FUNDAÇÃO GETULIO VARGAS
ESCOLA DE ADMINISTRAÇÃO DE EMPRESAS DE SÃO PAULO

Programa Institucional de Bolsas de Iniciação Científica (PIBIC)

Análise de sentimentos aplicada ao estudo do conteúdo das entrevistas dos candidatos à presidência no Brasil de 2022

VINÍCIUS DE ALCÂNTARA MOTA

JÚLIO CÉSAR BASTOS DE FIGUEIREDO

São Paulo – SP

2023

Análise de sentimentos aplicada ao estudo do conteúdo

das entrevistas dos candidatos à presidência no Brasil de 2022

Resumo

A análise de sentimentos pode ser considerada uma área de estudo relacionada tanto a mineração de dados, quanto a processamento de linguagem natural. Aplicando diferentes abordagens, como aprendizado de máquina e abordagem baseada no léxico, é possível inferir quais as opiniões, emoções e polaridade de determinado discurso. Mesmo com alguns pontos de atenção necessários na sua utilização, em particular, na área das ciências políticas, as técnicas de mineração de textos adquirem papel inovador e crucial, uma vez que diminuem o custo de análise ligado a grande quantidade de textos presentes nesse contexto de discursos. Esta pesquisa aplica a abordagem baseada em dicionário no estudo da polaridade dos discursos dos candidatos à presidência da república do Brasil em 2022, bem como no estudo de suas respectivas semelhanças e diferenças. As eleições presidenciais de 2022, foram marcadas pela considerável polarização, a diferença de votos em segundo turno entre os candidatos Luís Inácio Lula da Silva e Jair Messias Bolsonaro atingiu um valor excepcionalmente baixo, estabelecendo um novo marco desde a redemocratização em 1985. Para as comparações e inferências realizadas no estudo, foram selecionados os quatro candidatos mais expressivos em intenções de votos em entrevistas realizadas por plataformas relevantes, por meio de uma abordagem quantitativa com elaboração de software em Python. A partir disso, obteve-se semelhanças das escolhas de palavras em diversos discursos e diferenças significativas das polaridades apresentadas e dos sentimentos e opiniões expressos.

Palavras-chaves

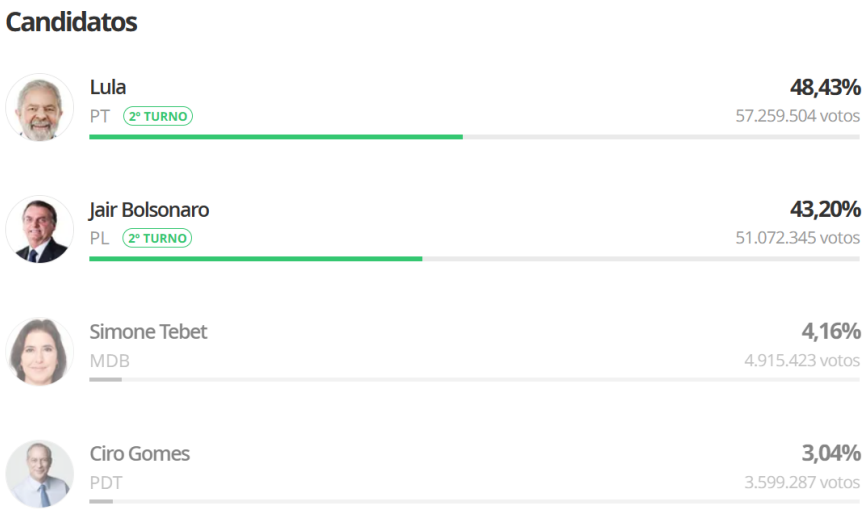
Análise de sentimentos, análise do discurso, discurso político, processamento de linguagem natural, polaridade do discurso

1. Introdução

No contexto político afetado pela pandemia do COVID-19, tanto na área econômica, com a redução de milhares de empregos formais (FAGUNDES, FELÍCIO E SCIARRETTA 2021), quanto na área da saúde, com mais de 700 mil mortes até o presente (PAINEL CORONAVÍRUS 2023), as eleições presidenciais tiveram um papel fundamental para garantir aos cidadãos a escolha do rumo que o país iria tomar no cenário de recuperação. Nessa conjuntura, foram debatidas diversas propostas, de maneira a refletir a diversidade de abordagens quanto aos desafios enfrentados pelos possíveis representantes políticos do mais alto cargo do poder executivo. Dessa maneira, resultando em um cenário de disputa intensa no contexto de votação, com uma diferença de apenas 2.1 milhões de votos no segundo turno entre os candidatos Bolsonaro e Lula, e pela primeira vez na história do Brasil, um presidente em exercício, ao finalizar seu primeiro mandato, perde a corrida presidencial para um ex-presidente.

Ao todo, 12 candidatos participaram das eleições a presidência em 2022, destacando-se quatro candidatos mais expressivos - em intenções de votos durante a campanha e votos válidos no primeiro turno, correspondendo a mais de 98% dos votos válidos durante o primeiro turno das eleições (Figura 1), abordados nessa pesquisa.

Figura 1 – Apuração em 1° turno dos 4 candidatos mais votados para a eleição presidencial de 2022



Fonte: G1- GLOBO, 2022

Com o objetivo de compreender as diferentes escolhas lexicais, sentimentos e opiniões expressas pelos candidatos, este trabalho apresenta um estudo por meio da análise de sentimentos sobre os discursos dos principais presidenciáveis. Para isto foi utilizado uma abordagem quantitativa, uma vez fundamental a aplicação de métodos de mineração de texto

pela grande quantidade de dados em forma textual analisados, correspondendo a mais de 150 mil palavras extraídas e mais de 65 mil palavras analisadas na totalidade dos discursos.

Segundo Iezzi et al. (2021) em seu trabalho *Text Analytics: Present, Past and Future*, visando explicitar o passado, presente e possível futuro para a área de análises de texto (e mineração de texto), pode-se afirmar que o período atual é marcado pela imensa disponibilidade de dados, explicado pelas integrações devido à internet e pelo surgimento de redes sociais. Esses dados são em maior parte não estruturados, como vídeos, áudios, textos e imagens e não podem ser processados diretamente pelos métodos tradicionais. Para isso, faz-se necessário a utilização de métodos e técnicas para retirar informações relevantes. (FAN et al., 2006).

Ao se tratar de dados em formato textual, é necessário a utilização de Processamento de Linguagem Natural (PNL) para extrair informações (KUMAR; KAR; ILAVARASAN, 2021). Este processamento pode ser aplicado a diversas áreas como em análises de textos jurídicos, diagnósticos médicos, avaliações de produtos em sites e discursos políticos, por exemplo. Com base nisso, pode-se compreender a Análise de Sentimentos como área de estudo da computação, relacionada a Processamento de Linguagem Natural (PLN) (MACHADO, 2018) e, da mesma forma, dentro da mineração de texto (MEDHAT; HASSAN; KORASHY, 2014).

A Análise de Sentimentos pode ser empregada a partir de diferentes abordagens para determinação dos sentimentos de uma sentença (MEDHAT; HASSAN; KORASHY, 2014), concentrando-se em métodos baseados em aprendizado de máquina, métodos baseados em léxico e métodos híbridos, sendo o último a combinação dos dois primeiros.

Não se limitando ao tema do estudo, a abordagem definida tem impacto em outras áreas do conhecimento. Na área da administração sua aplicação é fundamental pelo crescente número de informação geradas na WEB, o que torna a obtenção de informação, por exemplo, da avaliação de clientes sobre determinados produtos, uma tarefa complexa para análises de forma totalmente manual. Nos estudos realizados por Avanço e Nunes (2014) e Gonçalves (2016), são retratados a aplicação de métodos quantitativos e diferentes abordagens relacionadas a análise de sentimentos para extrair informações relevantes a partir de comentários em sites de recomendação de produtos, Buscapé e de reclamações de empresas, Reclame Aqui, respectivamente. Enquanto o estudo realizado por Gonçalves (2016) utiliza métodos de aprendizado de máquina para determinação dos sentimentos, Avanço e Nunes (2014) opta pela utilização da abordagem baseado no léxico, com a aplicação de diferentes dicionários na análise. Um dos benefícios da aplicação dessa abordagem, é a relativa facilidade de aplicação e de não

precisar de um conjunto de dados pré-rotulados, como em uma abordagem por aprendizado de máquina supervisionada.

Outro estudo que vale destaque, e que possui maior proximidade com esta pesquisa, é a utilização dessa forma de tratamento no campo da política. O trabalho realizado por Delavald (2018), buscou estudar a polaridade dos comentários feitos na rede social, Twitter, para entender qual a opinião dos brasileiros que utilizavam a plataforma, quanto a intervenção militar após o impeachment da ex-presidente Dilma Rousseff.

Por fim, para estudar as respectivas diferenças e semelhanças entre os discursos apresentados pelos quatro principais candidatos, analisou-se quais foram as palavras mais utilizadas no conjunto das entrevistas, as especificidades da escolha lexical e em que proporção os discursos dos presidenciáveis estavam polarizados entre negativo e positivo, de forma que os resultados possam contribuir para o conhecimento na área de análise de sentimentos, do estudo dos discursos políticos e em outras futuras análises e pesquisas.

Este trabalho é apresentado da seguinte forma: (i) apresentação da teoria sobre o tema, tendo em vista o foco do estudo em mostrar a aplicabilidade da análise de sentimentos na automatização de extração de informações a partir de documentos em linguagem natural e do discurso político, (ii) metodologia aplicada, consistindo de uma abordagem quantitativa, por meio da coleta de dados, tratamento e elaboração de software em Python com o propósito de realizar as análises, (iii) resultados alcançados, onde são apresentados os principais achados do estudos, (iv) discussão e problematização dos resultados das pesquisa a partir da ótica do autor e (v) conclusão, na qual são feitas as considerações finais quanto a realização do estudo.

2. Teoria

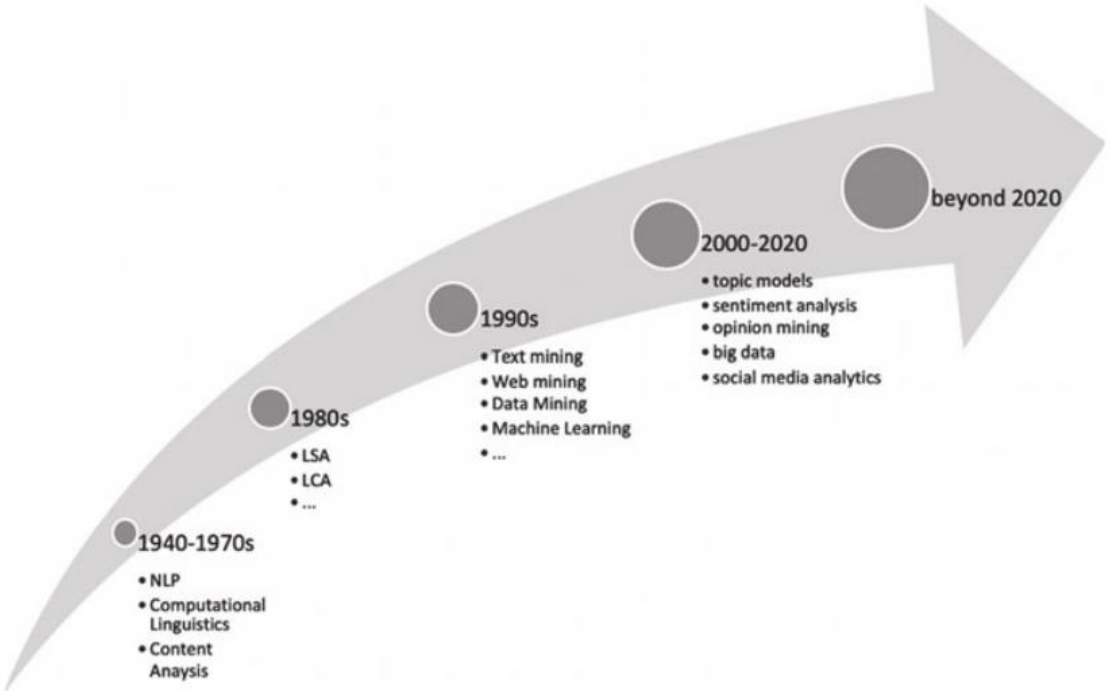
2.1 Text Analytics

A Análise de Texto (*Text Analytics*) pode ser considerada uma área de estudo ampla que engloba diferentes métodos e técnicas relacionadas a análises quantitativas de dados em forma de texto (Figura 2). Seu surgimento vem desde a criação do computador e atualmente possui ainda mais destaque devido ao crescente número de dados não estruturados presentes na internet, tanto de documentos disponibilizados, quanto de sites e redes sociais (IEZZI et al., 2021).

Pode-se afirmar que a origem da primeira análise quantitativa de textos remete desde antes ao desenvolvimento do primeiro computador. Autores como Benjamin Bourdon, Keading, Estoup,

1888, 1889 e 1907 respectivamente, faziam as primeiras publicações relacionadas ao tema, elaborando análises quantitativas sobre a frequência de palavras em determinados textos (IEZZI et al., 2021).

