Análise de Títulos e Conteúdos Políticos em Portais de Notícias Online

Maurício Jornada Bastos[1], **Valéria Q. Reis**[1,2]

[1]Faculdade de Computação – Universidade Federal de Mato Grosso do Sul, Brasil

[2]Instituto de Sistemas de Informação – Leuphana Universität Lüneburg, Alemanha

{m.jornada,valeria.reis}@ufms.br

Resumo. *Este trabalho investiga as características textuais de títulos de notícias de quatro grandes portais brasileiros — Folha de São Paulo, G1, Jovem Pan e CNN Brasil — durante o período eleitoral de 2022. O objetivo é identificar se existem "assinaturas" editoriais e temáticas que permitam diferenciar as fontes de notícia. Por meio de análises descritivas, como a frequência de termos e a métrica TF-IDF, foram revelados padrões distintos no vocabulário e no estilo de cada portal. Subsequentemente, um experimento de classificação de texto foi conduzido, utilizando modelos de linguagem baseados na arquitetura Transformer. O modelo* `neuralmind/bert-base-portuguese-cased`*, pré-treinado para o português do Brasil, demonstrou ser capaz de prever o portal de origem de uma notícia com um F1-Score médio de 82,5%, validando a hipótese de que os portais possuem características textuais distinguíveis. Os resultados indicam que, apesar de uma cobertura temática comum centrada nas eleições, nuances no foco editorial e no estilo linguístico são suficientes para uma diferenciação automática, contribuindo para o campo do Processamento de Linguagem Natural aplicado à análise de mídia.*

1. Introdução

Na concepção clássica da Filosofia, ideologia é compreendida como um conjunto de ideias, valores, crenças e representações que orientam a visão de mundo e o comportamento de indivíduos ou grupos sociais [Porfírio 2024]. Essas ideias exercem uma função social, ao moldar comportamentos e sustentar ou desafiar relações de poder, além de organizarem percepções coletivas da realidade. A ideologia também possui um caráter cultural, uma vez que é compartilhada e reproduzida entre os membros de um grupo por meio de instituições e práticas culturais.

A formação das ideologias ocorre em um contexto de interação entre experiências pessoais e as influências externas, como valores familiares, escolas, igrejas e veículos de comunicação. Esses fatores contribuem para a construção da identidade de cada indivíduo e refletem nas escolhas feitas no cotidiano, inclusive na preferência por determinados tipos de mídia.

A preferência por mídias alinhadas à orientação política dos leitores pode ser explicada pela teoria do viés de confirmação, que descreve a tendência de buscar, interpretar e recordar informações que reforcem crenças preexistentes. Essa inclinação cognitiva leva os indivíduos a preferirem fontes que sustentem sua visão de mundo e a evitarem aquelas que a desafiem. Tal comportamento resulta em bolhas informativas, nas quais as

pessoas interagem quase exclusivamente com ideias que ressoam com as suas, criando uma percepção distorcida de consenso e exacerbando polarizações. Estudos, como o de Sunstein (2001), evidenciam que a exposição limitada a fontes unilaterais não apenas reforça crenças existentes, mas também pode levar à radicalização ideológica ao longo do tempo [Sunstein 2001].

A eleição presidencial de 2018, no Brasil, foi um exemplo claro da intensificação da polarização política, com o eleitorado dividido entre apoiadores do governo anterior e do governo recém-eleito [Fuks and Marques 2022, Brasil 2023]. Essa polarização evidenciou o papel central da mídia online na disseminação de discursos alinhados a diferentes espectros ideológicos. Nesse cenário, torna-se relevante investigar como as notícias brasileiras retratam a política nacional, buscando padrões de similaridade entre as fontes de informação. Nesse contexto, é importante notar que, embora a televisão (63,7%) e as redes sociais (53,8%) ainda sejam as mídias de notícias mais acessadas pela população, os sites de jornais e os portais de notícias online figuram logo em seguida, sendo fontes frequentes de informação para 39,1% e 30,2% dos brasileiros, respectivamente [YouGov Profiles 2024]. Esses portais online, em particular, representam importantes fontes de estudo para o Processamento de Linguagem Natural, pois, diferentemente de muitas interações em redes sociais, tendem a prover textos mais longos, bem estruturados e compatíveis com a norma culta.

1.1. Objetivos

O objetivo deste trabalho é verificar a existência de características nas notícias de portais brasileiros que possam identificar cada um deles. Gostaríamos de responder às seguintes questões:

- **Q1 - Existem diferenças entre os tipos de notícias/conteúdos dos portais?** Se houver, a quê se devem essas diferenças? Ao número de palavras utilizadas, temas abordados, sentimento das manchetes?
- **Q2 - Há similaridades entre os portais?** É possível agrupar os portais de acordo com a semelhança de suas notícias/conteúdos?
- **Q3 - Do que tratam os conteúdos das notícias de cada portal?** Destacar os principais tópicos abordados por cada portal. Há similaridade entre eles?

Nossa hipótese é que os títulos das notícias, assim como os conteúdos das matérias, são bem definidos e diferenciados entre si, revelando os tópicos frequentemente abordados e os principais sentimentos comunicados ao público. A partir da confirmação da hipótese, seria possível treinar modelos para a classificação automática de notícias. Os modelos poderiam ser utilizados, por exemplo, por estudiosos de áreas sociais para quantificar e compreender melhor o impacto das mídias na sociedade.

Nesse sentido, este trabalho oferece contribuições em três frentes principais. Primeiro, apresenta um novo corpus de notícias políticas em português, coletado durante um período de alta relevância nacional. Segundo, realiza uma análise empírica detalhada que quantifica as diferenças estilísticas e temáticas entre quatro grandes portais de mídia brasileiros. Terceiro, valida a eficácia de um modelo de linguagem especializado em português para a tarefa de classificação de fontes de notícias, demonstrando a existência de "assinaturas editoriais" computacionalmente reconhecíveis.

Para atingir o objetivo proposto, um corpus de notícias foi criado e métodos de experimentação foram desenhados.

Os resultados desta pesquisa dialogam diretamente com os Objetivos de Desenvolvimento Sustentável (ODS) da ONU. A análise de vieses e características da mídia contribui para o **ODS 16 (Paz, Justiça e Instituições Eficazes)**, que, em sua meta 16.10, busca "assegurar o acesso público à informação e proteger as liberdades fundamentais". Ao desenvolver métodos que permitem identificar e compreender as "assinaturas" editoriais dos portais de notícias, o trabalho oferece uma ferramenta para promover a transparência da informação e a literacia mediática, fortalecendo as instituições democráticas. Indiretamente, a pesquisa também tangencia o **ODS 10 (Redução das Desigualdades)**, pois a polarização, muitas vezes amplificada pela mídia, pode acentuar divisões e desigualdades sociais.

2. Desenvolvimento

O desenvolvimento deste trabalho foi estruturado em três etapas metodológicas principais. A primeira etapa, **Construção da Base de Notícias (Corpus)**, detalha a origem e o tratamento dos dados. A segunda, **Análise dos Títulos: O que os Dados Revelam?**, investiga as características intrínsecas dos títulos de notícias. Por fim, a **Metodologia Experimental para Classificação de Portais** apresenta a abordagem de classificação automática utilizada para diferenciar os portais. As seções a seguir detalham cada uma dessas etapas.

