,779%	10,26%
Média	**4,50**	**0,476%**	**33,13**	**3,695%**	**17,92%**

Fonte: AUTOR (2022)

Adicionalmente, para poder entender quais os temas com maior índice de Pandemicidade e Economicidade, foi montado um quadro comparativo com os parâmetros médios de cada tema. Assim, foi possível direcionar o foco e analisar mais profundamente os temas que refletissem a maior quantidade de conteúdo sobre a pandemia e economia em seus textos.

Quadro 2: Comparativo de parâmetros por seção.

Gênero	Av. Qtd Pandêmica	Av. Pandemicidade	Av. Qtd Econômica	Av. Economicidade	Av. Pand/Econ
Introdução	19,13	0,680%	80,79	2,905%	23,79%
Ativ. Econômica	4,13	0,367%	24,52	2,179%	17,77%
Expec. dos empres. e consum.	5,08	0,695%	11,00	1,476%	55,58%
Mercado de Trabalho	3,25	0,505%	11,13	1,802%	33,18%
Inflação	1,13	0,141%	26,83	3,426%	4,36%
Pol. Monetária	0,88	0,094%	33,13	3,508%	2,88%
Pol. Fiscal	5,78	0,525%	24,09	2,210%	34,52%
Setor Externo	1,88	0,148%	35,13	2,988%	5,56%
Panorama Intern.	4,50	0,476%	33,13	3,695%	17,92%
Em foco	11,46	0,671%	32,25	1,796%	59,50%
Obs. Político	1,92	0,170%	3,75	0,352%	60,51%
Média	**5,38**	**0,407%**	**28,70**	**2,394%**	**28,69%**

Fonte: AUTOR (2022)

Tendo compilado e processado os dados obtidos, foi possível partir para a análise mais profunda dos resultados obtidos de maneira a entender se os impactos da pandemia foram refletidos na linguagem do corpus estudado e, em caso positivo, de que maneira isso ocorreu.

4. ANÁLISE DE RESULTADOS

Tendo efetuado o processamento dos dados e levantamento dos indicadores para todo o corpus, foi possível realizar alguns cruzamentos de dados adicionais para identificar de forma mais clara o reflexo do impacto da pandemia na economia brasileira no que se refere à linguagem.

Através da classificação das palavras do corpus em pandêmicas e econômicas, foi possível obter os valores das Densidades Pandêmica e Econômica para todo o corpus. Os indicadores foram calculados pela relação entre a quantidade de palavras de conteúdo pandêmico e econômico e o total de palavras no corpus. Como mencionado anteriormente, os valores obtidos para os indicadores de Densidade Pandêmica e Econômica, para o corpus como um todo foram, respectivamente, 0,60% e 3,17%. Vale ressaltar que o corpus possui um tamanho total de 276.521 palavras.

O primeiro resultado a ser analisado traz uma informação extremamente relevante a respeito da validade e aplicabilidade deste estudo. Ao comparar o indicador de Diversidade Lexical (TTR), que mede a variedade de palavras diferentes utilizadas no texto, para as palavras pandêmicas e econômicas, percebemos que o TTR pandêmico é ligeiramente maior do que o econômico, o que mostra uma maior redundância relativa no léxico econômico do que no pandêmico. A tabela abaixo mostra os resultados.

Quadro 3: Comparativo Type vs Token

	Pandêmico	Econômico
Token	1654	8768
Type	75	334
TTR	4,53%	3,81%

Fonte: Autor (2022)

Apesar do resultado ser muito próximo, sendo o TTR pandêmico minimamente mais alto do que o econômico, vale ressaltar que o corpus tem como fonte um veículo de informação predominantemente relacionado a economia. Dessa forma, a TTR pandêmico mais alto reforça a intensidade e variedade com as quais o assunto da Pandemia foi abordado ao longo das edições.

Inicialmente, foram elencadas as seções que, de forma geral, tiveram maior Densidade Pandêmica e Econômica. Para isso, com base no levantamento dos valores médios dos

parâmetros para cada seção, os mesmos foram ranqueados em ordem decrescente para evidenciar aqueles que obtiveram os maiores valores para cada um dos parâmetros.