Figura 2 – Linha do cronológica sobre técnicas de Análise de Texto (*Text Analytics*)



Fonte: IEZZI et al. (2021)

Para Grimmer e Stewart (2013), a análise de textos de forma automatizada surge como solução ao desafio do alto custo para realizar análises em grande quantidade de dados textuais, em particular o autor define como solução na área de ciências políticas. Segundo os autores para garantir a eficácia de sua aplicação é necessário estar atento às armadilhas que podem surgir durante a utilização desse método. Nesse aspecto é estabelecido pelos autores quatro princípios para a análise de textos de forma automatizada, que podem ser descritos respectivamente:

(i) Todos os métodos quantitativos para linguagem estão errados, mesmo que úteis: apesar de possibilitar a extração informações relevantes, a complexidade da linguagem implica que em todos os métodos existe uma falha inerente;

(ii) Os métodos amplificam a capacidade de compreensão pelos seres humanos, mas não os substituem: mesmo com diversas vantagens de aplicação, ainda é preciso que um pesquisador determine o problema a ser analisado, etapas do processo, a validade dos *outputs* e que seja necessária uma leitura detalhada em alguns casos;

(iii) Não existe um método melhor geral para ser aplicado em análise de textos: diferentes problemas de pesquisa implicam em métodos específicos a serem aplicados;

(iv) Validar diversas vezes: apesar de útil e mais rápido que uma abordagem não automatizada, as informações retiradas podem estar erradas e por esse motivo deve se atentar na validação do que é extraído.

Devido a grande quantidade de documentos textuais, o Processamento de Linguagem Natural e a Mineração de Texto passam a ser essenciais dado a impossibilidade de a análise do conteúdo ser feita de forma manual (KUMAR; KAR; ILAVARASAN, 2021). A Mineração de Texto se refere ao processo de obtenção de informações relevantes de dados não estruturados, podendo ser descrita como uma área da programação criada para solucionar a deficiência da compreensão da linguagem natural por computadores (MACHADO et al., 2010). O Processamento de Linguagem Natural é uma área de pesquisa que visa compreender como os computadores podem ser usados para interpretar a linguagem natural na forma de texto ou falada pela pessoas em diversos idiomas (CHOWDHURY, 2003).

2.1.1 Análise de Sentimentos

A partir desses aspectos pode se interpretar a Análise de Sentimentos como área de estudo da computação aplicada, mais especificamente, relacionada a Processamento de Linguagem Natural (PLN) (MACHADO, 2018) e, da mesma forma, dentro da Mineração de Texto (MEDHAT; HASSAN; KORASHY, 2014). A Análise de Sentimentos se concentra principalmente na determinação da polaridade, opinião, classificação da emoção, subjetividade e grau expressado nos discursos analisados. Segundo, Souza et al. (2017), possui como base teórica a interpretação das emoções do emissor da mensagem por meio de linguagem natural.

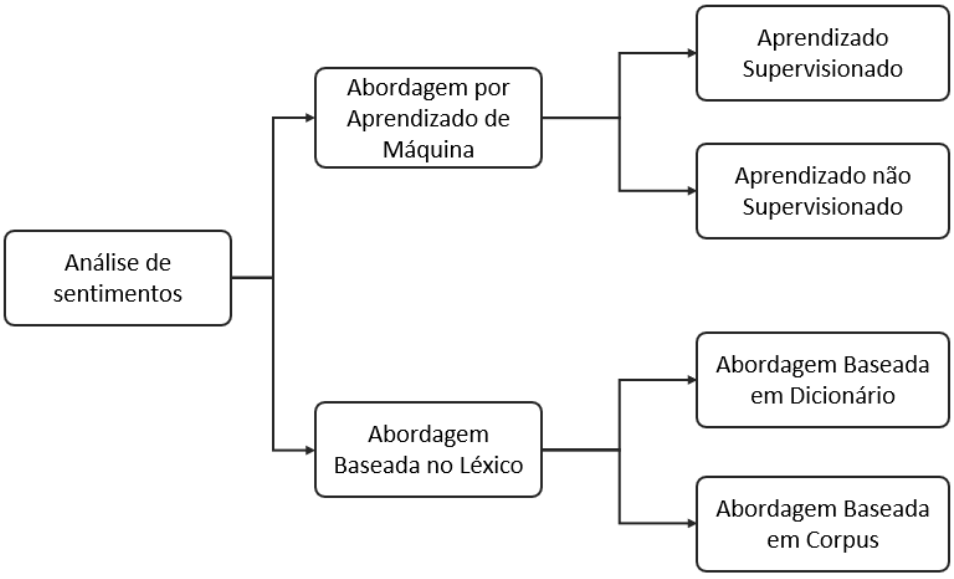
Dessa maneira, busca-se estudar a polaridade de um texto, ou seja, o grau de palavras positivas, negativas e neutras presentes, a opinião que indica o ponto de vista do autor sobre determinado assunto, a emoção que está relacionada a presença de um sentimento específico demonstrado no texto que remete a uma classificação (exemplo: raiva, felicidade, medo) e por último, o grau expressado, representando a intensidade de determinada polaridade ou emoção presentes no texto (GONÇALVES, 2015).

Um fator importante quanto a Análise de Sentimentos é a granularidade do texto analisado. Existem três classificações de granularidade quanto ao conjunto de textos: (i) nível do documento, (ii) nível da sentença e (iii) nível do aspecto. O nível do documento representa a análise do documento quanto a sua totalidade, ou seja, a polaridade do documento como um só,

enquanto, o nível de sentença visa classificar a polaridade por cada sentença do documento e por fim, o nível do aspecto diz respeito ao sentimento atribuído a cada aspecto presente na sentença. Exemplificando o último nível, a frase: “eu adorei a qualidade de imagem desse notebook, porém o processador é muito lento”, atribui diferentes polaridades para o notebook, relacionando a qualidade de imagem a uma avaliação positiva e o processador a uma avaliação de polaridade negativa (MEDHAT; HASSAN; KORASHY, 2014). Para Benevenuto et al., (2015), a escolha do nível do aspecto seria mais interessante para uma empresa avaliar a polaridade atribuída a avaliações de seus produtos ou serviços, pois existe a possibilidade das avaliações possuírem mais de um aspecto relacionado ao mesmo produto.

Ainda assim, segundo Medhat, Hassan e Korashy (2014), a Análise de Sentimentos pode ser dividida em três abordagens principais, sendo elas (i) métodos baseados em Aprendizado de Máquinas (*Machine Learning*), (ii) métodos baseados em léxicos e (iii) métodos híbridos, sendo o último uma combinação dos dois mencionados anteriormente.

Figura 3 – Técnicas de classificação de sentimentos



Fonte: Organizado pelo autor (2023) com base em Medhat, Hassan e Korashy (2014)

A abordagem por Aprendizado de Máquina utiliza os tradicionais modelos de Aprendizado de Máquina para realização da Análise de Sentimentos, podendo ser dividido em métodos supervisionados e não supervisionados (Figura 3).

Os métodos supervisionados exigem uma etapa de treinamento com uma grande quantidade de dados rotulados previamente, tornando mais elaborada a tarefa de inferência, enquanto os métodos não supervisionados, em contrapartida, não necessitam de dados previamente

rotulados para treinamento e teste, e muitas vezes usados quando existe uma dificuldade de obter a rotulação (BENEVENUTO et al., 2015). Ainda assim, ambos os métodos podem ser aplicados de maneira conjunta (MEDHAT; HASSAN; KORASHY, 2014).

Por outro lado, a abordagem por léxico não requer qualquer classificação do conjunto de palavras a ser estudado ou etapa de treinamento e pode ser dividido em dois métodos: métodos baseados em léxico e métodos baseados em dicionário (Figura 3). Os métodos baseados em corpus buscam encontrar palavras que remetem opiniões e os itens que elas modificam, utilizando padrões sintáticos para comparar uma lista de palavras que remetem a opiniões. Podendo se dividir em métodos semânticos e métodos estatísticos (GONÇALVES, 2015). Enquanto os métodos baseados em dicionários utilizam uma lista de palavras chaves com classificações previamente rotuladas quanto a aspectos e categorias, no qual a aplicação é reconhecida como sendo a mais intuitiva e de fácil aplicação (GRIMMER; STEWART, 2013).

Para a determinação da polaridade em um texto, tendo em vista o último método mencionado, utiliza-se um dicionário com a classificação das palavras quanto a positivo e negativo, e atribui-se uma pontuação à frequência das palavras no texto que estão presentes no dicionário, de forma a estabelecer quais palavras indicam a polaridade e quantas vezes se repetem. Cabe salientar que nesse método diversas palavras podem não ter uma polaridade relacionada ou podem ser classificadas como neutras, dependendo do dicionário em estudo (AVANÇO; NUNES, 2014).

Vale ressaltar que para a performance da classificação de um dicionário seja satisfatória, a classificação e as palavras pertencentes ao léxico devem ser coerentes com o contexto em que aplicadas. Nesse aspecto, caso exista uma diferença substancial do contexto de aplicação do dicionário e do documento estudado, sérios erros podem ocorrer quanto a inferência (GRIMMER; STEWART, 2013).

Como mencionado por Gonçalves (2015), com a recente popularização da Análise de Sentimentos, diversas aplicações foram desenvolvidas em áreas como comércio, turismo, saúde, economia e, como tema de análise desta pesquisa, política. Uma pesquisa realizada por Delavald (2018) buscou estudar a polaridade dos discursos relacionados à crise política brasileira e reações sobre uma intervenção militar, por meio de palavras-chaves na rede social Twitter, após o impeachment da ex-presidente Dilma Rousseff até o ano de 2017. Para isso, o autor utilizou a abordagem por dicionário, recorrendo à utilização do LIWC2007 em português do Brasil, elaborado a partir da versão em inglês, de forma colaborativa por três equipes de pesquisadores de diferentes instituições (DE CARVALHO, 2019). A principal conclusão quanto ao tema da pesquisa foi que, inicialmente, os usuários da rede social demonstraram

opiniões que repudiavam o tema “intervenção militar”, demonstrado por mais emoções negativas na Análise de Polaridade, entretanto, ao decorrer dos anos analisados, houve uma diminuição entre a diferença das classificações quanto à negativo e positivo dos discursos presentes no Twitter, de maneira a evidenciar uma tendência de aceitação por parte de mais usuários.

Savoy (2010) apresenta outro trabalho relevante ao tema: *Lexical Analysis of US Political Speeches*, em que é abordada uma análise e comparação entre 245 discursos dados pelos senadores John McCain e Barack Obama, durante os anos de 2007 e 2008, utilizando uma abordagem quantitativa. Os dados foram coletados por meio dos sites oficiais dos candidatos à presidência dos Estados Unidos e utilizando técnicas de pré-processamento, o autor analisou as palavras mais frequentes por candidatos e as comparou por métodos estatísticos para determinar sua relevância.

Ainda assim, as aplicações da área de estudo não se limitam aos casos mencionados, no estudo elaborado por Avanço e Nunes (2014), é visado determinar a polaridade das avaliações de produtos postados por usuários na plataforma de recomendação, Buscapé. Os autores decidiram quanto a utilização de 3 dicionários diferentes para determinação da polaridade de avaliações quanto à categoria de telefones móveis e smartphones. A principal dificuldade enfrentada na pesquisa foi de determinar a polaridade de comentários em que existiam diferentes polaridades a cada aspecto na sentença.

Nos três estudos citados, nota-se a aplicação de métodos quantitativos para extração de informações relevantes dos dados textuais em grandes quantidades e de forma automatizada em diferentes fontes de dados não estruturados.

2.2 Discurso político

A importância da linguagem e seu uso na política é crucial (SIM et al., 2013). Segundo o estudo realizado pelos autores em 2013, por meio da linguagem empregada no discurso político é possível identificar em proporções quais ideologias o candidato utiliza em seu posicionamento. O discurso político tem sido um tema relevante na pesquisa linguística, dado a sua complexidade e papel central na estruturação da sociedade, uma vez que tem o poder de moldar ideias e formar opiniões das pessoas e, por conseguinte, da sociedade. A escolha lexical nesse tipo de discurso não está ligada somente ao meio de comunicação e ao ato de fazer declarações públicas, mas, vale enfatizar, a capacidade de manipular a sociedade, legitimar poder político e dar relevância a atitudes e opiniões políticas (DYLGJERII, 2017). A seleção de palavras

utilizadas pelos candidatos, vai além da aleatoriedade, sendo um instrumento do político para representar seus objetivos, conforme Savoy (2010).

Segundo Pinto (2009), o discurso político é um tipo de discurso que tem sua eficácia fundamentada na habilidade de impor uma determinada visão da realidade, enfrentando constantes ameaças provenientes de significações adversas. Nesse sentido, esse discurso busca estabelecer e impor uma realidade diante do cenário de disputa entre diferentes candidatos e ideais, revelando uma urgência implícita à sua existência. É importante destacar que esse tipo de comunicação está sempre sujeito à desconstrução, ao mesmo tempo em que se constrói pela desconstrução do outro, sendo intrinsecamente um discurso de poder e representa visões de mundo e, de forma explícita, possui lados declarados.

3. Metodologia

Como objetivo desta pesquisa, a aplicação da Análise de Sentimentos ao estudo dos discursos políticos, foi adotada uma abordagem quantitativa para a investigação das palavras escolhidas por candidato e a determinação da polaridade, a partir de dados qualitativos, estes sendo o conteúdo das entrevistas dado pelos principais candidatos à presidência em 2022. Portanto, a abordagem é definida como quantitativa, dado a aplicação por dicionário na área de estudo da Análise de Sentimentos.

3.1 Coleta dos dados

Ao limitar o estudo aos (quatro) principais candidatos em intenção de votos e votos validos em primeiro turno, foram selecionadas as entrevistas no estilo sabatina, disponibilizadas na plataforma de vídeos, Youtube. Ao todo, 43 entrevistas foram encontradas dos quatro candidatos: Luiz Inácio Lula da Silva, Jair Messias Bolsonaro, Simone Nassar Tebet e Ciro Ferreira Gomes. Dado a desproporção de entrevistas por candidato, optou-se por selecionar quatro realizadas por candidato, de maneira que, necessariamente estivesse no estilo de entrevista, preferencialmente de forma presencial e que houvesse a participação de mais de um candidato à presidência mencionado.