2.1. Construção da Base de Notícias (Corpus)

A fundação de qualquer projeto de Processamento de Linguagem Natural é um corpus de dados robusto. Esta etapa do trabalho foi dedicada a esse objetivo e compreendeu dois processos fundamentais: primeiro, a coleta e curadoria dos dados brutos; e segundo, um estudo piloto de anotação manual que foi essencial para definir a direção experimental final do projeto.

2.1.1. Coleta Automatizada e Seleção das Notícias

Para a criação do corpus, foram coletados títulos e textos de notícias de quatro dos mais importantes portais de notícia do país: *Folha de São Paulo*[1], *G1*[2], *Jovem Pan*[3] e *CNN Brasil*[4]. A escolha do período, de 31 de agosto a 1º de novembro de 2022, foi estratégica para coincidir com a janela temporal do horário eleitoral gratuito no rádio e na televisão. Esse intervalo representa o auge do debate público e da cobertura midiática das eleições, caracterizado por discussões políticas intensas, sendo, portanto, ideal para a análise dos discursos veiculados.

O processo de raspagem de dados, essencial para a construção do dataset, foi desenvolvido integralmente na linguagem Python, com o suporte das bibliotecas *Beautiful Soup* e *Scrapy*. A escolha dessas ferramentas foi estratégica e baseada em suas funções

[1] https://www.folha.uol.com.br/
[2] https://g1.globo.com/
[3] https://jovempan.com.br/
[4] https://www.cnnbrasil.com.br/

complementares. O *Beautiful Soup* é uma biblioteca de análise sintática (*parsing*) de HTML, ideal para extrair com precisão informações específicas (como títulos e links) de uma página web já carregada. Em paralelo, o *Scrapy* é um *framework* de *web crawling* completo, projetado para gerenciar de forma robusta e assíncrona todo o processo de requisição, navegação entre múltiplas páginas e extração de dados em larga escala. Conforme detalhado adiante, o *Beautiful Soup* foi usado na coleta inicial de links, enquanto o *Scrapy* foi empregado para extrair o corpo completo de milhares de notícias.

Inicialmente, um *crawler* com *Beautiful Soup* foi configurado para navegar pelos portais e coletar os títulos das reportagens, as datas de publicação e os links correspondentes, armazenando esses dados em um arquivo CSV. Posteriormente, uma *spider* personalizada em *Scrapy* utilizou os links desse arquivo para acessar as páginas individuais e extrair o corpo completo de cada notícia, enriquecendo o dataset.

Uma etapa fundamental da curadoria foi o **tratamento de duplicatas**, visando endereçar a preocupação sobre como artigos editados ou atualizados seriam tratados. Para garantir que cada notícia fosse representada apenas uma vez no corpus, independentemente de edições posteriores, foi aplicado um processo de desduplicação após a coleta. Utilizando a biblioteca *pandas*, removemos todas as entradas que possuíam o mesmo link (URL), usando este campo como identificador único para cada matéria. Esta abordagem assegura que o dataset final não contém notícias duplicadas.

O processo inicial de raspagem resultou em um corpus bruto de 4.853 notícias. Foi observada, no entanto, uma grande disparidade na quantidade de notícias por portal, com um volume de 2.778 títulos provenientes apenas da CNN Brasil. Para evitar que o desbalanceamento excessivo pudesse viciar os modelos de classificação, optou-se por realizar uma curadoria no corpus. Foi aplicada uma técnica de subamostragem aleatória (*random subsampling*) sobre as notícias da CNN Brasil, reduzindo sua representação para 1.000 títulos.

Dessa forma, o corpus final, utilizado em todas as análises e experimentos subsequentes deste trabalho, foi consolidado em **3.075 notícias**. A distribuição final das notícias por portal é apresentada na Tabela 1.

Tabela 1. Distribuição final do corpus utilizado nos experimentos.

Portal	**Quantidade**
Folha de São Paulo	653
G1	505
Jovem Pan	917
CNN Brasil	1.000
Total	**3.075**

2.1.2. Estudo Piloto: Classificação Manual do Sentimento

A abordagem inicial deste trabalho visava classificar os títulos das notícias de acordo com o sentimento veiculado (positivo, neutro ou negativo). Para construir um conjunto de dados para esta tarefa, foi realizado um processo de anotação manual. A definição da amostra foi parte de um processo iterativo de calibração entre os dois anotadores (o aluno

e a orientadora), visando estabelecer critérios de classificação consistentes. O processo ocorreu em rodadas: lotes de aproximadamente 50 títulos de um portal eram anotados independentemente, seguidos por uma reunião de consenso para discutir divergências e refinar as diretrizes de anotação. Este ciclo foi repetido para os diferentes portais até que uma consistência metodológica fosse atingida. Este "acordo interavaliador satisfatório" não foi definido de forma subjetiva, mas sim pela obtenção de um nível de concordância estatística — mensurado pelo índice Fleiss' Kappa, como detalhado adiante — classificado, no mínimo, como "moderado". Os números finais — 100 títulos da Folha de S. Paulo, 100 da Jovem Pan, 51 do G1 e 50 da CNN Brasil (totalizando 301 títulos) — refletem o total de amostras utilizadas ao término desta fase de calibração, o que justifica a variação na quantidade por portal.

O processo de anotação foi realizado por dois avaliadores (o aluno e a orientadora). Cada um, de forma independente, classificava os títulos em uma de três categorias: **-1 (Negativo)**, **0 (Neutro)** ou **1 (Positivo)**. Quinzenalmente, os anotadores se reuniam para debater as classificações divergentes, discutir os critérios e refinar o padrão de anotação até que se chegasse a um consenso para cada título.

Para validar a confiabilidade e a consistência do processo antes da fase de debate, foi mensurado o acordo interavaliador utilizando o índice Fleiss' Kappa [Fleiss 1971]. Esta métrica estatística avalia o grau de concordância entre múltiplos avaliadores, corrigindo a concordância que ocorreria puramente por acaso. Seus valores variam de -1 a 1, onde 1 indica concordância perfeita. Os resultados do cálculo de concordância para as amostras de cada portal foram os seguintes:

- **G1:** $k = 0,934$
- **Folha de São Paulo:** $k = 0,639$
- **CNN Brasil:** $k = 0,617$
- **Jovem Pan:** $k = 0,500$

De acordo com a escala de interpretação proposta por Landis e Koch [Landis and Koch 1977], os resultados indicam um acordo "quase perfeito" para o G1, "substancial" para a Folha de São Paulo e CNN Brasil, e "moderado" para a Jovem Pan. Os valores, em geral, demonstram que o processo de classificação foi consistente. No entanto, como será discutido na Seção 2.3.1, os desafios relacionados à obtenção de um grande volume de dados para esta tarefa motivaram uma mudança de escopo no trabalho.

2.2. Análise Inicial: O que os Títulos Revelam?

A fase inicial de análise e pré-processamento dos dados textuais é crucial para extrair informações relevantes e preparar os dados para etapas posteriores, como modelagem ou classificação. Neste trabalho, a análise descritiva focou em compreender a estrutura e o vocabulário dos títulos de notícias, visando identificar padrões proeminentes em cada um dos quatro portais analisados.