Tabela 3 - Ranking Pandemicidade por seção

Seção	Av. Pandemicidade
Expec. Dos emp. E consum.	0,695%
Introdução	0,680%
Em foco	0,671%
Pol. Fiscal	0,525%
Mercado de Trabalho	0,505%
Panorama Intern.	0,476%
Média	**0,407%**
Ativ. Econômica	0,367%
Obs. Político	0,170%
Setor Externo	0,148%
Inflação	0,141%
Pol. Monetária	0,094%

Fonte: AUTOR (2022)

Conforme observado na tabela acima, a seção que teve maior incidência relativa de tokens pandêmicos, em média, foi “Expectativa dos Empresários e Consumidores”, que trata de maneira geral, a respeito da confiança dos participantes da economia real sobre o futuro do ambiente de negócios e consumo no país, baseando-se principalmente na análise da evolução dos chamados Índices de Confiança. Em segundo e terceiro lugar estão, respectivamente os temas “Introdução” e “Em foco: IBRE”. O primeiro traz no início de cada edição um breve resumo com os assuntos tratados em cada seção ao longo do boletim, enquanto o segundo faz uma análise mais aprofundada a respeito de um assunto que foi considerado destaque no mês de referência.

Tendo definido as seções mais impactadas, em termos de linguagem, pela pandemia, foi possível analisar em quais textos, dentro de cada uma destas seções, o impacto foi mais percebido. Sendo assim, da mesma forma que os temas foram ranqueados de acordo com sua Pandemicidade, o procedimento foi repetido para os textos pertencentes a cada um deles.

Abaixo o ranking de Pandemicidade para os textos das três seções com maior Densidade Pandêmica média, evidenciando os cinco primeiros de cada um.

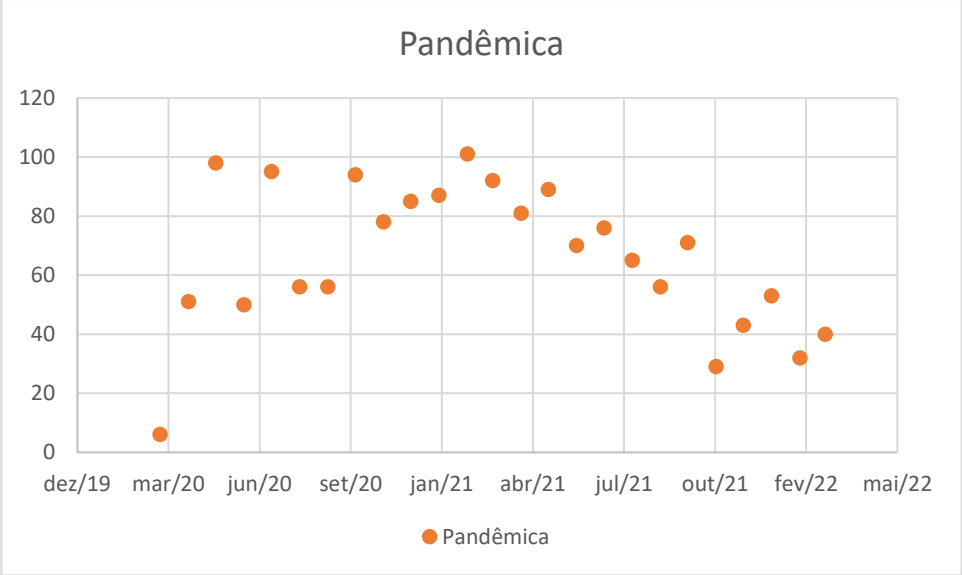
Tabela 4 - Ranking Pandemicidade – Principais textos

Tema: Expectativa de Empresários e Consumidores		Tema: Introdução		Tema: Em foco: IBRE	
Seção	Pandemicidade	**Seção**	**Pandemicidade**	**Seção**	**Pandemicidade**
Fev2021Expectativa	1,615%	Jan2021INTRO	1,316%	Mai2020IBRE	1,750%
Mai2021Expectativa	1,226%	Jul2021INTRO	1,017%	Mar2021IBRE	1,578%
Jul2021Expectativa	1,223%	Jun2021INTRO	0,997%	Set2021IBRE	1,123%
Ago2021Expectativa	1,108%	Abr2021INTRO	0,966%	Jul2020IBRE	1,057%
Jan2021Expectativa	1,057%	Fev2021INTRO	0,958%	Mar2022IBRE	0,967%

Fonte: Autor (2022)

Pode-se observar que dentre os textos destacados acima pela alta Pandemicidade, a maioria (60%) é referente ao período do primeiro semestre do ano de 2021, período este que coincide com a segunda onda da pandemia de coronavírus no Brasil, quando, após uma redução no número de casos, a pandemia voltou a avançar, resultando em médias móveis de casos novos ainda maiores do que as vistas na primeira onda.