Desse modo, totalizou-se em 16 entrevistas por candidato distribuídos conforme a Tabela 1. Vale ressaltar que não foi encontrado para o candidato Luiz Inácio Lula da Silva, quatro entrevistas que seguiam os critérios estabelecidos de modalidade presencial e com participação de outros candidatos, uma vez que foram concedidas menos entrevistas pelo candidato. Assim escolheu-se as entrevistas concedidas ao UOL e a rádio Liberal FM.

Tabela 1 - Entrevistas por candidato à presidência da república do Brasil em 2022

Candidato	Entrevista	Duração	Data
Jair Bolsonaro	Correio braziliense e Tv Brasilia	1:21:53	08/09/2022
	TV Record: segundo turno	1:21:48	23/10/2022
	TV Record: primeiro turno	41:31	26/09/2022
	SBT - programa do Ratinho	30:59	13/09/2022
Ciro Gomes	CNN	53:45	01/09/2022
	TV Record: primeiro turno	41:44	27/09/2022
	Correio braziliense e Tv Brasilia	41:45	22/09/2022
	SBT - programa do Ratinho	31:34	19/09/2022
Lula	Uol Entrevista	1:38:43	27/07/2022
	Rádio Liberal FM	1:21:03	28/01/2022
	CNN	56:09	12/09/2022
	SBT - programa do Ratinho	31:43	22/09/2022
Simone Tebet	Correio braziliense e Tv Brasilia	1:10:12	06/09/2022
	CNN	53:37	29/08/2022
	TV Record: primeiro turno	41:48	28/09/2022
	SBT - programa do Ratinho	31:22	20/09/2022

Fonte: elaborado pelo autor (2023)

A fim de obter os discursos em forma de texto, foi utilizado a plataforma do Youtube para a transcrição das falas. Para tal, criou-se um *software*, em linguagem Python, utilizando uma Interface de Programação de Aplicação (*Application Programming Interface*) para obtenção, nomeação e armazenamento de forma automatizada das transcrições de todos as entrevistas em vídeo, as quais foram verificadas em seguida durante a etapa de pré-processamento dos dados.

3.2 Elaboração do Software

Com o objetivo de elaborar o *software* para a análise de sentimentos, foi escolhida a linguagem de programação Python. A decisão foi tomada devido a linguagem ser de propósito geral, o que possibilita a utilização para uma larga variedade de aplicações, existir diversas pesquisas e bibliotecas publicadas para o Processamento de Linguagem Natural, *open source* – em código aberto - e possuir poderosas ferramentas para manipulação e visualização de dados.

Pode-se dividir a elaboração em três partes: (i) pré-processamento, (ii) desenvolvimento de análises e (iii) visualização.

A importação dos dados em forma de arquivo TXT foi realizada por meio da biblioteca Pandas, que após a etapa de pré-processamento, resultou em um *DataFrame* (objeto principal da biblioteca semelhante a uma matriz, com colunas e linha rotuladas) com cada coluna especificando a palavra, o candidato e a entrevista em que foi expressa.

3.3 Pré-processamento

A etapa de pré-processamento é muito importante para qualquer atividade em relação a mineração de textos (AVANÇO; NUNES, 2014). Algumas etapas importantes são: a remoção de *stopwords*, transformação em tokens, correções quanto a ortografia, homogeneização da fonte - uma vez, que é reconhecido pelo sistema letras minúsculas e maiúsculas como símbolos com significados diferentes - e outras possíveis transformações. As *stopwords* são palavras que não agregam sentido específico para a frase, podendo ser conectivos, artigos e proposições e por isso não são relevantes (SOUZA et al., 2017).

Na pesquisa em específico, foram aplicados os métodos mencionados. As *stopwords* foram retiradas conforme a lista disponibilizada pela biblioteca, *Natural Language Tool Kit* (NLTK), mais a adição de termos a serem retirados, identificados como não relevantes. Além disso foram aplicados os métodos de *Stemming*, utilizando a NLTK, e *Lemmatization* e POS *tagging* por meio da biblioteca SpaCy.

Pode-se definir *Stemming* como o processo de transformação da palavra para o seu radical (elemento que contém o significado básico da palavra), ignorando a classe morfológica. Por exemplo, pedra, pedreiro e pedrada possuem o mesmo radical: ‘pedr’, e consequentemente sofreriam o mesmo resultado no processamento. Por outro lado, *Lemmatization* é um método mais sofisticado que busca transformar a palavra em sua forma primitiva ou, em caso de um verbo, sua forma infinitiva, de maneira a não alterar a sua classe morfológica. Enquanto, *Parts of Speech* (POS) *tagging* é um método de reconhecimento das partes que compõem o discurso, por meio da análise morfológica do conjunto de palavras (RAVI et al., 2015). Tanto o POS *tagging* quanto o *Lemmatization* do discurso foram elaborados utilizando um modelo de aprendizado de máquina pré-treinado por meio da biblioteca spaCy.

Por fim, para identificar as palavras mais relevantes, foi utilizado um método estatístico que evidencia a relevância da palavra no discurso. O *Term frequency-inverse document frequency* (TF-IDF) é uma medida estatística que avalia a importância de uma palavra em um documento, considerando a frequência dessa palavra no documento e a raridade que ela é apresentada em uma coleção de documentos. O TF-IDF é calculado multiplicando duas métricas: *Term Frequency* (TF) e *Inverse Document Frequency* (IDF) (GONÇALVES, 2016). O *Term Frequency* (Frequência do Termo) mede a frequência com que uma palavra aparece em um documento específico. Essa métrica assume que quanto mais vezes uma palavra aparece em um documento, mais importante ela é para aquele documento (1). Enquanto o *Inverse Document Frequency* (Frequência Inversa de Documento) mede importância de um termo comparando o

número de documentos que o termo aparece com o número total de documentos (2). O IDF é calculado para cada termo no conjunto de documentos, e dessa maneira quanto menos documentos contiverem um determinado termo, maior será o IDF atribuído a ele.

A fórmula do TF é definida como:

$$TF(termo, documento) = \frac{Número\ de\ vezes\ que\ o\ termo\ ocorre\ no\ documento}{(Número\ total\ de\ termos\ no\ documento)} \quad (1)$$

A fórmula do IDF é definida como:

$$IDF(termo) = log(\frac{Número\ total\ de\ documentos}{\text{Número de documentos que contêm o termo}}) \quad (2)$$

Dessa maneira, a TF-IDF é resultado pela multiplicação de ambas as métricas atribuindo uma importância a determinado termo e levando em consideração o respectivo documento e demais documentos do conjunto. Por exemplo, para comparar a relevância das palavras ‘Guedes’ e ‘Brasil’ em um documento, podemos usar o TF-IDF para calcular a importância de cada palavra. Suponha que o documento tenha 100 palavras. Além disso, suponha que a coleção de documentos tenha 4 documentos, dos quais 1 contém a palavra ‘Guedes’ e todos contém a palavra ‘Brasil’. Então, o TF-IDF de cada palavra é dado por:

$$TF - IDF(Guedes) = (\frac{5}{100}) \times log(\frac{4}{1}) = 0.03$$

$$TF - IDF(Brasil) = \left(\frac{20}{100}\right) \times log(\frac{4}{4}) = 0$$

Pode-se perceber que o TF-IDF do termo ‘Guedes’ é maior que o de ‘Brasil’, o que significa que ‘Guedes’ é mais relevante para o documento do que ‘Brasil’, mesmo possuindo uma frequência menor.

3.4 Desenvolvimento de análise

A fim de estudar a escolha lexical, optou-se pela comparação das palavras mais utilizadas, a comparação das palavras mais importantes, a correlação do uso de palavras entre os candidatos e a comparação entre dois candidatos por meio de um gráfico de dispersão da frequência relativa. Para isso determinou-se a aplicação a base de dados na sua forma primitiva (processo de *lemmatization*) e quanto a determinadas classes morfológicas.

A correlação de Pearson é uma importante medida estatística que mede a relação entre duas variáveis contínuas. O coeficiente de correlação assume valores em um intervalo entre -1 e +1, sendo que um valor próximo ou igual a 0 indica que não há associação entre duas variáveis, enquanto um valor próximo de 1 indica que existe uma relação linear positiva e -1 uma relação linear negativa. Para a análise foram selecionadas as frequências relativas da utilização das palavras por cada candidato.

Para a determinação da polaridade do discurso foi utilizado a abordagem por dicionário. Métodos baseados em dicionários constituem-se de um conjunto de palavras classificadas quanto a polaridade, podendo ser positiva, negativa ou neutra. Segundo, Grimmer e Stewart (2013), a aplicação da abordagem possui resultados satisfatórios quando a classificação das palavras no dicionário é coerente com o contexto do documento analisado, uma vez que, as palavras têm significados diferentes dado o domínio na qual está inserida. Como exemplo dado pelo autor, a palavra 'câncer' pode não ter uma polaridade negativa quando utilizada em um contexto técnico.

Além disso, devido a escolha dos dicionários utilizados na análise, em específico, o LlWC2015 *Brazilian Portuguese*, foram possíveis de inferência a frequência de menção por parte dos candidatos aos temas: família, morte, saúde e religião, e a utilização de palavras que remetem a: raiva, ansiedade e tristeza, visto a classificação dessas categorias nesse dicionário.

O *Linguistic Inquiry and Word Count* (LIWC, 2001) foi desenvolvido com o objetivo de fornecer um método efetivo para estudar as emoções e os componentes estruturais, cognitivos e procedurais presentes nos discursos verbais e escritos (PENNEBAKER; FRANCIS; BOOTH, 2001). A segunda versão, desenvolvida em 2001, trazia ainda a possibilidade de categorizar as falas do discurso nas dezenas de classificações estabelecidas por palavra.

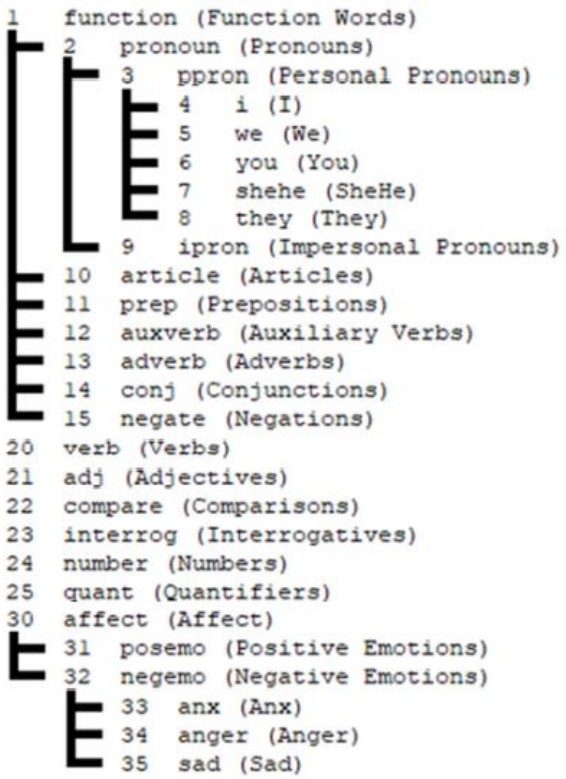
Ao longo dos anos, diferentes versões do LIWC foram desenvolvidas, valendo destaque a primeira versão em português*:* *Brazilian Portuguese* LIWC 2007 *Dictionary*, desenvolvido via tradução de forma colaborativa por 3 equipes: Unisinos, empresa Checon Pesquisa e NILC (DE CARVALHO, 2019). E a segunda versão, utilizada neste trabalho, LlWC2015 Brazilian Portuguese, desenvolvida por De Carvalho (2019), utilizando a criação do dicionário a partir das categorias selecionadas da versão em inglês, LIWC_2015en (PENNEBAKER et al., 2015), e depois as unindo em um único dicionário final.

De acordo com De Carvalho (2019), a versão do dicionário de 2015 em português apresentou melhorias significativas a versão de 2007, tanto em relação a classificação correta da polaridade,

comprovada estatisticamente com nível de confiança de 95%, quanto ao tempo necessário de processamento, dado a quantidade menor de palavras na lista, em uma redução de até 87% do tempo de processamento nos testes apresentados no estudo.

O dicionário em português, assim como sua versão em inglês possui classificação da lista de palavras em categorias e subcategorias, conforme a Figura 4, totalizando 35 classificações das palavras no dicionário. Para a análise de sentimentos, foram utilizadas as categorias *posemo* (emoções positivas) e *negemo* (emoções negativas) para classificação entre positivo (1) e negativo (-1) das palavras apresentas no discurso. Além disso, foi possível perceber a proporção de palavras que remetem a ansiedade, tristeza e raiva presentes no discurso – subclassificação das palavras negativas classificadas no dicionário, bem como a relação das palavras escolhidas com temas como: família, morte, saúde e religião classificadas no LIWC 2015.

Figura 4 - Divisão das categorias do dicionário LIWC 2015



Fonte: De Carvalho (2019)

Todo o código desenvolvido neste trabalho pode ser acessado por meio da plataforma de hospedagem de códigos de programação, GitHub, no repositório:
https://github.com/ProjetosPesquisa/PIBIC2023

3.5 Visualização de dados

Por último, a etapa final do processo de elaboração foi a apresentação dos dados em forma visual. De acordo com Rougier et al. (2014), a visualização de dados científica pode ser entendida como a interface gráfica que liga o público-alvo as informações extraídas dos dados e deve ser utilizado para passar de melhor maneira a mensagem ao público-alvo. No presente projeto foi escolhida a utilização da biblioteca Matplotlib para a representação gráfica.

4. Resultados

A partir da metodologia aplicada, os principais resultados da pesquisa podem ser classificados quanto a (i) escolha lexical, onde foram analisadas as palavras mais utilizadas por candidato, a correlação e as palavras mais importantes, e pela (ii) análise da polaridade dos discursos, onde foram analisadas a classificação da polaridade aplicada em níveis da sentença e do discurso.