2.2.1. Limpeza e Preparação dos Textos para Análise

O pré-processamento dos títulos foi conduzido com a biblioteca **spaCy**, uma ferramenta robusta para Processamento de Linguagem Natural (PLN). O objetivo primordial foi lim-

par o texto, padronizar as palavras e reduzir a dimensionalidade do vocabulário, eliminando elementos que não contribuem significativamente para o valor semântico. Para isso, aplicou-se um método que consistiu em diversas etapas para cada título. Primeiramente, o texto foi **normalizado**, convertido para minúsculas e com espaços em branco no início e fim removidos. Em seguida, o spaCy realizou a **tokenização**, segmentando o texto em tokens (palavras, pontuação, etc.). Procedeu-se à **remoção de pontuação e espaços**, excluindo tokens que representam pontuação, espaços e quebras de linha. Também foi aplicada a **remoção de stopwords**, eliminando palavras de alta frequência, mas de baixo conteúdo semântico (como artigos, preposições, conjunções, conforme definido pelo modelo de português do spaCy). Adicionalmente, foi feito um **filtro alfabético** para manter apenas tokens compostos exclusivamente por caracteres alfabéticos, removendo números ou símbolos isolados. Por fim, a **lematização** foi aplicada, onde cada token restante foi reduzido ao seu lema, a forma base ou canônica da palavra (por exemplo: "correndo" e "correu" seriam reduzidos a "correr"; "casas" a "casa"). A lematização agrupa diferentes flexões da mesma palavra, otimizando a precisão da contagem de frequência e a eficácia da análise.

O resultado desse pré-processamento foi armazenado em uma nova coluna no *dataframe*, onde cada entrada consiste em uma lista de tokens lematizados, limpos e filtrados do título original.

Com os títulos pré-processados, a etapa seguinte da análise descritiva concentrou-se na quantificação da frequência das palavras para identificar os termos mais recorrentes em cada portal. Este procedimento envolveu uma **iteração por portal**, onde o *dataframe* foi percorrido, filtrando os dados para cada um dos rótulos de portal (0 a 3) individualmente. Posteriormente, ocorreu a **agregação de tokens**, onde, para cada portal, todas as listas de tokens dos títulos foram concatenadas em uma única lista grande, formando o vocabulário total processado para aquela fonte. A seguir, a **contagem de frequência** foi realizada, utilizando a biblioteca `collections.Counter` para contar a ocorrência de cada palavra na lista agregada de tokens. Concluindo esta etapa, procedeu-se à **identificação dos mais frequentes**, determinando as 10 palavras mais comuns para cada portal.

A apresentação dessas palavras mais frequentes, juntamente com suas contagens, forneceu uma visão inicial sobre o conteúdo e o foco temático de cada portal de notícias, permitindo observar distinções no vocabulário e, consequentemente, nos tipos de notícias que tendem a publicar. A contagem total de palavras processadas por portal também foi utilizada como um indicador da quantidade de dados textuais analisados para cada fonte. Esta análise descritiva serve como base para compreensões mais aprofundadas na sequência do trabalho.

2.2.2. Análise de Comprimento e Distribuição

A análise descritiva dos títulos por portal revelou características distintas em termos de volume de dados e extensão dos títulos.

A **contagem total de títulos** por portal demonstrou uma distribuição desigual. A **CNN Brasil** apresentou o maior número de títulos (1.000), seguida pela **Jovem Pan**

(917), **Folha de São Paulo** (653) e, por fim, o **G1** (505). Em termos percentuais, a CNN Brasil responde por aproximadamente 32,52% dos títulos, a Jovem Pan por 29,82%, a Folha de São Paulo por 21,24% e o G1 por 16,42%. Essa variação no volume de dados entre as fontes é um fator a ser considerado em análises subsequentes.

Ao examinar o **comprimento dos títulos**, tanto em caracteres quanto em palavras, observaram-se padrões interessantes. As estatísticas descritivas, apresentadas na Tabela 2, indicam que, em média, os títulos da **CNN Brasil** são os mais longos, com uma média de 130,16 caracteres e 20,86 palavras por título. Isso é significativamente maior do que os outros portais, que apresentam médias de comprimento mais próximas entre si. Por exemplo, a **Folha de São Paulo** tem uma média de 76,79 caracteres e 12,75 palavras, a **Jovem Pan** 82,08 caracteres e 13,32 palavras, e o **G1** 93,95 caracteres e 15,26 palavras.

Ao examinar o **comprimento dos títulos**, tanto em caracteres quanto em palavras, observaram-se padrões interessantes. As estatísticas descritivas, apresentadas na Tabela 2, indicam que, em média, os títulos da **CNN** (`label 2`) são os mais longos, com uma média de 130,16 caracteres e 20,86 palavras por título. Isso é significativamente maior do que os outros portais, que apresentam médias de comprimento mais próximas entre si. Por exemplo, a **Folha de São Paulo** (`label 0`) tem uma média de 76,79 caracteres e 12,75 palavras, a **Jovem Pan** (`label 1`) 82,08 caracteres e 13,32 palavras, e a **G1** (`label 3`) 93,95 caracteres e 15,26 palavras.

Tabela 2. Estatísticas descritivas do comprimento dos títulos por portal (média e desvio padrão).

Portal	Média[a]	Desvio Padrão[a]	Média[b]	Desvio Padrão[b]
Folha de São Paulo	76,79	12,09	12,75	2,33
Jovem Pan	82,08	15,17	13,32	2,75
CNN	130,16	32,93	20,86	5,58
Globo	93,95	21,79	15,26	3,72

[a] Valores em caracteres.
[b] Valores em palavras.

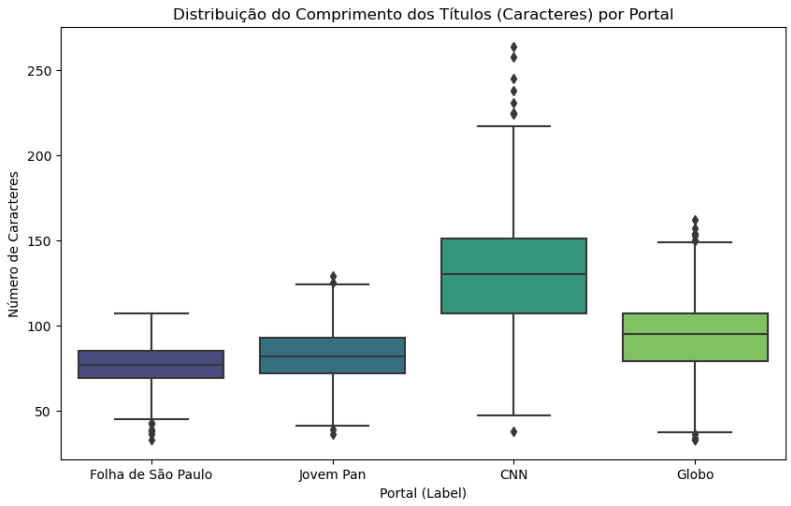

Figura 1. Distribuição do comprimento de caracteres por portal.