Gráfico 7 – Gráfico Dispersão Temporal - Termos Pandêmicos

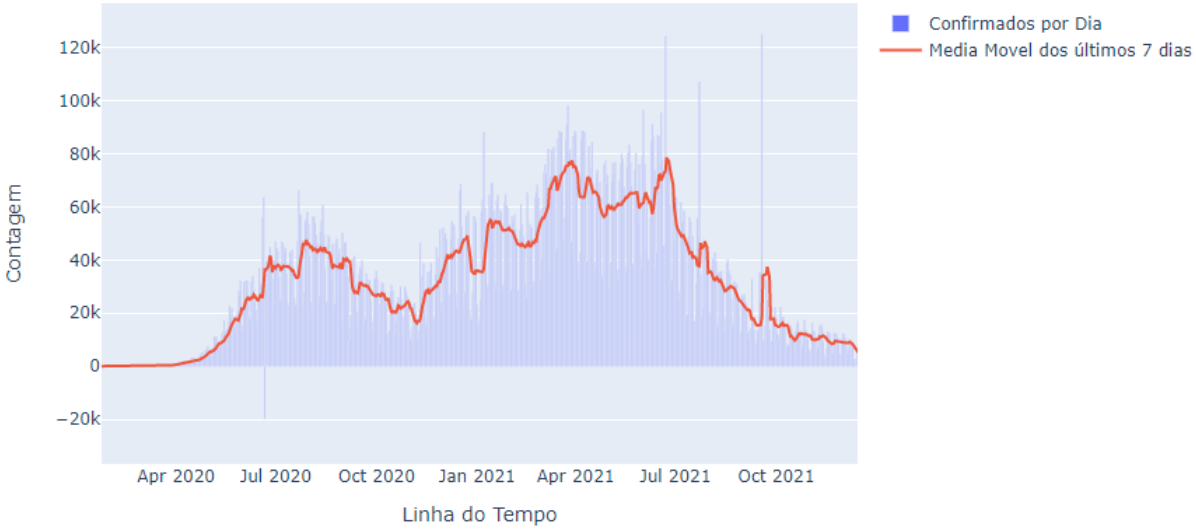


Fonte: Autor (2022)

Para ilustrar essa correlação entre o pico de pandemicidade lexical identificado no corpus com o período da segunda onda da pandemia no Brasil, podemos comparar o gráfico 7, acima, que mostra a dispersão temporal dos tokens pandêmicos no período da pandemia com o gráfico 8, abaixo, que mostra o número de novos casos de COVID-19 no Brasil. É possivel observar certa similaridade entre os dois gráficos, ao perceber que tanto o pico de novos casos

quanto o período com maior incidência de tokens pandêmicos se encontram no primeiro semestre de 2021

Gráfico 8 – Novos Casos - Pandemia



Fonte: USP (2022)

Outro resultado interessante que pode ser observado é a presença de apenas um texto do primeiro semestre de 2020, período em que se iniciou a pandemia no Brasil, entre aqueles com maior Densidade Pandêmica. Uma hipótese é que, apesar do impacto causado pela pandemia neste período ter sido inédito, intenso e, portanto, exaustivamente retratado em qualquer forma de mídia na época, as palavras, expressões e jargões comumente utilizados para tratar a respeito do COVID-19 ainda não estavam tão fortemente incorporadas a linguagem econômica, sendo seu uso acentuado ao longo do tempo.

A análise realizada a partir do indicador de Densidade Pandêmica foi repetida para a Densidade Econômica. Assim, foi possível identificar as seções com maior Economicidade e, dentro delas, os textos com os níveis mais altos para o indicador.