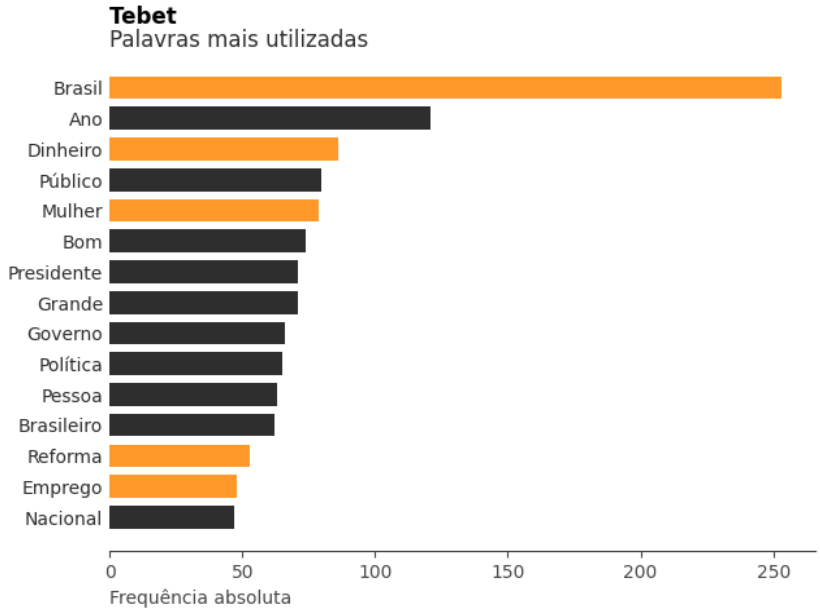
4.1 Análise da escolha lexical

A escolha lexical de um texto diz respeito ao conjunto de palavras que um locutor escolhe para transmitir a mensagem. Por meio da aplicação de métodos quantitativos para estudar esta escolha é possível aplicar diferentes formas de visualização e compreender a informação como demonstrado neste estudo.

Selecionando apenas adjetivos e substantivos, nas Figuras 5-8 são apresentadas as palavras mais faladas por candidato, com destaque aos termos mais significativos, identificados pela frequência em todos os discursos. Analisando todos os candidatos, a palavra ‘Brasil’ é o substantivo mais falado, com exceção a Lula, em que o substantivo mais falado é ‘país’ e em segundo lugar ‘Brasil’. Além disso, palavras como ‘ano’ e que remetem ao público ouvinte como ‘pessoa(s)’, ‘brasileiro(s)’ e ‘povo’ são amplamente utilizados por todos os candidatos, provavelmente como uma estratégia para evocar um senso de urgência no respectivo ano e incentivar os eleitores a tomar posição frente às propostas apresentadas.

Na Figura 5 é apresentada as palavras mais utilizadas pela candidata Tebet. Na figura, nota-se a forte utilização das palavra ‘mulher’, possivelmente relacionada a um apelo a esta parcela da população, pouco representada proporcionalmente no cenário político brasileiro (LIMA, 2022) e relacionada ao fato de ser a única candidata mulher dentre os quatro com maiores candidatos em intenções de votos e votos no primeiro turno. Além disso, é possível perceber a enunciação de termos que apontam para o cenário econômico brasileiro e a necessidade de mudança, como ‘dinheiro’, ‘reforma’ e ‘emprego’, sendo os dois últimos não utilizados com a mesma frequência por outros candidatos.

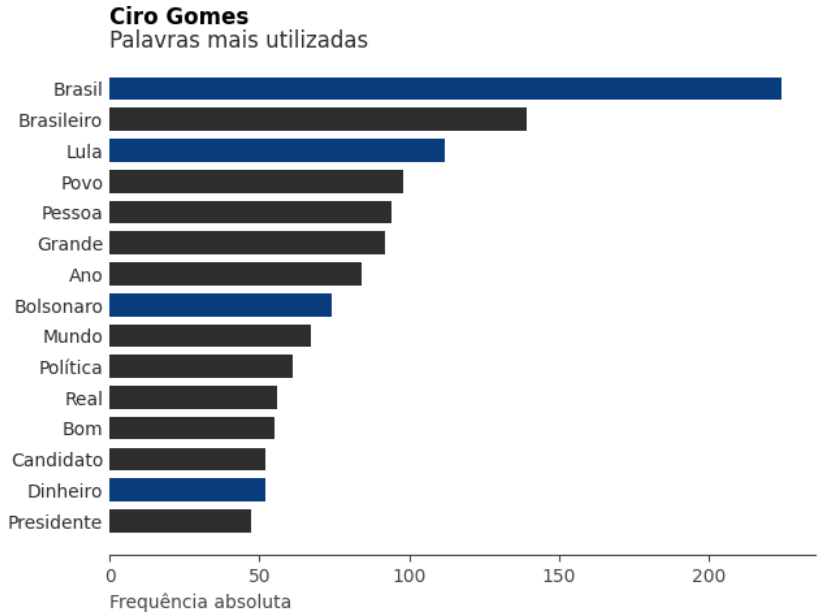
Figura 5 - 15 Palavras mais utilizadas Simone Tebet



Fonte: elaborado pelo autor (2023)

Na Figura 6, onde apontado as palavras mais faladas por Ciro Gomes, destaca-se a referência aos dois principais candidatos em intenções de voto ao longo da campanha, Lula, em primeiro lugar e Jair Bolsonaro em seguida, evidenciando a tentativa de se posicionar como oposição a ambos os candidatos, como uma terceira via (GRANJEIA, 2022). Outro ponto importante foi a utilização do termo ‘dinheiro’ por parte do candidato, mais uma vez apontando para a preocupação do cenário econômico.

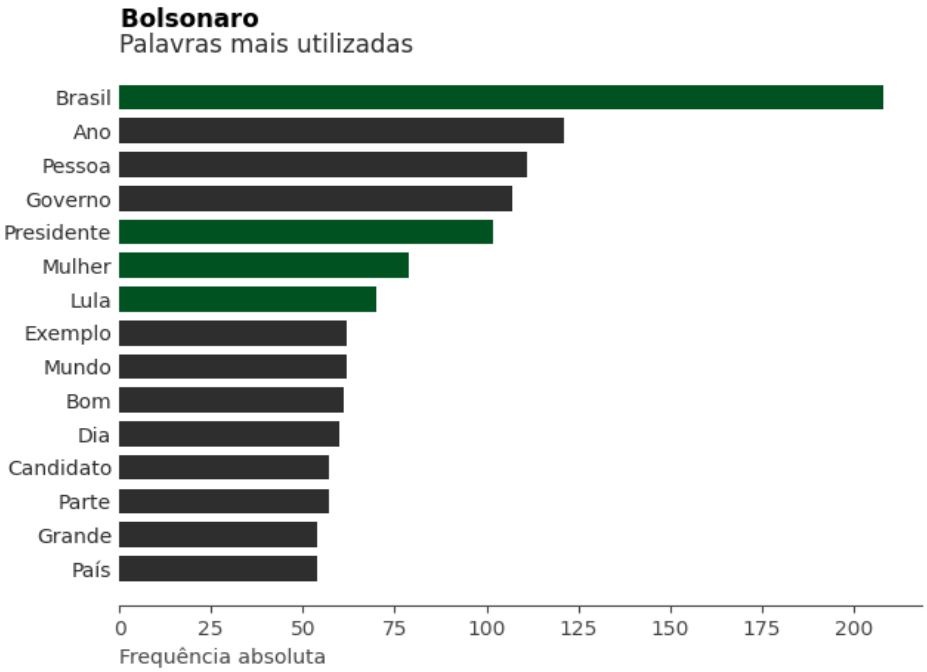
Figura 6 - 15 Palavras mais utilizadas Ciro Gomes



Fonte: elaborado pelo autor (2023)

A Figura 7 apresenta as palavras mais utilizadas pelo ex-presidente Jair Bolsonaro, de modo que, existe uma maior utilização dos termos ‘mulher’, uma vez, constituinte de um grupo sensível de votos ao candidato e onde possuía maior número de rejeição (DA ROCHA, 2022), e ‘Lula’, candidato com maior intenção de votos e principal adversário político.

Figura 7 - 15 Palavras mais utilizadas Jair Bolsonaro

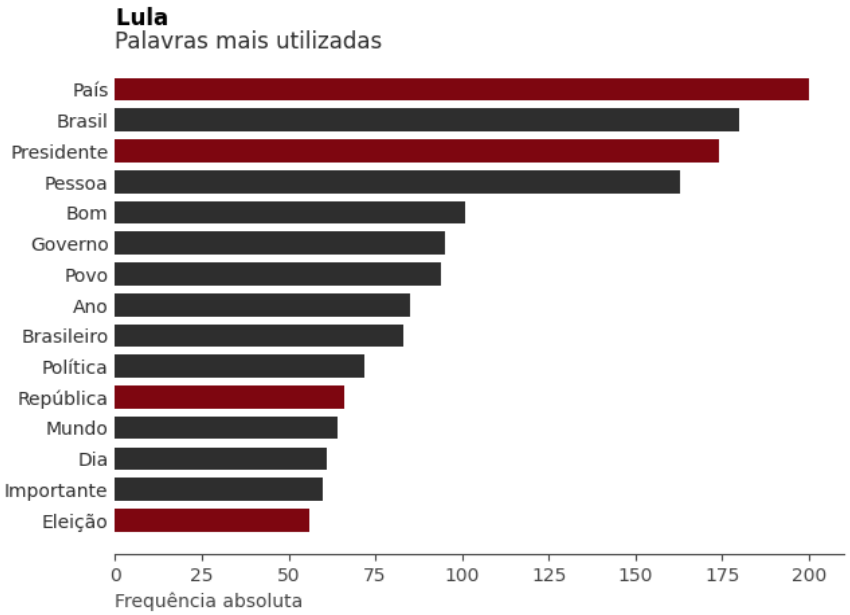


Fonte: elaborado pelo autor (2023)

Na Figura 8, vemos as 15 palavras mais citadas pelo candidato Lula. Na figura, é possível notar as duas palavras mais utilizadas: ‘país’ e ‘Brasil’. Destaca-se o emprego da palavra ‘presidente’, sendo o candidato Lula o candidato que mais utilizou este termo. Pode-se também perceber que os termos ‘Bolsonaro’, ‘Tebet’ e ‘Ciro’ não são uma das 15 palavras mais utilizadas pelo candidato, ao contrário do que visto nos discursos dos outros candidatos Bolsonaro e Ciro Gomes, que citam frequentemente o candidato Lula.

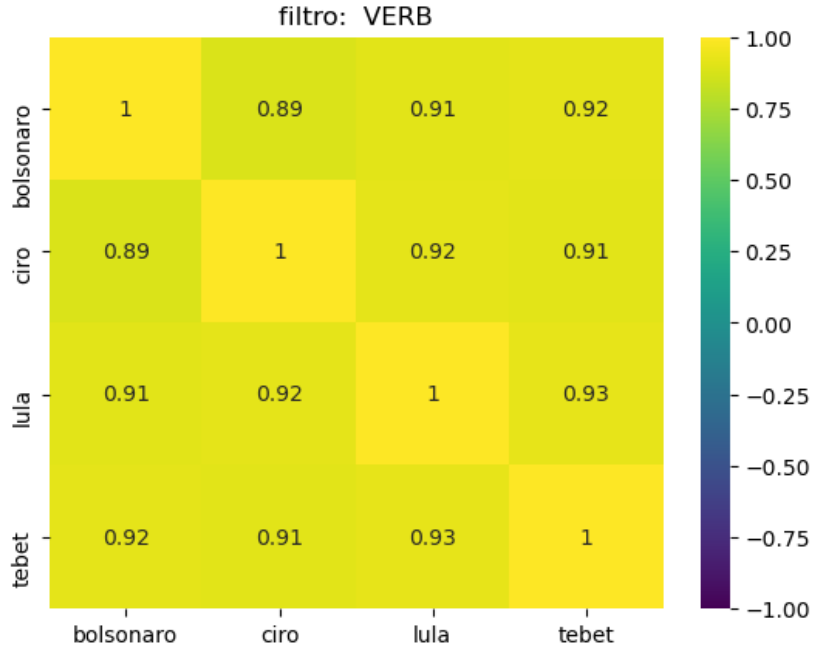
Analisando a correlação da escolha de verbos, não é possível inferir diferenças significativas entre a escolha de palavras dos candidatos, dado os verbos utilizados como ‘fazer’, ‘poder’ e ‘falar’. Pela análise, obteve-se uma correlação do uso de verbos maior que 0.9 para todos os candidatos (Figura 9). Isto que pode ser explicado pela escolha de verbos ser similar no contexto de entrevistas.