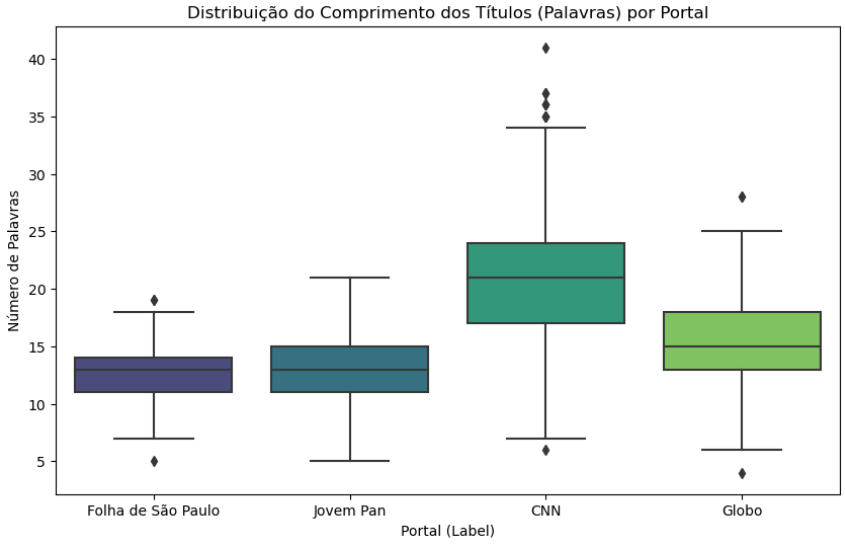

Figura 2. Distribuição do comprimento de palavras por portal.

A **distribuição do comprimento dos títulos** por meio de *box plots* (Figuras 1 e 2) corrobora essas observações, mostrando que a **CNN** possui uma dispersão maior e valores de comprimento significativamente mais altos, o que sugere um estilo editorial que favorece títulos mais detalhados ou explicativos. Os demais portais exibem distribuições de comprimento mais concentradas e com medianas menores, indicando uma tendência a títulos mais concisos. A presença de *outliers* (pontos fora das caixas) em todos os portais sinaliza alguns títulos excepcionalmente curtos ou longos, mas a tendência geral de cada portal é clara.

A contagem e proporção dos títulos por portal, juntamente com as análises de comprimento, fornecem um panorama inicial da composição do *dataset*. A CNN, por ter o maior volume de dados e títulos mais longos, pode, naturalmente, apresentar um vocabulário mais diversificado em suas análises de frequência de palavras e TF-IDF, comparado aos portais com menor número de títulos ou títulos mais curtos. Essa diferença no perfil dos dados é um aspecto importante para interpretar os resultados das análises textuais subsequentes.

2.2.3. Identificando as Palavras-Chave de Cada Portal (TF-IDF)

Após o pré-processamento dos dados textuais e a análise inicial de frequência, empregou-se a técnica TF-IDF (*Term Frequency-Inverse Document Frequency*). O TF-IDF é crucial para identificar a importância de um termo em um "documento" — neste contexto, o conjunto de títulos de um portal de notícias — em relação a um corpus maior, que engloba os títulos de todos os portais analisados. Diferentemente da simples contagem de frequência, que apenas indica quão comum uma palavra é, o TF-IDF atribui maior valor a termos que são frequentes em um documento específico, mas raros no corpus geral. Isso permite destacar palavras distintivas que verdadeiramente caracterizam e diferenciam o conteúdo de cada fonte de notícia.

Para a aplicação desta técnica, os dados pré-processados precisaram ser agrupa-

dos. Primeiramente, todos os títulos limpos de um mesmo portal foram reunidos em um único "super-documento". Ao final desse processo, tínhamos quatro "super-documentos", um representando o vocabulário total de cada portal.

Esses quatro documentos foram então processados com a biblioteca *scikit-learn* para o cálculo do TF-IDF. Durante essa etapa, dois filtros importantes foram aplicados para focar apenas nas palavras relevantes. O primeiro filtro removeu termos que apareciam em **quase todos** os portais (como palavras genéricas sobre política), pois estes não ajudam a diferenciar um portal do outro. O segundo filtro descartou termos **excessivamente raros** (que surgiram em apenas um portal), que não são estatisticamente relevantes para encontrar um padrão. Ao filtrar esses "ruídos", o cálculo do TF-IDF pôde focar nas palavras que são, de fato, distintivas de cada portal. O resultado foi uma tabela de pontuação, onde cada palavra do vocabulário recebeu um score de importância para cada um dos quatro portais.

Com a matriz TF-IDF calculada, procedeu-se à identificação dos termos mais relevantes para cada portal individualmente. Para cada linha da matriz (correspondente a um portal), os scores TF-IDF foram extraídos, e os índices dos maiores scores foram recuperados para associá-los aos termos correspondentes no vocabulário. A Tabela 3 apresenta os 10 termos com os maiores scores TF-IDF para cada um dos portais analisados.

Tabela 3. Principais termos por portal com maior score TF-IDF.

Posição	**Folha de São Paulo**		**Jovem Pan**		**CNN**		**Globo**	
	Termo	Score	Termo	Score	Termo	Score	Termo	Score
1	derrota	0,1874	confira	0,2761	levantamento	0,3723	matar	0,2648
2	bolsonarismo	0,1781	instituto	0,1652	erro	0,3689	briga	0,2648
3	citar	0,1730	pan	0,1561	percentual	0,3585	ms	0,2181
4	evangélico	0,1586	freita	0,1458	margem	0,3551	confirar	0,2025
5	fake	0,1297	prefeito	0,1458	face	0,2129	morrer	0,1892
6	folha	0,1297	liderar	0,1458	ouvir	0,1620	carro	0,1869
7	news	0,1297	piso	0,1440	dizer	0,1551	pm	0,1513
8	zema	0,1297	criticar	0,1361	tribunal	0,1482	balear	0,1513
9	avançar	0,1247	zema	0,1361	apresentar	0,1345	sc	0,1513
10	trar	0,1247	eletrônico	0,1166	comentar	0,1138	igreja	0,1387

A análise dos termos com maior score TF-IDF para cada portal revela nuances importantes sobre seus focos temáticos e abordagens. A **Folha de São Paulo** apresenta termos como "derrota", "bolsonarismo", "evangélico", "fake" e "news", sugerindo uma cobertura que frequentemente aborda política, polarização e questões sociais e religiosas, com uma possível ênfase em discussões sobre desinformação ("fake news"). O termo "folha" também aparece, o que é esperado e reforça a autorreferência em seus títulos.

Já a **Jovem Pan** se destaca por termos como "confira", "instituto", "pan" e "eletrônico". "Confira" e "pan" (autorreferência ao portal) podem indicar um estilo mais direto e promocional no engajamento com o leitor, enquanto "instituto" e "eletrônico" podem estar relacionados a notícias sobre pesquisas de opinião, tecnologia ou mesmo temas eleitorais (urnas eletrônicas, por exemplo). A repetição de "zema" e outros nomes políticos como "prefeito" e "freita" (possivelmente ligado a "Freitas" ou similar) sugere uma forte cobertura de política local e estadual.

A **CNN** exibe termos como “levantamento”, “erro”, “percentual”, “margem” e “tribunal”. Isso indica uma inclinação para reportagens investigativas, análises de dados, pesquisas e notícias sobre o sistema judiciário. Os termos sugerem um jornalismo que busca aprofundamento e precisão, frequentemente citando fontes e dados estatísticos.

Por fim, o **G1** apresenta termos mais voltados para o cotidiano e ocorrências, como “matar”, “briga”, “morrer”, “carro”, “pm” e “balear”. Isso aponta para uma cobertura mais voltada a notícias de segurança pública, acidentes, crimes e eventos locais que impactam diretamente a população. A presença de “igreja” e “ms” (Mato Grosso do Sul) ou “sc” (Santa Catarina) pode indicar um foco em notícias regionais ou temas específicos ligados a comunidades e religião.