Tabela 5 - Ranking Economicidade por seção

Tema	Av. Economicidade
Panorama Intern.	3,695%
Pol. Monetária	3,508%
Inflação	3,426%
Setor Externo	2,988%
Introdução	2,905%
Média	2,394%
Pol. Fiscal	2,210%
Ativ. Econômica	2,179%
Mercado de Trabalho	1,802%
Em foco	1,796%
Expec. dos empresários e consum.	1,476%
Obs. Político	0,352%

Fonte: AUTOR (2022)

As seções com maior Densidade Econômica foram “Panorama Internacional”, “Política Monetária” e “Inflação”. O resultado não surpreende. De forma geral, devido a fonte do corpus ser um boletim de macroeconomia, a Economicidade é consideravelmente maior do que a Pandemicidade na maioria dos textos, com algumas poucas exceções. As seções supramencionadas tratam de assuntos altamente especializados da economia, de forma que, apesar de haver alguma referência ao tema da pandemia, essas ocorrências são ofuscadas pela maior frequência de termos e jargões específicos da economia que fazem parte do vocabulário frequente na discussão de tais assuntos.

5. DISCUSSÃO

O presente trabalho teve como principal contribuição o desenvolvimento de um novo método de análise linguística de um corpus formado por mais de um domínio de assunto, no caso, o econômico e pandêmico. Normalmente, em uma análise de PLN, um corpus é estruturado de forma que seja focado em um tema específico e cuja análise permita entender as particularidades do comportamento da língua neste tema. Dessa forma, um método que permita analisar um corpus que englobe mais de um domínio, permitindo obter resultados a partir da comparação entre os temas, sofre da dificuldade de encontrar estudos acadêmicos anteriores que apoiem a análise.

Considerando a escassez de estudos, tanto em estudos gerais de PLN para a língua portuguesa, quanto para um método mais específico como o aplicado neste estudo, identificou-se apenas um estudo acadêmico que possui uma abordagem similar ao problema de tratar um corpus que seja composto por lexicalidades distintas. O estudo de Assis et al (2021) analisou o comportamento do léxico biomédico em um corpus formado por textos que tratam sobre a COVID-19 dentro do contexto médico.

De forma similar ao presente estudo, os autores tem como base para sua análise de dados uma linguagem de especialidade, notadamente, a biomédica. Eles refletem sobre as especificidades de se trabalhar com um corpus formado por uma linguagem deste perfil quando citam que na visão de Perna, Delgado e Finatto, (2010 p.138), "(...) a partir da observação da linguagem especializada em corpora que se percebe mais francamente como a observação de termos é somente um pequeno passo na observação do texto especializado". O trecho trazido reflete sobre a necessidade de uma análise mais profunda, detalhada e customizada ao trabalhar com linguagens especializadas, que justifiquem estudos como este.

Retomando a reflexão proposta pelo estudo de Assis et al (2021), identifica-se que os autores definiram como necessário a definição de um novo indicador que pudesse analisar o comportamento do léxico biomédico dentro do corpus, como explicitado a seguir:

> A observação inicial sobre o comportamento lexical das linguagens de especialidades nos textos do corpus e a indefinição das medidas de densidade lexical nesse contexto levou a busca de um novo indicador da lexicalidade no corpus. Propomos aqui o conceito de "lexicalidade biomédica" ou "densidade lexical biomédica" para identificar com maior segurança o espaço lexical que é das especialidades biomédicas e diferenciá-lo de um léxico fronteiriço revelado em gêneros menos técnicos. (ASSIS ET AL, 2021)

Esta abordagem se mostrou válida e necessária também neste estudo. Para identificar o impacto da pandemia sobre o cenário econômico brasileiro foi necessário estabelecer os indicadores de Densidade Lexical Pandêmica e Econômica, utilizando da mesma métrica definida por Assis et al (2021) em seu estudo.

Conforme visto acima, o estudo de Assis et al (2021) criou uma nova forma de mensurar a densidade lexical, representada pelo índice Lex-Biomed, que visava analisar o comportamento do léxico biomédico dentro do corpus. Dado a proximidade do tema da biomedicina com o da pandemia, analisado neste estudo, somado ao fato de que o indicador foi construído tendo como base um corpus a respeito do COVID-19, um possível desenvolvimento futuro para este trabalho seria aplicar o indicador Lex-Biomed na análise do corpus construído neste trabalho, para auxiliar a mapear o comportamento do léxico pandêmico frente ao econômico, bem como comparar as diferenças entre os métodos.