Figura 8 – 15 Palavras mais utilizadas Lula



Fonte: elaborado pelo autor (2023)

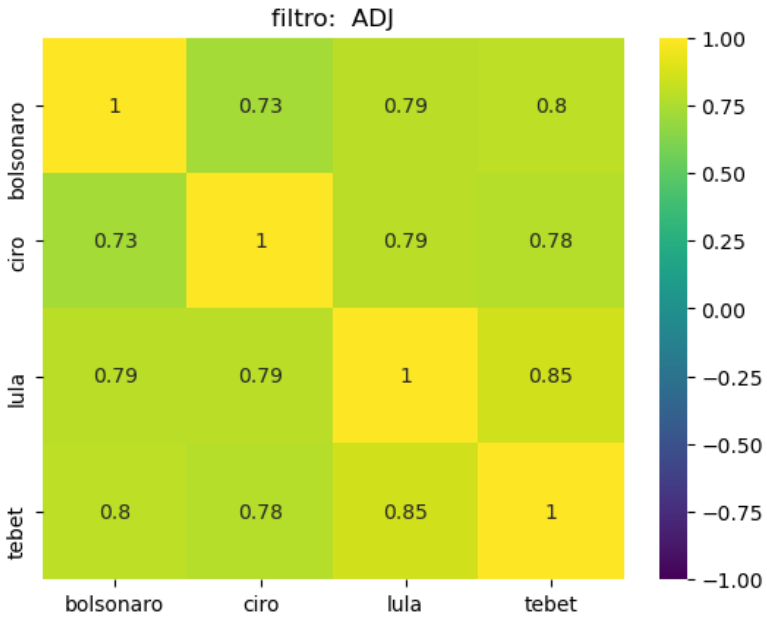
Figura 9 – Correlação da frequência de verbos



Fonte: elaborado pelo autor (2023)

Na Figura 10, foi analisada a correlação do uso de adjetivos pelos candidatos, como por exemplo: ‘grande’, ‘importante’, ‘nacional’ e ‘social’. Pode-se notar que os candidatos Lula e Tebet apresentaram a correlação mais alta, enquanto Ciro e Bolsonaro a mais baixa entre si, o que representa uma maior similaridade e diferença, respectivamente, entre a escolha de adjetivos, mesmo que pouco expressiva.

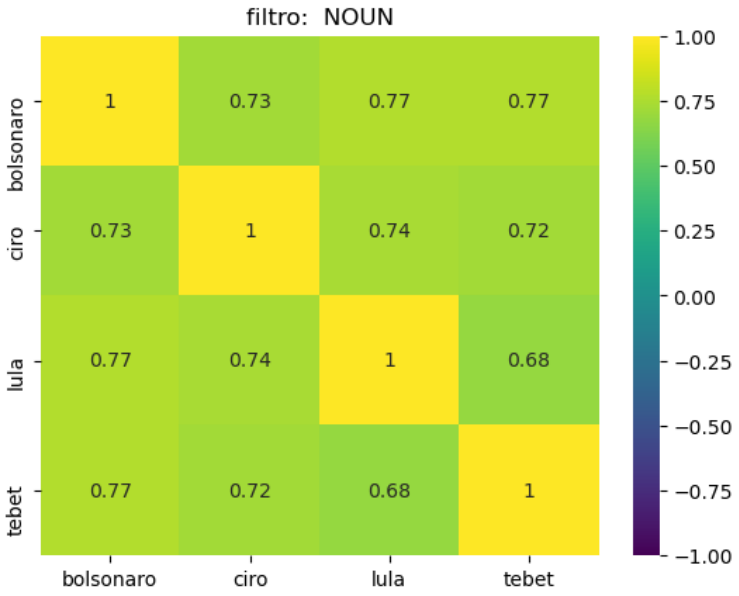
Figura 10 – Correlação da frequência de adjetivos



Fonte: elaborado pelo autor (2023)

Analisou-se também a correlação da escolha dos substantivos (Figura 11). Apesar de todos os candidatos apresentarem correlações altas dado a frequência e escolha, houve maior disparidade neste caso entre Lula e Tebet e semelhança entre Bolsonaro e Tebet e Bolsonaro e Lula.

Figura 11 – Correlação da frequência de substantivos



Fonte: elaborado pelo autor (2023)

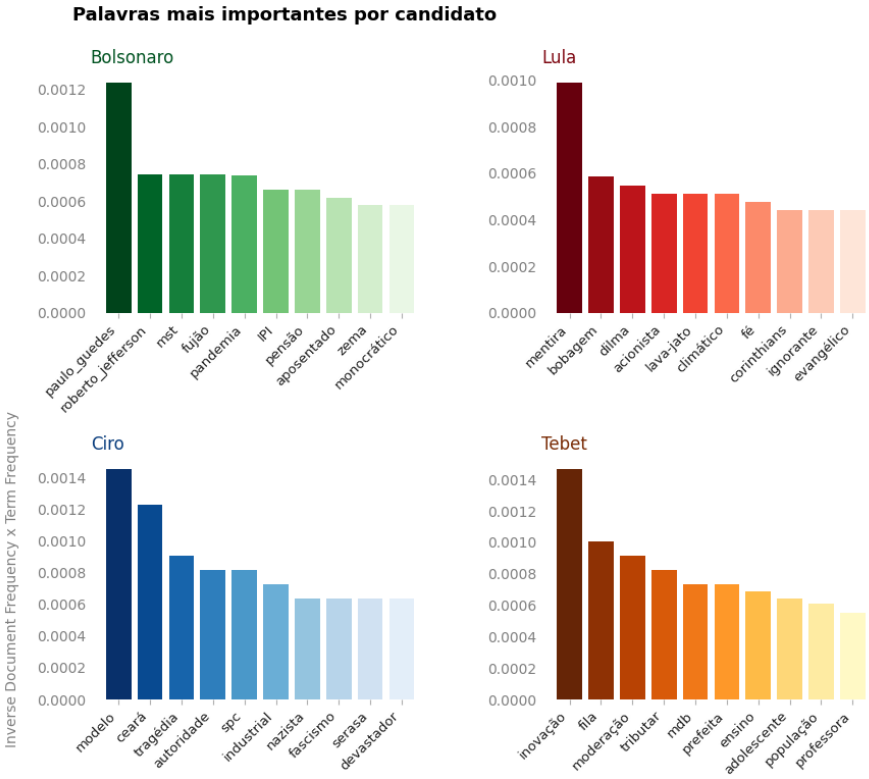
Na figura 12 são representadas as palavras mais importantes por candidato, selecionando apenas adjetivos e substantivos. Nesse aspecto foram consideradas as quatro entrevistas por candidato

como um único documento, enquanto o corpus (conjunto de documentos) o conjunto dos discursos de cada presidenciável. A partir disso, é possível notar que as palavras mais importantes dado o TF-IDF são diferentes para todos os candidatos, em contraposição da análise das palavras mais frequentes, que apresenta alta similaridade.

Na fala do ex-presidente Bolsonaro, destaca-se a presença das palavras ‘pandemia’, motivo que gerou diversas crises e insatisfações em seu mandado (CNN BRASIL, 2021), ‘Paulo Guedes’, Ministro do superministério da Economia em seu mandato, ‘MST’ Movimento dos Trabalhadores Rurais Sem Terra e ‘Roberto Jefferson’, ex-deputado do PTB. No discurso do atual presidente Lula, nota-se a presença dos termos ‘mentira’, ‘Dilma’, ‘Lava-Jato’ dentre as 10 palavras com maior importância. Analisando o discurso de Ciro Gomes, distingue-se os termos ‘Ceará’, estado que exerceu cargo de governador, ‘SPC’, o serviço de proteção ao crédito, mais uma vez relacionado ao aspecto econômico e os termos ‘nazismo’ e ‘fascismo’. Por último, na fala de Simone Tebet, destaca-se os termos ‘inovação’, ‘moderação’, ‘ensino’ e ‘prefeita’, cargo que atuou no município de Três Lagoas – MS em 2004.

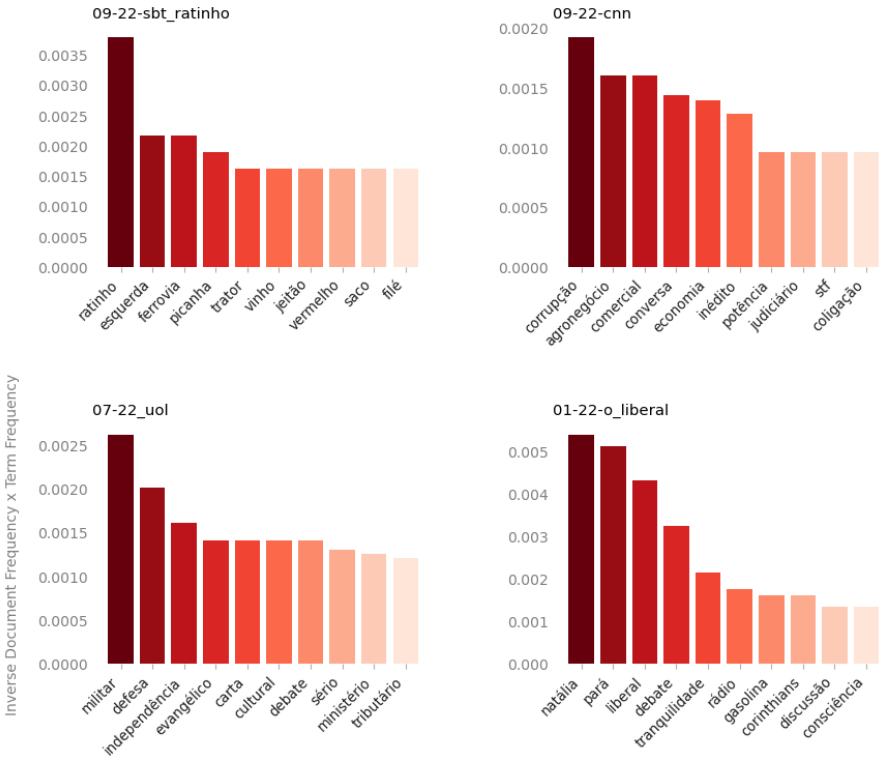
Outra aplicação abordada na pesquisa, foi a consideração das entrevistas de cada candidato como documentos separados e o conjunto das 4 entrevistas, como o corpus estudado. A Figura 13 apresenta as palavras mais importantes faladas pelo candidato Lula para cada entrevista. Na figura é possível perceber a disparidade de importância no contexto do meio de comunicação. Dessa maneira enquanto na CNN, existe uma maior ênfase em temas como corrupção, agronegócio e economia, no programa do Ratinho existe um apelo a temas como esquerda, picanha e vinho, explicados pelo segundo estar num contexto mais popular, exibido em canal de televisão aberto. Vale ressaltar que a escala não representa uma base quantitativa precisa, regra também aplicada no caso anterior, e não possibilita uma comparação direta entre palavras de documentos diferentes, dado a desproporção de termos falados em cada entrevista.

Figura 12 – 10 palavras mais importantes por candidato, com base no TF-IDF



Fonte: elaborado pelo autor (2023)

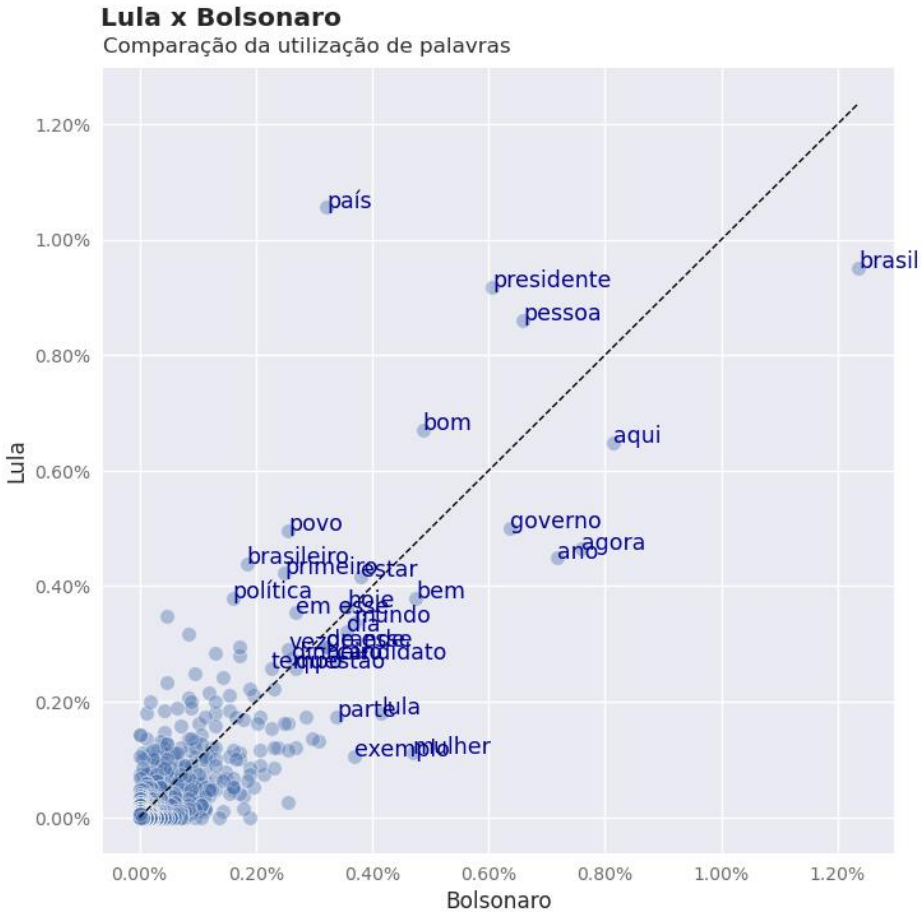
Figura 13 – Palavras mais importantes das entrevistas do candidato Lula



Fonte: elaborado pelo autor (2023)

A frequência das palavras utilizadas por cada candidato, pode ser expressa também, de maneira comparativa por um gráfico de dispersão, cada ponto correspondendo a uma palavra e a posição da palavra gráfico condiz com a frequência de cada candidato em um eixo x e y. De maneira que quanto mais utilizada por apenas um candidato, maior a proximidade de um dos eixos e mais distante da diagonal, e quanto mais utilizada por ambos, maior a proximidade com a diagonal. Na Figura 14, é apresentado a comparação entre Lula e Bolsonaro, analisando apenas substantivos adjetivos e advérbios. Pode-se notar que a palavra 'Brasil', é falada muito por ambos, porém tende a ser mais falada relativamente por Bolsonaro, enquanto 'país' é mais falada por Lula, conforme indicado nas análises anteriores. Além disso, vale destaque as palavras que indicam tempo utilizadas por ambos, como 'hoje', 'ano', 'agora' e 'dia' e todas mais utilizadas por Bolsonaro.