Em síntese, a análise TF-IDF corrobora e aprofunda as observações iniciais de frequência, demonstrando que, enquanto alguns termos são de esperada autorreferência (“folha”, “pan”), a maioria dos termos distintivos reflete as linhas editoriais e os focos temáticos predominantes de cada portal. Esta técnica permite uma compreensão mais refinada do vocabulário que caracteriza e diferencia cada fonte de notícia no panorama midiático.

2.2.4. Análise de Frequência com Nuvem de Palavras

A Nuvem de Palavras constitui uma técnica de visualização de dados textuais amplamente utilizada para representar a frequência de termos em um dado *corpus*. Nesta representação gráfica, a proeminência visual de uma palavra (geralmente seu tamanho) é diretamente proporcional à sua frequência no conjunto de textos analisado. Essa abordagem permite uma rápida identificação dos termos mais recorrentes, fornecendo informações imediatas sobre os tópicos dominantes ou o vocabulário característico de um conjunto de documentos.

Para a geração das nuvens de palavras correspondentes a cada portal de notícias, foram empregados os tokens resultantes da etapa de pré-processamento dos títulos. De maneira similar ao preparo dos dados para a análise TF-IDF, todos os tokens associados aos títulos de um mesmo portal foram concatenados em uma única string, utilizando espaços como delimitadores. Este procedimento foi replicado individualmente para cada um dos quatro portais em estudo, resultando em um “documento” textual consolidado para cada fonte, a partir de seus títulos pré-processados. Uma verificação foi incluída no processo para prevenir a tentativa de gerar uma nuvem a partir de um texto vazio, garantindo a robustez da análise mesmo na ausência de dados processados para algum portal específico.

A geração das visualizações foi implementada utilizando a biblioteca `wordcloud` em Python. Para cada string de texto agregada por portal, uma instância da classe `WordCloud` foi configurada com parâmetros visuais específicos, incluindo largura, altura, cor de fundo (`white`) e um mapa de cores (`viridis`). Adicionalmente, o parâmetro `max_words` foi definido como 100, limitando a nuvem às cem palavras mais frequentes para otimizar a clareza e a interpretabilidade da visualização.

A manipulação e exibição das visualizações foram realizadas com a biblioteca `matplotlib.pyplot`. Para cada nuvem gerada, uma figura foi criada, e a imagem da

nuvem foi apresentada utilizando a função `plt.imshow`. Os eixos da plotagem foram desativados para focar exclusivamente na nuvem de palavras. Um título dinâmico, indicando o portal correspondente, foi adicionado a cada figura. As nuvens geradas foram salvas em arquivos `.png` para inclusão posterior no documento.

A análise visual das nuvens de palavras para cada portal (Figuras 3 a 6) revela padrões distintos nos termos mais frequentes, refletindo potencialmente as abordagens editoriais ou os focos temáticos de cada fonte durante o período analisado.

No caso da **Folha de São Paulo** (Figura 3), a nuvem destaca termos como "Bolsonaro", "Lula", "eleição", "eleitoral", "campanha" e "político". Termos relacionados a instituições e processos, como "governo", "partido", "ministro" e "voto", também aparecem com relevância. Observa-se uma concentração em nomes de figuras políticas proeminentes e vocabulário diretamente associado ao processo eleitoral e ao cenário político nacional.

Figura 3. Nuvem de Palavras da Folha de São Paulo.

Para a **Jovem Pan** (Figura 4), a proeminência de "Bolsonaro" e "Lula" é similar à da Folha, mas a nuvem também apresenta termos como "governo", "Brasil", "eleição", "político" e "ministro" com destaque. Palavras como "defender", "justiça" e "pedir" sugerem uma cobertura que pode focar em ações governamentais, questões judiciais ou pedidos e demandas políticas. A presença de "campanha" e "eleitoral" reforça o foco no contexto das eleições.

Figura 4. Nuvem de Palavras da Jovem Pan.

A nuvem da **CNN** (Figura 5) apresenta termos como “presidente”, “percentual”, “eleitoral”, “candidato”, “pesquisa” e “margem”. A forte presença de “percentual”, “pesquisa” e “margem” indica um foco significativo na cobertura de pesquisas de intenção de voto e na análise numérica do cenário eleitoral. Termos como “debate” e “entrevista” sugerem uma abordagem que inclui a interação direta com candidatos e discussões sobre temas relevantes.

Figura 5. Nuvem de Palavras da CNN.

Finalmente, a nuvem do **G1** (Figura 6) exibe palavras como “político”, “eleição”, “governo”, “polícia”, “candidato” e “Brasil”. A palavra “polícia” aparece com um destaque notável em comparação com os outros portais, sugerindo uma cobertura mais frequente ou aprofundada de temas relacionados à segurança pública, operações policiais ou investigações que envolvem figuras políticas. Termos como “matar” e “violência” reforçam essa possível inclinação.

Figura 6. Nuvem de Palavras da Globo.

Em suma, as nuvens de palavras fornecem uma representação visual concisa da frequência dos termos nos títulos de cada portal. Embora todos os portais abordem o tema eleitoral e político central, as diferenças na proeminência de certos termos, como "percentual" e "pesquisa" na CNN, ou "polícia" na Globo, sugerem nuances em suas coberturas. A Folha e a Jovem Pan parecem ter um foco mais equilibrado entre nomes de políticos e o vocabulário eleitoral/político geral, enquanto a CNN se destaca pela cobertura baseada em dados de pesquisas, e o G1 pela inclusão de temas ligados à segurança e justiça. Essa análise visual complementa as análises de frequência de palavras e TF-IDF, oferecendo uma compreensão rápida das áreas temáticas prioritárias em cada fonte de notícia.

2.3. Metodologia Experimental para Classificação de Portais

A fase experimental foi guiada pela busca de padrões textuais que pudessem caracterizar os portais, culminando no desenvolvimento de um classificador de fontes.

2.3.1. Desafios da Abordagem Inicial: Análise de Sentimento

A primeira abordagem considerada foi a análise de sentimento dos títulos, com o objetivo de verificar se o tom (positivo, neutro ou negativo) poderia ser um diferenciador entre os portais. Para isso, uma parte do corpus foi anotada manualmente. Contudo, essa abordagem enfrentou desafios significativos. A coleta de um volume de dados suficiente para treinar um modelo de sentimento de forma eficaz para cada portal se mostrou inviável dentro do escopo do projeto, resultando em um dataset desbalanceado e com poucas amostras para algumas classes. Tal desbalanceamento prejudicou o desempenho dos treinamentos preliminares, mesmo com o uso de técnicas de aumento de dados (*text augmentation*). Diante desses obstáculos, o foco do trabalho foi redirecionado para uma tarefa de classificação mais direta e robusta, aproveitando a totalidade do corpus coletado.

2.3.2. Abordagem Final: Classificação de Notícias por Portal de Origem

A abordagem final, que se tornou o experimento central deste trabalho, consistiu em treinar modelos de PLN para prever o portal de origem de uma notícia baseando-se unicamente em seu título. Foram avaliados três modelos da arquitetura Transformer, selecionados por sua relevância e desempenho em tarefas de PLN em português:

Modelos Avaliados.