A diferença entre o estudo de Assis et al (2021) e o proposto neste trabalho é a de que, no primeiro, o corpus era composto de apenas um domínio do conhecimento, o biomédico, e os autores analisaram as diferenças de comportamento entre as especialidades médicas e os gêneros textuais identificados naquele corpus. Já este estudo conduz uma análise buscando identificar o impacto causado pelo léxico pandêmico no econômico, em um corpus naturalmente econômico. Ou seja, aborda a relação entre léxicos distintos e como essa relação se traduz em impactos reais e passíveis de análise.

Portanto, apesar da escassez de conteúdo acadêmico acerca do tema, foi possível identificar similaridades nos estudos recém mencionados e inclusive aplicar métricas utilizadas em tais estudos no presente trabalho. Considera-se então que o aporte teórico de Assis et al (2021) foi um conteúdo norteador para desenhar o escopo deste trabalho.

6. CONCLUSÃO

Este trabalho estudou o comportamento da linguagem na intersecção entre os domínios econômico e da saúde.

Com ajuda de ferramentas computacionais e linguísticas detectou-se, no período da pandemia de Covid-19, expressões típicas da lexicalidade pandêmica e buscou-se examinar sua distribuição ao longo de 25 boletins econômicos da FGV, e das 10/11 seções que compõem cada edição do boletim. Seguindo o linguista clássico J.R. Firth, em sua monumental obra *Papers in Linguistics, 1934-1951*, partimos da hipótese de que o modo como os usuários escolhem as palavras na comunicação não é aleatório e apresenta regularidades, o que nos dias de hoje pode ser detectado com ajuda de recursos de processamento de linguagem natural (FIRTH, 1957). O problema inicial de pesquisa era mensurar o impacto da pandemia na conjuntura econômica brasileira através de um mapeamento das palavras usadas em textos do domínio econômico. Através da elaboração de um corpus composto de 262 textos pudemos extrair dados capazes de identificar como itens lexicais expressando conteúdos tipicamente pandêmicos foram utilizados em associação a seções do boletim tratando de temas prevalentemente econômicos como expectativas dos empresários e consumidores, política fiscal, mercado de trabalho, etc. Também foi possível revelar no período estudado alguns dos momentos em que a pandemia se tornou tema preferencial na discussão do cenário econômico brasileiro.

Os resultados obtidos e analisados através do processamento do corpus construído para este trabalho permitem concluir que o método estabelecido no estudo foi satisfatoriamente bem sucedido ao identificar os níveis de Lexicalidade Pandêmica e Econômica dentro do texto, além de mapear momentos de desvio desses parâmetros, permitindo analisá-los de forma mais detalhada para entender suas possíveis causas e estabelecer algumas relações com a evolução cronológica da pandemia de COVID-19 no país.

Dessa forma, o método se mostra uma ferramenta válida para uma análise de PLN onde se deseja efetuar uma comparação de dois domínios de assuntos distintos, buscando entender como um se comporta em relação ao outro e, após isso, abrindo caminho para que seja identificada a natureza da relação entre eles. É interessante ressaltar que em um estudo de PLN os corpora geralmente pertencem a um domínio apenas. Assim, a construção, classificação e análise de um corpus que transita entre dois domínios distintos é de certa forma inovadora,

podendo ser aprimorada e aprofundada posteriormente para alcançar análises extremamente relevantes.

Como sugestão de trabalho futuro, deve-se destacar a importância de se refinar o método aqui apresentado, aumentando a qualidade e o poder de seu processamento de dados, de forma que se possa fazer uso máximo das informações estatísticas e metadados extraídos do corpus. No caso deste trabalho, por exemplo, há espaço para uma melhor utilização dos dados a respeito das POS do corpus que, apesar de terem sido levantadas, não chegaram a ser analisadas detalhadamente. Além disso, um maior poder de processamento de dados poderia agregar ao estudo uma análise semântica das expressões formadas por mais de uma palavra, ao analisa-las em conjunto, considerando seu contexto e significado no texto.

7. REFERÊNCIAS BIBLIOGRÁFICAS

ALUÍSIO, Sandra Maria; ALMEIDA, Gladis Maria de Barcellos. O que é e como se constrói um corpus? Lições aprendidas na compilação de vários corpora para pesquisa linguística. 2006. Disponível em: http://revistas.unisinos.br/index.php/calidoscopio/article/view/6002. Acesso em: 20 nov. 2021.