Figura 14 – Comparação da utilização de palavras: Lula e Bolsonaro

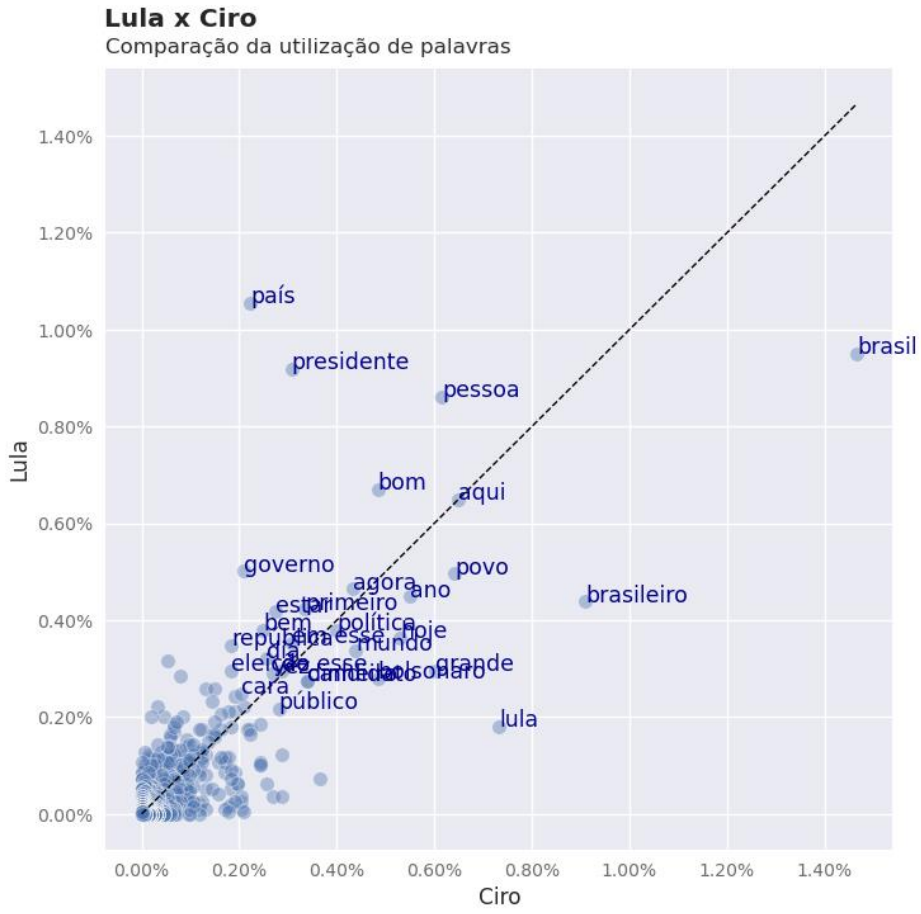


Fonte: elaborado pelo autor (2023)

A comparação entre Lula e Ciro, mostrado na Figura 15, identifica uma maior dispersão das palavras, levando em consideração a figura anterior. O nome 'Lula' é mais falado por Ciro do

que pelo próprio candidato Lula. Chama-se a atenção também que Ciro, refere-se mais a população como ‘brasileiro(s)’ e Lula utiliza mais o termo ‘pessoa(s)’.

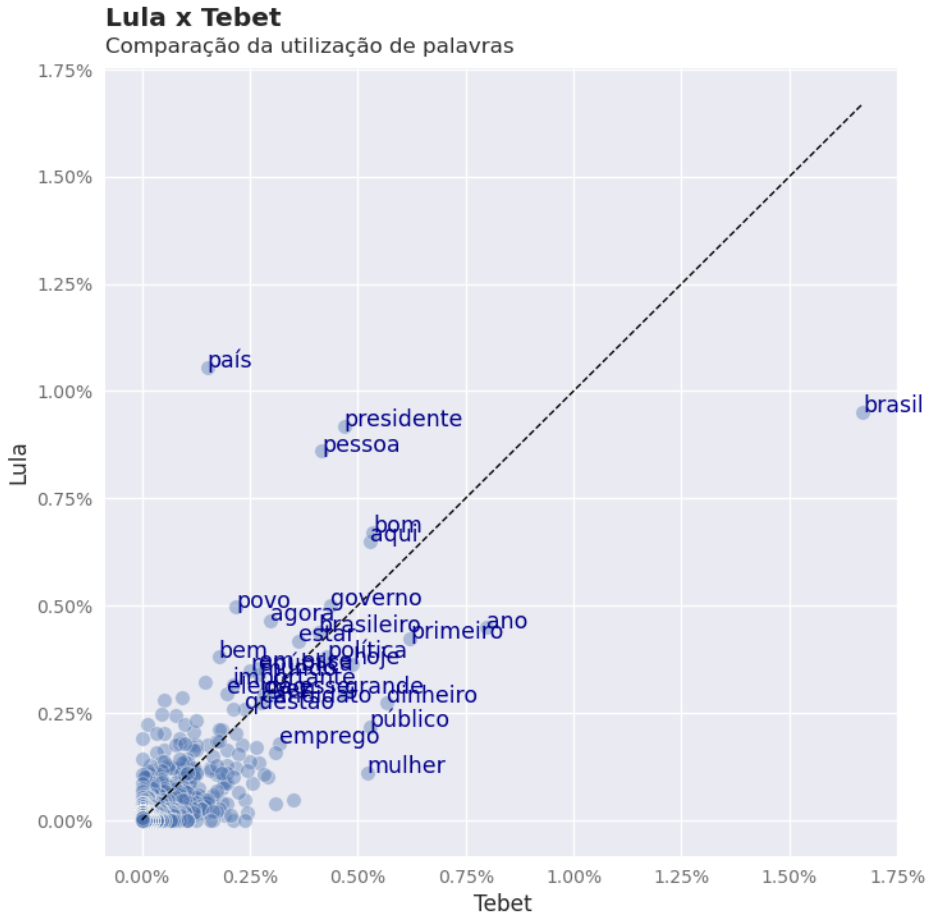
Figura 15 – Comparação da utilização de palavras: Lula e Ciro



Fonte: elaborado pelo autor (2023)

Na figura 16, compara-se a frequência das palavras entre Tebet e Lula. É possível notar uma menor dispersão das palavras entre os candidatos, ou seja, uma similaridade maior da utilização quando comparado Lula a Bolsonaro e Ciro. Além disso o termo ‘presidente’ é mais utilizado por Lula, enquanto ‘dinheiro’, ‘emprego’ e ‘mulher’ mais utilizado por Tebet.

Figura 16 – Comparação da utilização de palavras: Lula e Tebet

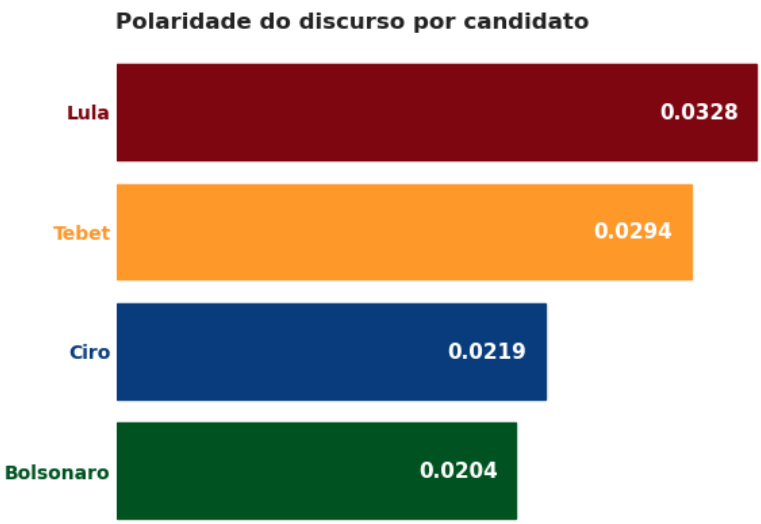


Fonte: elaborado pelo autor (2023)

4.2 Análise da polaridade dos discursos

A Figura 17, apresenta a polaridade dos discursos dado todas as entrevistas de cada candidato. Nota-se que o candidato e primeiro lugar no segundo turno, Luiz Inácio Lula, apresentou maior polaridade positiva e o ex-presidente Jair Bolsonaro a menor polaridade positiva, de maneira que os demais tiveram polaridade intermediaria entre os dois, Simone Tebet e Ciro Gomes, respectivamente.

Figura 17– Polaridade dos discursos por candidato

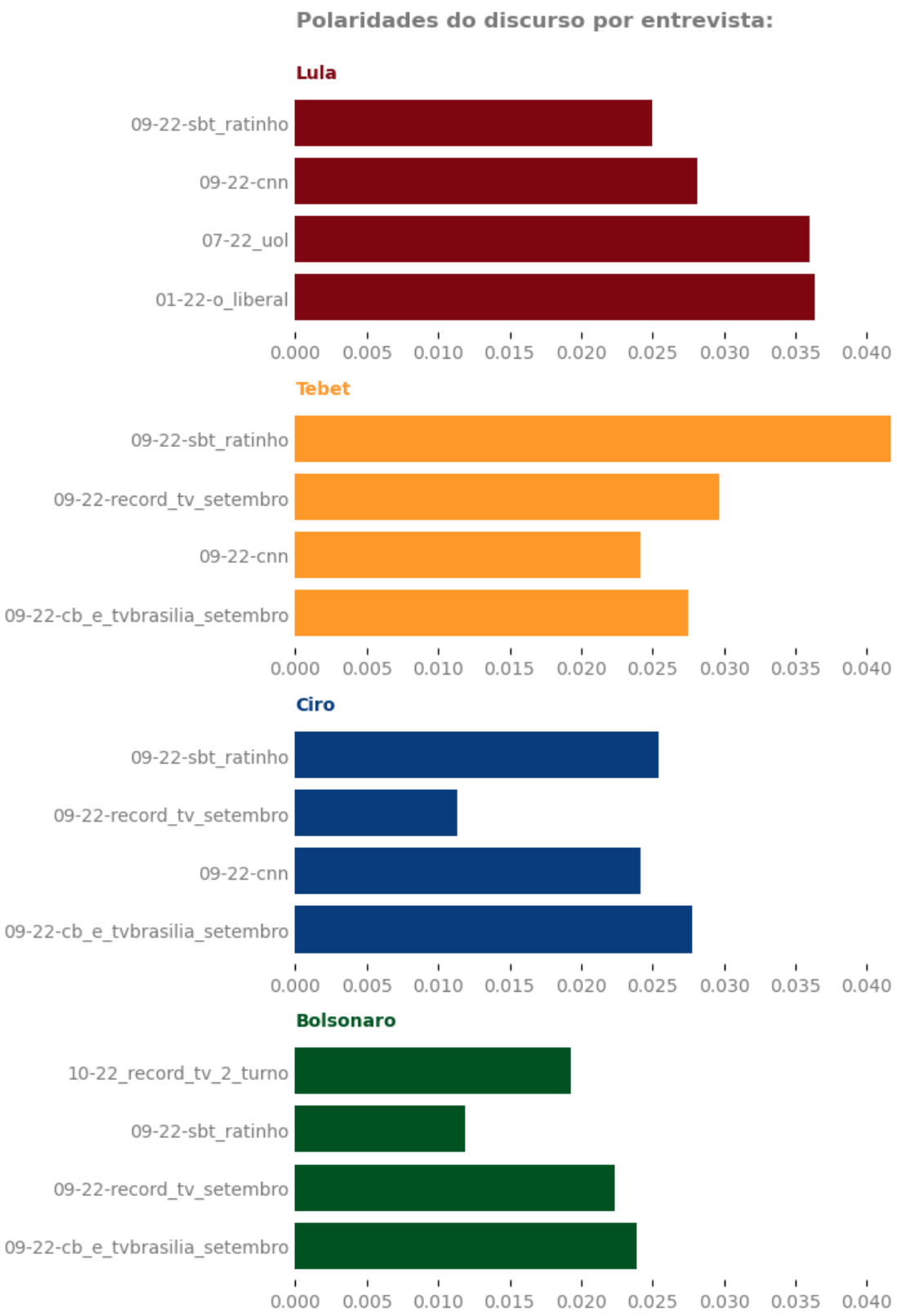


Fonte: elaborado pelo autor (2023)

Destaca-se ainda a comparação da polaridade nos discursos de cada candidato, revelando variações distintas com base nas entrevistas. Na Figura 18 é possível notar que quando comparado todas as entrevistas, o maior grau de polaridade é visto na entrevista com Simone Tebet realizadas no Programa do Ratinho, em setembro de 2022, enquanto o menor grau de polaridade, apresentado por Jair Bolsonaro no mesmo meio de comunicação. Quanto ao candidato Lula, apresenta-se uma menor diferença das polaridades entre os discursos em comparação aos demais presidenciáveis e nota-se por parte do candidato Ciro Gomes uma menor polaridade na sabatina realizada pelo Jornal da Record em setembro de 2022, quando comparado apenas as suas entrevistas.