- `lxyuan/distilbert-base-multilingual-cased-sentiments-student`: Uma versão mais leve (destilada) e multilíngue, com foco em sentimentos.
- `distilbert-base-multilingual-cased`: O modelo DistilBERT base, também multilíngue.
- `neuralmind/bert-base-portuguese-cased`: Uma versão do BERT pré-treinada especificamente com um grande corpus de textos em português do Brasil.

Método de Validação. Para garantir a robustez e a capacidade de generalização dos resultados, foi utilizada a metodologia de validação cruzada estratificada (*Stratified K-Fold*) com $k = 5$. Essa técnica divide o dataset em 5 subconjuntos (ou *folds*), utilizando iterativamente 4 para treino e 1 para validação, e assegura que a proporção das classes (portais) seja mantida em cada dobra. O treinamento em cada dobra foi realizado por 8 épocas, com um tamanho de lote (*batch size*) de 64, taxa de aprendizado de 2×10^{-5} e decaimento de peso (*weight decay*) de 0.01.

3. Resultados

Esta seção detalha o desempenho dos modelos de classificação na tarefa de identificar o portal de origem dos títulos.

3.1. Desempenho Geral dos Modelos

Os resultados da validação cruzada para os três modelos são consolidados na Tabela 4. As métricas apresentadas são a média e o desvio padrão do F1-Score (ponderado), calculados sobre os resultados das 5 dobras do experimento. O F1-Score é uma média harmônica entre precisão e recall, sendo uma métrica robusta para avaliar o desempenho geral em problemas de classificação.

Tabela 4. Resultado consolidado da validação cruzada (5-folds). Média e desvio padrão do F1-Score ponderado.

Modelo	F1-Score Ponderado (Média $\pm$ DP)
`distilbert-base-multilingual-cased`	$0,781 \pm 0,024$
`lxyuan/...-sentiments-student`	$0,753 \pm 0,016$
`neuralmind/bert-base-portuguese-cased`	$\mathbf{0,825 \pm 0,022}$

O modelo `neuralmind/bert-base-portuguese-cased` alcançou o melhor desempenho, com um F1-Score médio de 82,5% e um desvio padrão baixo, indicando consistência nos resultados. O desempenho superior deste modelo, que é pré-treinado especificamente para o português do Brasil, sugere fortemente que sua especialização linguística foi um fator decisivo para capturar as nuances, o vocabulário e os estilos editoriais de cada portal nacional.

3.2. Análise de Desempenho por Classe

Para aprofundar a análise, a Tabela 5 apresenta o relatório de classificação detalhado para o melhor modelo (`neuralmind/bert-base-portuguese-cased`). As métricas de precisão, recall e F1-Score foram calculadas para cada classe (portal) e agregadas ao longo das 5 dobras da validação cruzada.

Tabela 5. Relatório de classificação agregado para o modelo `neuralmind/bert-base-portuguese-cased`.

Portal (Label)	Precisão	Recall	F1-Score	Suporte
Folha de S. Paulo (0)	0,80	0,68	0,73	653
Jovem Pan (1)	0,76	0,84	0,80	917
CNN Brasil (2)	0,93	0,95	0,94	1000
G1 (3)	0,73	0,69	0,71	505
Acurácia		0,82		
Média Ponderada	**0,82**	**0,82**	**0,82**	**3075**

Os resultados por classe são reveladores. O modelo foi excepcionalmente eficaz em identificar notícias da **CNN Brasil** (F1-Score de 0,94), o que está alinhado à análise descritiva, que mostrou que este portal possui um estilo de título mais longo e um vocabulário distintivo. A **Jovem Pan** também foi bem classificada (F1-Score de 0,80), possivelmente devido ao seu estilo editorial característico e ao uso de termos específicos. Em contraste, **Folha de S. Paulo** e **G1** apresentaram as pontuações mais modestas (F1-Score de 0,73 e 0,71, respectivamente), indicando que seus estilos podem ser mais generalistas ou possuir maior sobreposição, tornando a classificação mais desafiadora para o modelo.

4. Discussão

Os resultados obtidos permitem uma discussão aprofundada das questões de pesquisa que nortearam este trabalho.

Q1 - Existem diferenças entre os tipos de notícias/conteúdos dos portais? A resposta é afirmativa. As diferenças vão além do conteúdo superficial e se manifestam em características textuais mensuráveis. A análise descritiva (Seção 2.2) foi o primeiro indicativo, revelando variações no comprimento médio dos títulos, no vocabulário e nos temas prioritários. A análise TF-IDF, em particular, foi crucial para destacar os termos que conferem uma identidade única a cada portal. O sucesso do experimento de classificação (Seção 3) é a prova final de que essas diferenças são consistentes e aprendidas por um modelo de IA. A capacidade de um modelo prever a fonte de uma notícia com 82,5% de

F1-Score demonstra que os portais não são intercambiáveis, mas possuem "assinaturas" editoriais distintas.

Q2 - Há similaridades entre os portais? Apesar das diferenças marcantes, também existem similaridades. O modelo de classificação, embora preciso, não foi perfeito, e seus erros fornecem pistas sobre as semelhanças. O desempenho inferior na distinção entre a Folha de S. Paulo e o G1 sugere que esses dois veículos podem compartilhar um estilo jornalístico mais tradicional ou generalista, com menos marcadores textuais exclusivos em seus títulos quando comparados à CNN (foco em dados) ou à Jovem Pan (foco em política local e linguagem direta). Portanto, é possível inferir a existência de um "agrupamento" estilístico onde Folha e G1 estão mais próximos entre si do que dos outros dois portais.

Q3 - Do que tratam os conteúdos das notícias de cada portal? Durante o período eleitoral analisado, a política nacional foi o tema dominante em todos os portais. Contudo, a análise de conteúdo (TF-IDF e nuvens de palavras) revelou sub-focos temáticos distintos, que respondem a esta questão. A *Folha* concentrou-se em análises do cenário político ("bolsonarismo", "derrota", "fake news"). A *Jovem Pan* teve uma cobertura mais factual e ligada a atores específicos ("zema", "prefeito", "confira"). A *CNN Brasil* adotou uma abordagem quantitativa, com ênfase em pesquisas eleitorais ("levantamento", "percentual", "margem"). Por fim, o *G1*, apesar do seu foco político, também apresentou uma cobertura significativa de temas de segurança pública e cotidiano em nível nacional e local ("matar", "briga", "pm", "ms").

4.1. Impacto e Contribuições da Pesquisa

A validação da hipótese de que "assinaturas editoriais" são mensuráveis gera dois impactos centrais: um acadêmico, na área de PLN, e um social, ligado à literacia midiática.

Do ponto de vista **acadêmico e científico**, a principal contribuição é a **prova empírica** de que o estilo editorial dos portais é um padrão que pode ser aprendido por modelos de IA. O sucesso do modelo `neuralmind/bert-base-portuguese-cased` (82,5% F1) não apenas define um forte *baseline* para a tarefa de classificação de fontes em português, mas também reforça a importância de usar modelos de linguagem treinados para o idioma, em vez de alternativas multilíngues genéricas. Este trabalho oferece, portanto, um método de análise de estilo validado para o jornalismo brasileiro, servindo de base para futuras investigações sobre viés e enquadramento.

Do ponto de vista **social e prático**, o impacto é a **promoção da literacia midiática**. Ao provar que os portais *são* diferentes em seu vocabulário (como o foco da CNN em dados ou do G1 em segurança), o estudo ajuda a sociedade a entender como as "bolhas informativas" são formadas pela linguagem. A classificação automática de fontes oferece uma ferramenta para cientistas sociais e de comunicação analisarem a mídia em larga escala, superando as análises manuais.