ASSIS, Karhyne S. Padilha de; SILVA, Camila das Merces; LEITE, Janaína da Silva; NOGUEIRA, Wellington Araujo; NOSE FILHO, Kenji; TAKAHATA, André K.; STEINBERGER-ELIAS, Margarethe. Lexicalidade biomédica e sua mensuração em um corpus sobre COVID-19 em língua portuguesa. In: SIMPÓSIO BRASILEIRO DE TECNOLOGIA DA INFORMAÇÃO E DA LINGUAGEM HUMANA (STIL), 13. , 2021, Evento Online. Anais [...]. Porto Alegre: Sociedade Brasileira de Computação, 2021 . p. 39-46. DOI: https://doi.org/10.5753/stil.2021.17782.

FIGUEIREDO FILHO, Dalson Britto; SILVA JÚNIOR, José Alexandre da. Desvendando os Mistérios do Coeficiente de Correlação de Pearson (r). **Revista Politica Hoje**, Recife, v. 1, p. 115-146, 24 ago. 2022. Disponível em: https://periodicos.ufpe.br/revistas/politicahoje/article/viewFile/3852/3156. Acesso em: 24 ago. 2022

FIRTH, J. R. Papers in linguistics: 1934-1951. London: Oxford University Press, 1957. 233 p.

HASAN, R. Rationality in everyday talk: From process to system. In Svartvik J. Directions in Corpus Linguistics. Berlin: Mounton de Gruyter; 1992. p. 257-307.

IBGE. . **IPCA**: índice nacional de preços ao consumidor amplo. Índice Nacional de Preços ao Consumidor Amplo. 2022. Disponível em: https://www.ibge.gov.br/estatisticas/economicas/precos-e-custos/9256-indice-nacional-de-precos-ao-consumidor-amplo.html?=&t=conceitos-e-metodossource=landing&utm_medium=explica&utm_campaign=desemprego. Acesso em: 18 ago. 2022.

IBGE. . **PNAD Contínua**: pesquisa nacional por amostra de domicílios. Pesquisa Nacional por Amostra de Domicílios. 2022. Disponível em: https://www.ibge.gov.br/estatisticas/sociais/trabalho/9173-pesquisa-nacional-por-amostra-de-domicilios-continua-trimestral.html?=&t=o-que-e&utm_source=landing&utm_medium=explica&utm_campaign=desemprego. Acesso em: 18 ago. 2022.

IBRE FGV. Boletim Macro. 2022b. Disponível em: https://portalibre.fgv.br/boletim-macro. Acesso em: 26 maio 2022.

IBRE FGV. Quem Somos. 2022a. Disponível em: https://portalibre.fgv.br/quem-somos. Acesso em: 26 maio 2022.

JOHANSSON, Victoria. **Lexical Diversity and Lexical Density in speech and writing**: a developmental perspective. Suécia: Working Papers, 2008. p. 61-79. Disponível em: https://journals.lub.lu.se/LWPL/article/view/2273/1848. Acesso em: 11 jun. 2022.

LEITE, Janaína da Silva; TAKAHATA, André Kazuo; STEINBERGER-ELIAS, Margarethe. Elaboração de corpus biomédico em Português sobre o Covid-19. 2020. Disponível em: http://www.jhi-sbis.saude.ws/ojs-jhi/index.php/jhi-sbis/article/view/821. Acesso em: 01 maio 2022.

McENERY, T. e WILSON, A. 1996. Corpus linguistics. Edinburgh,Edinburgh University Press.

Perna, L. Cristina; Delgado, K. Heloísa; Finatto, J. Maria. (2010) "Linguagens Especializadas em CORPORA. Modos de Dizer e Interfaces de Pesquisa". EDIPUCS-Editora Universitária da Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre, p.138.

SARDINHA, Tony Berber. **LINGÜÍSTICADE CORPUS**: histórico e problemática. 02. ed. São Paulo: Delta, 2000. p. 323-367. Disponível em: https://revistas.pucsp.br/index.php/delta/article/view/39903/26975. Acesso em: 09 jul. 2022.

USP. . **COVID-19**: monitoramento ⁝ brasil. Monitoramento – Brasil. 2022. Disponível em: https://ciis.fmrp.usp.br/covid19/brasil/. Acesso em: 29 out. 2022