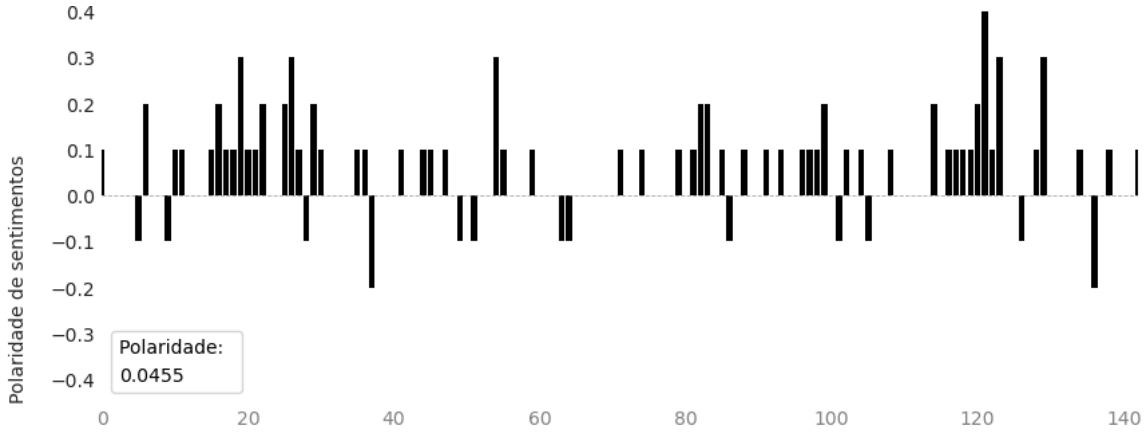
Além do exposto, é possível entender qual a variação da polaridade por meio da classificação da polaridade ao longo de trechos do discurso e identificar em quais partes houve uma alteração acentuada do sentimento expresso entre positivo e negativo. A título de exemplificação, na Figura 19, é mostrado o primeiro discurso do então presidente eleito, Lula, onde existe a predominância de termos que remetem a emoções positivas e momentos no discurso em que existe uma diminuição notável do grau de polaridade, resultando em uma polaridade maior que a média vista em cada candidato.

Figura 18 – Polaridade dos discursos por entrevista



Fonte: elaborado pelo autor (2023)

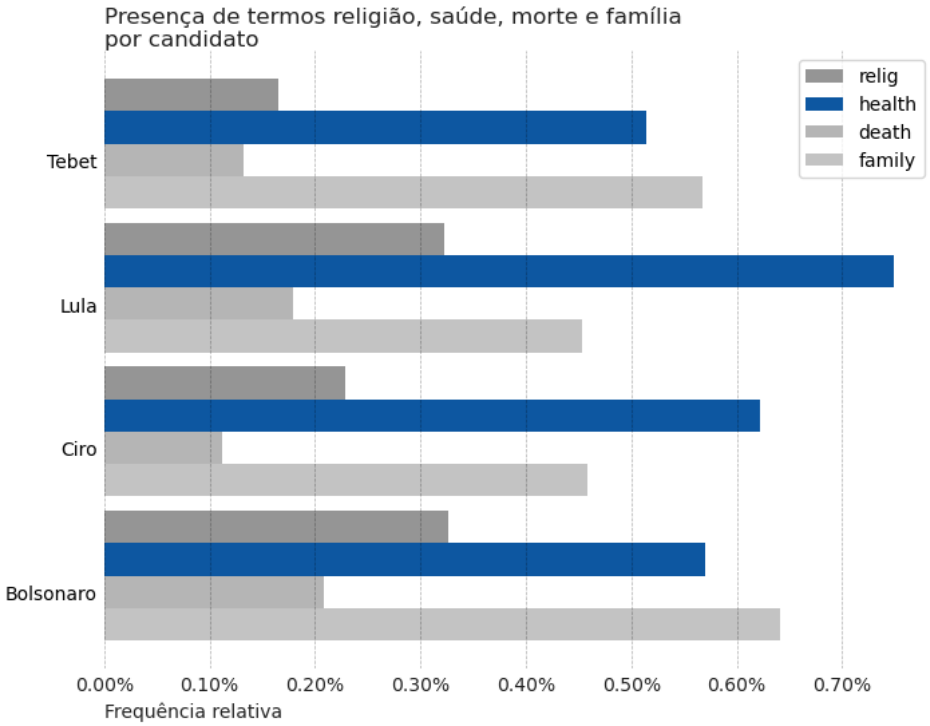
Figura 19 - Classificação da polaridade ao longo do primeiro discurso após a vitória do candidato Lula



Fonte: elaborado pelo autor (2023)

A Análise de Sentimentos possibilita também extrair opiniões e sentimentos a partir de documentos. Na Figura 20, foi observado, uma diferença entre a utilização de palavras que indicam religião, saúde, morte e família. Na figura em questão, é destacado o termo saúde e nota-se o emprego de maneira expressiva pelo vencedor da eleição, Luiz Inácio Lula e utilizado com menor frequência por Simone Tebet.

Figura 20– Análise de palavras que remetem a saúde por candidato

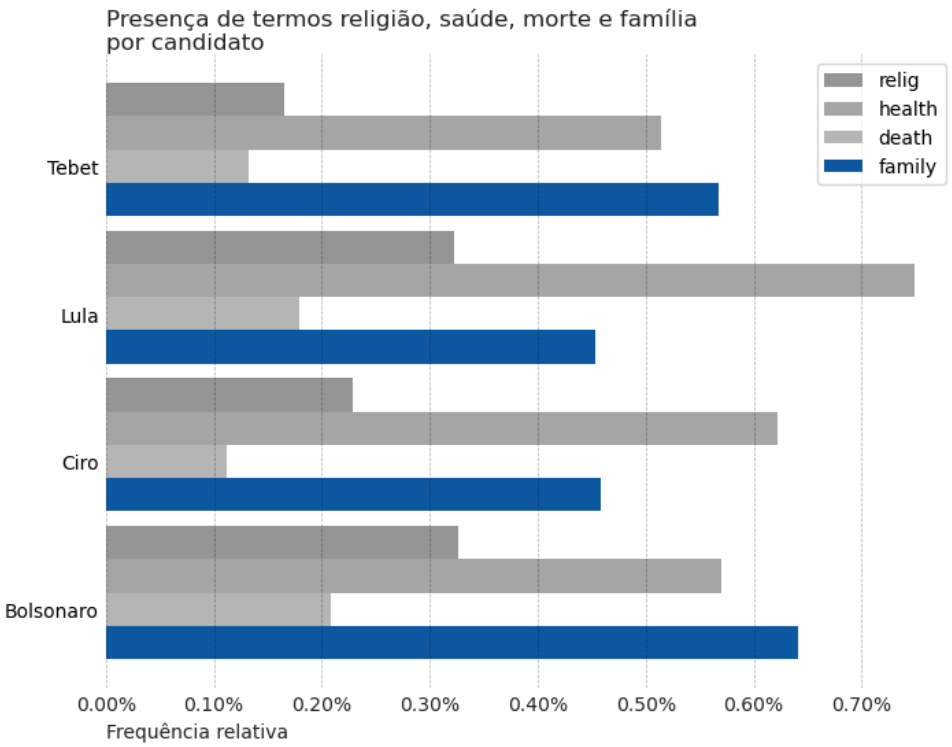


Fonte: elaborado pelo autor (2023)

Quanto ao termo família, Jair Bolsonaro, seguido de Simone Tebet, foram os que utilizaram mais palavras que demonstram essa opinião (Figura 21). Na figura, ainda é perceptível que

Bolsonaro é o candidato que mais apresenta termos que remetem opiniões de morte, em seguida de Lula. Essa mesma ordem se aplica na utilização de palavras que indicam religião, utilizadas com maior intensidade pelos candidatos mencionados e em menor por Tebet e Ciro, com menor número de votos, respectivamente no primeiro turno.

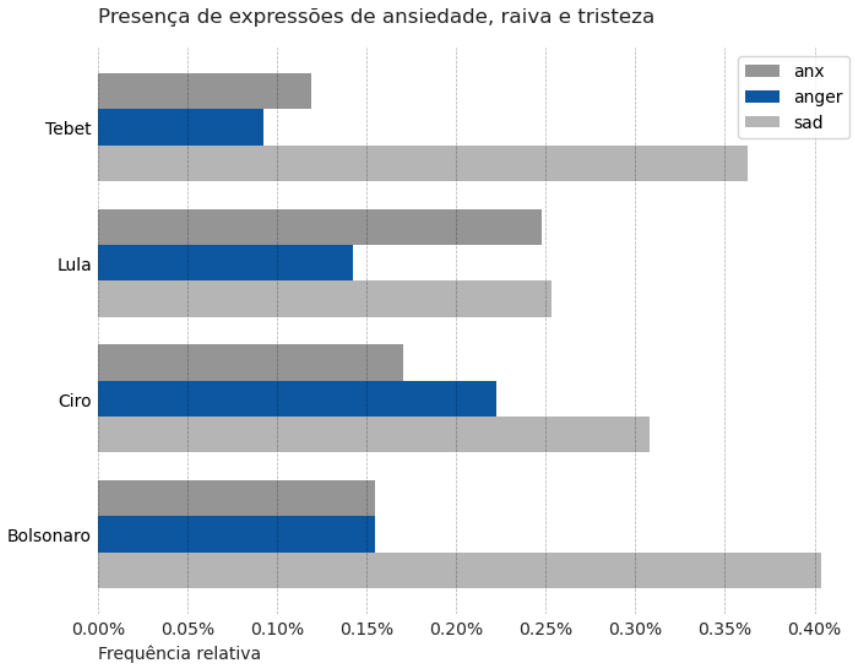
Figura 21 – Análise de palavras que remetem a família por candidato



Fonte: elaborado pelo autor (2023)

A Figura 22 apresenta o comparativo da frequência de termos que demonstram sentimentos de ansiedade, raiva e tristeza, dando destaque a comparação de raiva. Ciro Gomes foi o candidato que mais demonstrou raiva nas entrevistas e Tebet a que menos demonstrou.

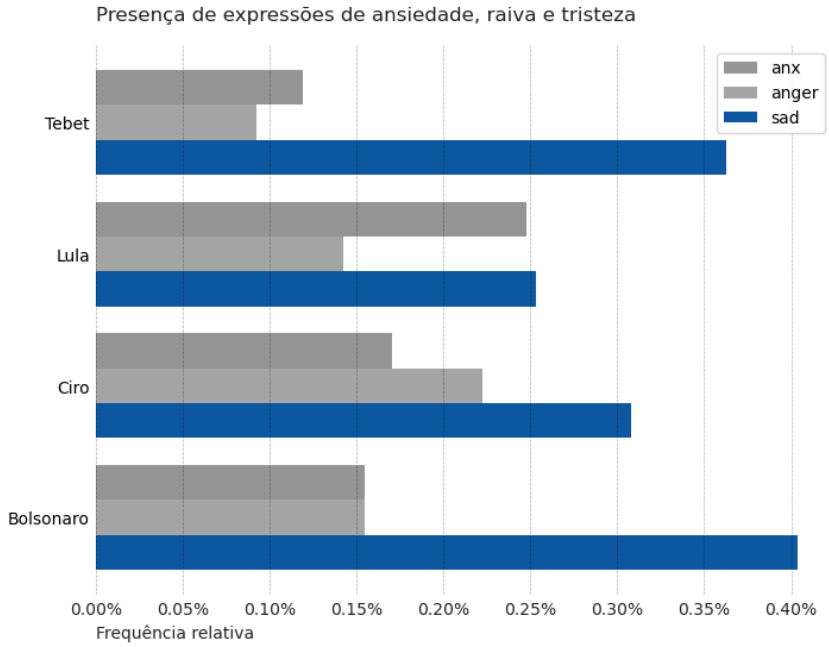
Figura 22 – Análises de expressões que remetem a ansiedade por candidato



Fonte: elaborado pelo autor (2023)

Na Figura 23, indica-se que em relação a tristeza, Bolsonaro foi o que mais demonstrou em seus discurso e Lula, o que menos demonstrou. Ainda assim, pode-se notar que em todos os candidatos, o sentimento de tristeza presente nos discursos foi mais alto quando comparado a ansiedade e raiva e Lula apesar de apresentar menor sentimento de tristeza, foi o que mais apresentou palavras que indicam o sentimento de ansiedade.

Figura 23 – Análises de expressões que remetem a tristeza por candidato



Fonte: elaborado pelo autor (2023)

5. Discussão

Analisando as palavras mais faladas, é possível reconhecer contrastes e semelhanças na forma pretendida de gerar convencimento e direcionar a vitória da eleição presidencial. A partir das Figuras 5-8, torna-se evidente a semelhança da escolha lexical em que 5 das 15 palavras mais faladas são comuns entre todos os candidatos, sendo elas 'Brasil', 'presidente', 'ano', 'bom' e 'pessoa' e uma minoria é mais falada exclusivamente, de tal forma que algumas das palavras mais utilizadas representam temas repetitivamente debatidos por esses candidatos durante as entrevistas analisadas na campanha, como questões quanto à economia e a forma de apelo ao eleitorado.

Quando comparado as palavras mais importantes dado o TF-IDF (Figura 12), percebe-se uma diferença dos termos destacados entre os candidatos e diferença com as respectivas palavras mais faladas. Isso pode ser fundamentado pelo fato de que os termos mais falados (Figuras 5-8), em razão de serem comuns a todos os discursos, não possuem tanta importância dado a medida estatística utilizada, que além de contar a frequência da palavra, leva em consideração a exclusividade nos discursos dos candidatos. Dessa forma, evidenciado palavras utilizadas com menor frequência, porém mais exclusivas ao discurso do candidato, como 'Paulo Guedes' para Bolsonaro, 'Dilma' para Lula, 'Ceará' para Ciro, e 'moderação' para Tebet (Figura 12).

A correlação de palavras classificadas em diferentes classes morfológicas, por meio do *POS tagging*, possibilita identificar semelhança em maior grau entre verbos, adjetivos e substantivos escolhidos pelos candidatos, respectivamente (Figuras 9-11). De maneira similar, a análise da escolha lexical, também pode ser representada pela presença do termo quando comparado entre dois candidatos (Figuras 14-16). Nessa análise, apresentada por um gráfico de dispersão, observa-se que além de entender quanto um termo é utilizado, é possível identificar em qual discurso está mais presente. Dessa forma, quando comparada as escolhas lexicais de Lula com os demais candidatos, existem diferenças na frequência, mas não de modo acentuado, dado a alta correlação de verbos, substantivos e adjetivos nos discursos.