4.2. Considerações Éticas e Legais da Classificação

É necessária, ainda, uma reflexão crítica sobre as implicações éticas e legais de se classificar portais de notícias. Afinal, a atribuição de "assinaturas editoriais" poderia ser interpre-

tada de forma pejorativa ou mesmo gerar responsabilidade legal? Esta é uma preocupação central em estudos de mídia e PLN.

Do ponto de vista **legal**, o risco de responsabilização é mitigado por três pilares centrais deste trabalho. Primeiro, a pesquisa se ampara na **liberdade de pesquisa e de expressão**, garantias constitucionais, não fazendo acusações difamatórias, mas apresentando achados empíricos. Segundo, a metodologia é **transparente e replicável**; o estudo não afirma que um portal é "melhor" ou "pior", mas sim que ele é "diferenciável" com base em dados quantitativos. Terceiro, os dados utilizados (títulos de notícias) são **informações publicamente disponíveis**, não havendo violação de privacidade.

Duas leis principais de dados do Brasil podem ser citadas neste contexto. A **Lei de Acesso à Informação (LAI - Lei 12.527/2011)** não se aplica diretamente, pois regula o acesso a dados de entidades públicas, e os portais analisados são empresas privadas. Já a **Lei Geral de Proteção de Dados (LGPD - Lei 13.709/2018)** foca na proteção de dados pessoais de indivíduos. Este trabalho é aderente à LGPD, pois o objeto de análise não são os dados pessoais dos jornalistas ou leitores, mas sim o produto editorial da pessoa jurídica (o portal). Mesmo na limitação discutida sobre "artigos de opinião", o foco da análise permanece no texto público, e não no indivíduo que o escreveu.

Do ponto de vista **ético**, a principal "ameaça" não é legal, mas sim o **risco da super-simplificação**. Um leitor leigo poderia interpretar os resultados (ex: G1 focado em "polícia") como uma definição completa do portal, o que seria uma redução injusta. A responsabilidade ética do pesquisador, aqui cumprida, é delimitar claramente o escopo do trabalho (apenas títulos, apenas período eleitoral), como feito na Seção de Limitações.

Finalmente, a classificação não deve ser vista como uma ameaça, mas como uma contribuição à **literacia midiática**. Ao demonstrar objetivamente que existem diferenças estilísticas, este trabalho fornece ferramentas para que a sociedade compreenda melhor a pluralidade da mídia e o funcionamento das "bolhas informativas", dialogando diretamente com o **ODS 16 (Paz, Justiça e Instituições Eficazes)** da ONU, que visa assegurar o acesso público à informação.

5. Trabalhos Relacionados

A análise de ideologia apresentada no texto de Porfírio (2024) dialoga diretamente com os objetivos do presente trabalho ao fornecer uma base teórica para compreender como os sistemas ideológicos se manifestam e influenciam os discursos em diversos contextos, incluindo o jornalístico [Porfírio 2024]. A visão crítica de ideologia, por exemplo, permite observar como as representações de mundo promovidas por diferentes portais de notícias podem estar alinhadas a interesses de grupos específicos, moldando percepções coletivas e legitimando narrativas hegemônicas.

Essa perspectiva é essencial para a análise de sentimento aplicada aos títulos de notícias, pois ajuda a identificar não apenas padrões linguísticos, mas também os vieses ideológicos que permeiam os textos. A proposta de compreender ideologia como um instrumento que organiza ou manipula ideias, dependendo do contexto, fornece um enquadramento teórico que enriquece a abordagem metodológica deste trabalho, contribuindo para a detecção automática do viés presente nas mensagens veiculadas pelos portais.

O estudo de Fuks & Marques (2022) e o Relatório de Edelman Trust Barome-

ter contribuem significativamente para o entendimento da polarização política no Brasil ao diferenciar as dimensões afetiva e ideológica desse fenômeno, destacando como ele se manifesta de maneira assimétrica e contextual no país [Fuks and Marques 2022, Brasil 2023]. Essa análise fornece um pano de fundo essencial para o desenvolvimento de métodos automatizados de análise de viés em portais de notícias, como o proposto neste trabalho.

Enquanto Fuks & Marques destacam que a polarização no Brasil possui características próprias — como a predominância do antagonismo afetivo e a fragmentação partidária —, nosso foco recai sobre como esses elementos são refletidos na cobertura jornalística e nas narrativas midiáticas. A crescente radicalização ideológica, especialmente no espectro político à direita, e o papel central das lideranças políticas na definição de polarizações no Brasil são aspectos que podem influenciar diretamente o viés percebido nos títulos de notícias e no alinhamento ideológico implícito de determinados portais.

Dessa forma, o presente estudo avança ao propor ferramentas para a detecção automática desses vieses, buscando promover maior transparência e acessibilidade das informações ao público. A análise de Fuks & Marques serve, portanto, como uma base teórica para compreender como o cenário político brasileiro pode moldar os discursos veiculados pela mídia, indicando a relevância de se observar o alinhamento ideológico no contexto das notícias como um reflexo da polarização existente na sociedade.

Ribeiro et al. (2018) propõem duas abordagens complementares para verificar o viés em notícias. A primeira envolve a identificação indireta de viés com base no perfil dos leitores, assumindo que a notícia reflete o alinhamento ideológico de seus consumidores. A segunda abordagem propõe a identificação de viés por meio da análise do conteúdo textual, sendo adotada nesta pesquisa devido à independência de fontes externas. Enquanto análises manuais são mais precisas, soluções automáticas, especialmente aquelas baseadas em Processamento de Linguagem Natural (PLN) e Aprendizagem Profunda, oferecem escalabilidade. No entanto, limitações persistem, como a predominância de abordagens baseadas em análise de sentimentos, que podem falhar na detecção de enquadramento em cenários complexos. Além disso, há uma lacuna na literatura em relação à análise de vieses midiáticos em notícias de múltiplos idiomas e na identificação de vieses relacionados a atributos definidos pelo usuário, como avaliações específicas de candidatos em eleições.

Além da análise de sentimento, outra vertente de pesquisa para identificar características da mídia é a **estilometria computacional**, que se dedica à análise quantitativa do estilo de escrita. Essa área é frequentemente aplicada na tarefa de atribuição de autoria, mas seus princípios são diretamente relevantes para a classificação de fontes de notícias, onde o "estilo" de um portal funciona como uma "impressão digital". Trabalhos como o de Koppel, Schler & Argamon (2009) estabeleceram as bases para o uso de características textuais na identificação de autores [Koppel et al. 2009]. Mais recentemente, com o advento de modelos de linguagem como o BERT, a capacidade de capturar nuances estilísticas se tornou ainda mais poderosa. Estudos como o de Zellers et al. (2019) demonstram o potencial de modelos Transformer para diferenciar textos gerados por humanos e por máquinas, uma tarefa que também depende da identificação de padrões estilísticos sutis [Zellers et al. 2019]. O presente trabalho se alinha a essa vertente, aplicando um modelo BERT especializado em português para a tarefa de classificar a fonte da notícia, tratando cada portal como um "autor" com um estilo editorial distinto.