A diferença entre os candidatos, também pode ser percebida pelo estudo da polaridade de seus discursos. Quando comparado o ex-presidente Jair Bolsonaro com o atual presidente Lula, verifica-se uma polaridade 37% menor do ex-presidente ao seu atual, dado os discursos analisados (Figura 15). Essa polaridade menor, está relacionada, parcialmente, pelo maior emprego de Bolsonaro de palavras que remetem a tristeza, quando comparado a todos (Figura 23). Ainda assim, existe uma variação por cada candidato quanto ao emprego de termos que remetem a família, religião e saúde. Enquanto Lula e Ciro utilizaram mais palavras relacionadas

à saúde, Bolsonaro e Tebet apresentaram mais palavras vinculadas à família e Bolsonaro e Lula, uma forte ênfase na utilização de termos ligados à religião. (Figura 20-21).

Ademais, é importante se atentar aos erros inerentes ao modelo de classificação do conjunto de documentos estudados, uma vez que a complexidade da linguagem torna o erro algo inerente à aplicação de métodos quantitativos para a análise de textos (GRIMMER; STEWART, 2013). Elucidando o caso em estudo, a abordagem por dicionário possui restrições quanto ao domínio do conhecimento em que o dicionário, contendo um conjunto de palavras classificadas, foi originalmente criado. Durante a realização da pesquisa, não foram encontrados dicionários em português do Brasil específicos para o vocabulário utilizado no contexto político. Dessa maneira, a solução surge a partir da aplicação do LlWC2015 Brazilian Portuguese (DE CARVALHO, 2019), adaptado da versão em inglês desenvolvida com propósito de estudar de maneira eficiente os componentes emocionais e cognitivos presentes nos discursos (PENNEBAKER et al., 2015). Outro ponto importante, devido a abordagem, é a dificuldade em determinar a polaridade de uma sentença por não considerar o contexto em que determinada palavra que demonstra polaridade está inserida. Mesmo assim, a abordagem possui vantagens significativas quanto a outras, por conta da simplicidade de entendimento e facilidade de aplicação em diferentes problemas (GRIMMER; STEWART, 2013). Segundo os autores, essa contrariedade não pode ser solucionada pela adoção de uma metodologia em específico, devido a não existir um modelo quantitativo melhor geral, dado a diferentes perguntas de pesquisa e necessidades de aplicações.

6. Conclusão

Dado o objetivo geral da pesquisa, pode-se concluir que a análise de sentimentos é uma área de estudo cada vez mais importante, em que a todo momento dados textuais são gerados em grande quantidade na WEB. Por meio dela é possível a identificação de aspectos da emoção, opinião e sentimentos de documentos, que poderiam apenas antes ser entendidos por análises qualitativas. A sua aplicação na área das ciências políticas, em mais específico nos discursos políticos, mostra-se promissor. Mesmo que existam poucos estudos relacionando as duas áreas de conhecimento no Brasil, com a evolução de métodos e aplicações mais simples e mais assertivas, existe um grande potencial na área acadêmica.

Os discursos políticos analisados, mostraram que não só existem semelhanças entre a escolha lexical por cada candidato de ideologias diferentes, mas diferenças que caracterizam o discurso, dado imposição de visão de realidade como objetivo a ser alcançado (PINTO, 2009). Muitas delas, são percebidas, por meio da importância das palavras atribuídas comparativamente e de

maneira mais expressiva, essas diferenças se materializam por meio da análise de sentimentos, como visto em diferentes polaridades entre candidatos e entre discursos. Outra possibilidade de aplicação abordada foi a categorização de palavras de acordo com temas que elas representam, como família, saúde e religião. Evidenciando divergências de temas debatidos por candidato de forma automatizada e possível de replicação em outros estudos e documentos de contexto similar.

Como mencionado anteriormente, uma das limitações do estudo, refere-se ao erro inerente ao modelo de extração de informações a partir de dados qualitativos, em razão da complexidade da linguagem. Em vista disso a etapa validação, tem importância significativa para minimizar e evitar tais erros, conforme aplicada no estudo (GRIMMER; STEWART, 2013). Uma ressalva adicional diz respeito aos dados coletados dos discursos de cada candidato. Como a participação de entrevistas depende da disposição e escolha do candidato, dificilmente houve entrevistas do mesmo veículo de comunicação comum aos quatro presidenciáveis. Por esse motivo, selecionou-se as entrevistas mais pertinentes, dado critérios objetivos estabelecidos, como de veículos relevantes, preferencialmente em estilo presencial e com mais de 3 candidatos participantes, realizadas no ano das eleições.

Dessa maneira, com este trabalho, espera-se contribuir com um entendimento mais aprofundado sobre o tema da análise de sentimentos e sua aplicação nos discursos de figuras políticas, em um cenário que marca a democracia do país e a disputa pelo poder, além de colaborar em futuras pesquisas com elaboração do software e disponibilização em repositório aberto. Como futuros estudos e ampliação do tema abordado na pesquisa, sugere-se o mesmo objeto de estudo – o discurso político - com a utilização de outros métodos para a análise de sentimentos e mineração de textos, tendo em vista a crescente popularização de ferramentas de processamento de linguagem natural, surgimento de novas tecnologias e avanços na área de Machine Learning e modelos de geração de textos.

Por fim, os resultados obtidos neste trabalho podem ser expandidos para outros campos, por exemplo, a administração. De maneira acessível, uma empresa poderia utilizar as técnicas de mineração de texto e análise de sentimentos para entender a partir de uma base de dados, como os seus consumidores avaliam a empresas e os serviços ou produtos ofertado. Nesse aspecto é possível monitorar o sentimento e opiniões do público em relação à marca em plataformas de mídias sociais e em plataformas de avaliação de produtos, identificando padrões e tendencias dificilmente perceptíveis de forma manual, e de mesmo modo, extrair *feedbacks*, insatisfações

e problemas específicos nos canais de atendimento ao cliente, além de diversas outras aplicações em que é necessário o processamento de dados em forma textual.

7. Referências

AVANCO, Lucas Vinicius; NUNES, Maria das Graças Volpe. Lexicon-based sentiment analysis for reviews of products in Brazilian Portuguese. In: **2014 Brazilian Conference on Intelligent Systems**. IEEE, 2014. p. 277-281.

BENEVENUTO, Fabrício; RIBEIRO, Filipe; ARAÚJO, Matheus. Métodos para análise de sentimentos em mídias sociais. **Sociedade Brasileira de Computação**, 2015.

CHOWDHURY, G. G. Natural language processing. **Annual Review of Information Science and Technology**, v. 37, p. 51-89, 2003.

DATAFOLHA: rejeição a Bolsonaro em gestão da pandemia bate recorde e chega a 54%. **CNN Brasil**, 17 mar. 2021. Disponível em: https://www.cnnbrasil.com.br/politica/datafolha-rejeicao-a-bolsonaro-em-gestao-da-pandemia-bate-recorde-e-chega-a-54/. Acesso em: 23 jun. 2023.

DA ROCHA, Guilherme. Na TV, Bolsonaro foca no voto feminino com Michelle em destaque e Lula prega contra ódio político. **Valor Econômico**, São Paulo, 13 set. 2022. Disponível em: https://valor.globo.com/politica/eleicoes-2022/noticia/2022/09/13/na-tv-bolsonaro-foca-no-voto-feminino-com-michelle-em-destaque-e-lula-prega-contra-odio-politico.ghtml. Acesso em: 25 maio 2023.

DE CARVALHO, Flavio Matias Damasceno. **DESENVOLVIMENTO DO DICIONARIO LIWC 2015 EM PORTUGUÊS DO BRASIL**. 2019. Tese de Doutorado. Centro Federal de Educação Tecnológica Celso Suckow da Fonseca.

DELAVALD, Gabriel Sehn. **Uma análise de dados das reações à crise política brasileira no twitter**. 2018. Monografia (Ciência da computação) - Universidade Federal do Rio Grande do Sul, Porto Alegre, 2018

DIFERENÇA de votos entre Lula e Bolsonaro é a menor da história. **Correio do Povo**, 30 out. 2022. Disponível em: https://www.correiodopovo.com.br/not%C3%ADcias/pol%C3%ADtica/elei%C3%A7%C3%B5es/diferen%C3%A7a-de-votos-entre-lula-e-bolsonaro-%C3%A9-a-menor-da-hist%C3%B3ria-1.915391. Acesso em: 2 fev. 2023.

DYLGJERI, Ardita. Analysis of speech acts in political speeches. **European Journal of Social Sciences Studies**, 2017.

ELEIÇÃO para Presidente. **G1 - GLOBO**, 4 out. 2022. Disponível em: https://g1.globo.com/politica/eleicoes/2022/apuracao/presidente.ghtml. Acesso em: 3 fev. 2023.

FAGUNDES, Álvaro; FELÍCIO , César; SCIARRETTA, Toni. Marcas da pandemia: Com 10 milhões de casos em menos de um ano, Brasil é o terceiro país com mais infectados por covid-19. Na esteira da crise sanitária, economia sofre ainda com vacinação lenta e dúvidas

sobre auxílio emergencial. **Valor Econômico**, São Paulo, 18 fev. 2021. Disponível em: https://valor.globo.com/coronavirus/a-economia-na-pandemia/. Acesso em: 28 maio 2023.

FAN, W.; WALLACE, L.; RICH, S. Tapping The Power of Text Mining: Communications of the ACM. **Research Gate**, p. 77-80, 2006.

DE SOUZA, Karine França; PEREIRA, Moisés Henrique Ramos; DALIP, Daniel Hasan. Unilex: Método léxico para análise de sentimentos textuais sobre conteúdo de tweets em português brasileiro. **Abakós**, v. 5, n. 2, p. 79-96, 2017.

GONÇALVES, Cristiano de Andrade. **Análise de sentimentos em reclamações: uma aplicação no maior site de reclamações do Brasil**. 2016. Tese de Doutorado.

GONÇALVES, Pollyanna de Oliveira. **Um benchmark para comparação de métodos para análise de sentimentos**. 2015. Dissertação (Mestrado em Computação) - Universidade Federal de Minas Gerais, Belo Horizonte, 2015

GRANJEIA, Julianna. Sob pressão, Ciro Gomes tenta se reafirmar como terceira via. **Valor Econômico**, Santos, 15 maio 2022. Disponível em: https://valor.globo.com/politica/noticia/2022/05/12/sob-presso-ciro-gomes-tenta-se-reafirmar-como-terceira-via.ghtml. Acesso em: 25 maio 2023.

GRIMMER, Justin; STEWART, Brandon M. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. **Political analysis**, v. 21, n. 3, p. 267-297, 2013

IEZZI, Domenica Fioredistella; CELARDO, Livia. Text analytics: Present, past and future. In: **Text Analytics: Advances and Challenges**. Springer International Publishing, 2020. p. 3-15.

KUMAR, Sunil; KAR, Arpan Kumar; ILAVARASAN, P. Vigneswara. Applications of text mining in services management: A systematic literature review. **International Journal of Information Management Data Insights**, v. 1, n. 1, p. 100008, 2021.

LIMA, Paola. **Mulheres na política**: ações buscam garantir maior participação feminina no poder. São Paulo: Agência Senado, 27 maio 2022. Disponível em: https://www12.senado.leg.br/noticias/infomaterias/2022/05/aliados-na-luta-por-mais-mulheres-na-politica#:~:text=De%20acordo%20com%20o%20IBGE,de%2015%25%20dos%20cargos%20eletivos. Acesso em: 14 jun. 2023.

MACHADO, Aydano P. et al. Mineração de texto em Redes Sociais aplicada à Educação a Distância. **Colabor@-A Revista Digital da CVA-RICESU**, v. 6, n. 23, 2010.

MACHADO, Mateus Tarcinalli. **Estudo e avaliação de métodos de análise de sentimentos baseada em aspectos para textos opinativos em português**. 2018. Tese de Doutorado. Universidade de São Paulo.

MEDHAT, Walaa; HASSAN, Ahmed; KORASHY, Hoda. Sentiment analysis algorithms and applications: A survey. **Ain Shams engineering journal**, v. 5, n. 4, p. 1093-1113, 2014.

PAINEL Coronavírus. **CORONAVÍRUS BRASIL**. São Paulo, 18 jun. 2023. Disponível em: https://covid.saude.gov.br/. Acesso em: 18 jun. 2023.

PENNEBAKER, James W.; FRANCIS, Martha E.; BOOTH, Roger J. Linguistic inquiry and word count: LIWC 2001. **Mahway: Lawrence Erlbaum Associates**, v. 71, n. 2001, p. 2001, 2001.

PENNEBAKER, James W. et al. **The development and psychometric properties of LIWC2015**. 2015.

PINTO, Céli Regina Jardim. Elementos para uma análise de discurso político. **Barbarói: revista do Departamento de Ciências Humanas e do Departamento de Psicologia. Santa Cruz do Sul, RS. N. 24 (jan./jun. 2006), p. 78-109**, 2006.

RAVI, Kumar; RAVI, Vadlamani. A survey on opinion mining and sentiment analysis: tasks, approaches and applications. **Knowledge-based systems**, v. 89, p. 14-46, 2015.

ROUGIER, Nicolas P.; DROETTBOOM, Michael; BOURNE, Philip E. Ten simple rules for better figures. **PLoS computational biology**, v. 10, n. 9, p. e1003833, 2014.

SAVOY, Jacques. Lexical analysis of US political speeches. **Journal of Quantitative Linguistics**, v. 17, n. 2, p. 123-141, 2010.

SIM, Yanchuan et al. Measuring ideological proportions in political speeches. In: **Proceedings of the 2013 conference on empirical methods in natural language processing**. 2013. p. 91-101.