6. Limitações e Ameaças à Validade

Todo trabalho de pesquisa possui limitações que devem ser transparentemente declaradas. A primeira ameaça à validade externa é o **viés temporal do corpus**. Os dados foram coletados during um período eleitoral de dois meses. Esse contexto específico certamente influenciou os tópicos e o tom da cobertura, como o foco em candidatos e pesquisas. Os padrões de vocabulário aqui identificados podem não ser generalizáveis para a cobertura jornalística dos mesmos portais em períodos não eleitorais.

A segunda limitação reside no **escopo do texto analisado**. O estudo se concentrou exclusivamente nos títulos das notícias. Embora um título seja uma síntese poderosa do conteúdo, ele não captura a complexidade total do argumento, o enquadramento da matéria ou as fontes citadas no corpo do texto. Uma análise mais profunda, incluindo o conteúdo completo, poderia revelar padrões de viés mais sutis e complexos.

Uma terceira limitação refere-se à **mistura de gêneros textuais** no corpus. Durante o processo de raspagem, não foi realizada uma distinção entre notícias factuais (que seguem a linha editorial do portal) e artigos de opinião (que refletem a visão particular de colunistas). A inclusão desses artigos no conjunto de dados representa uma ameaça direta à validade da conclusão, pois o modelo de classificação pode ter aprendido a identificar o estilo de um autor específico (colunista) em vez da "assinatura" editorial do portal como um todo. Esse ruído na composição do corpus pode ter dificultado a diferenciação entre os portais.

Finalmente, a **mudança de escopo metodológico** representa uma limitação. A dificuldade inicial em prosseguir com a análise de sentimento, devido à escassez de dados anotados, impediu la exploração quantitativa do viés sentimental, que era um dos objetivos iniciais. Embora a pivotação para a classificação de portais tenha sido bem-sucedida, a questão do sentimento permanece como um tópico para trabalhos futuros.

Adicionalmente, uma reflexão crítica sobre o processo revela pontos que poderiam ser abordados de maneira diferente. Com o conhecimento do resultado final, onde o modelo `neuralmind/bert-base-portuguese-cased` se mostrou vastamente superior, uma estratégia mais eficiente teria sido focar os experimentos exclusivamente em modelos pré-treinados para o português do Brasil desde o início, em vez de avaliar modelos multilíngues. Outro ponto de reflexão reside na etapa de coleta de dados. Sabendo da disparidade de volume entre os portais, um esforço poderia ter sido direcionado para balancear melhor o corpus durante a raspagem, talvez estendendo o período de coleta para os portais com menor volume, em vez de aplicar uma subamostragem posterior. Tais ajustes poderiam, potencialmente, aprimorar ainda mais o desempenho do classificador, especialmente para as classes com menor suporte.

7. Trabalhos Futuros

A partir dos resultados e das limitações identificadas neste estudo, emergem diversas oportunidades para investigações futuras. A seguir, detalham-se três caminhos principais:

- **Análise do Texto Completo:** Enquanto este trabalho focou nos títulos, uma expansão natural seria aplicar as mesmas técnicas de classificação e análise no corpo completo das notícias. Isso permitiria capturar nuances mais sutis, como o enquadramento (framing) dos fatos, as fontes citadas e os argumentos desenvolvidos, potencialmente revelando padrões de viés que não são evidentes apenas no título.

- **Expansão Temporal e Temática do Corpus:** O corpus atual é focado em um período eleitoral específico. Seria de grande valia coletar dados dos mesmos portais em períodos não eleitorais para verificar se as "assinaturas" estilísticas são consistentes ao longo do tempo ou se variam drasticamente com o contexto noticioso. A inclusão de outras editorias (ex: Economia, Cotidiano) também testaria a generalização dos modelos.
- **Interpretabilidade e Análise de Viés Explícito:** Embora o modelo tenha classificado as fontes com sucesso, ele opera como uma "caixa-preta". A aplicação de técnicas de interpretabilidade, como LIME (Local Interpretable Model-agnostic Explanations) ou SHAP (SHapley Additive exPlanations), seria um avanço significativo. Tais métodos poderiam revelar exatamente quais palavras ou frases em um título mais influenciaram a decisão do modelo, transformando a detecção de um padrão implícito em uma análise explícita dos termos que caracterizam o viés ou o estilo de cada portal.

8. Conclusão

Este trabalho validou a hipótese de que os títulos de notícias de grandes portais brasileiros possuem "assinaturas" editoriais e temáticas distintas, demonstrando empiricamente que tais diferenças são robustas o suficiente para permitir a classificação de sua origem com alta precisão. A principal evidência desta contribuição foi o desempenho do modelo `neuralmind/bert-base-portuguese-cased`, especializado em português do Brasil, que alcançou um F1-Score médio de **82,5%**. Este sucesso não apenas corrobora a tese central, mas também ressalta a importância de utilizar modelos de linguagem específicos para o idioma em análise, capazes de capturar suas particularidades.

As análises descritivas e os resultados da classificação se mostraram complementares: as diferenças de vocabulário e estilo — como o foco da CNN em dados ou do G1 em segurança — não são apenas percepções qualitativas, mas padrões concretos que sustentam o desempenho do modelo. Mais do que um exercício técnico, o trabalho oferece um método empírico para quantificar "assinaturas editoriais", contribuindo para a área de Processamento de Linguagem Natural com um caso de uso validado e oferecendo à sociedade uma ferramenta para futuras investigações sobre a pluralidade da mídia, o viés de enquadramento e a formação de bolhas informativas no ecossistema de notícias brasileiro.

A pesquisa também documentou uma jornada metodológica realista, na qual os desafios práticos na coleta de dados para a análise de sentimento levaram a um redirecionamento estratégico do foco experimental. Esta adaptabilidade, por si só, fortalece o relato da pesquisa como um reflexo fiel do processo científico.

Do ponto de vista do aprendizado, o desenvolvimento deste TCC proporcionou uma imersão prática e teórica em Ciência de Dados e PLN. Refletindo sobre o processo, é notável que o **trabalho árduo** da pesquisa não se concentrou onde era inicialmente esperado. Graças à experiência profissional prévia, a etapa de engenharia de dados e coleta (web scraping) foi uma tarefa direta. O verdadeiro desafio metodológico e o que demandou meses de esforço foi a **fase experimental**. O treinamento e a otimização dos modelos Transformer foram um processo iterativo e intenso, exigindo dezenas de ciclos de calibração de hiperparâmetros e testes com múltiplas arquiteturas para que o desempenho final de 82,5% fosse alcançado. O desafio inicial com a análise de sentimento e a

subsequente mudança de escopo ensinaram uma lição valiosa sobre a adaptabilidade na pesquisa, onde os obstáculos práticos devem guiar a reformulação de hipóteses. Acima de tudo, o projeto aprofundou a capacidade de análise crítica sobre o ecossistema de mídia, transformando um problema socialmente relevante em um desafio de modelagem computacional bem definido.

Por fim, cabe registrar que a elaboração deste documento contou com o auxílio de um modelo de linguagem de inteligência artificial (Gemini, do Google) como ferramenta de apoio. O uso da tecnologia foi direcionado para tarefas específicas de assistência à escrita, como o aprimoramento da clareza e coesão textual, a revisão de trechos para adequação à norma culta, a formatação de código em LaTeX e a estruturação de parágrafos para expandir ideias, sempre a partir de diretrizes e conteúdos conceituais fornecidos pelo autor. É fundamental destacar que a responsabilidade integral pelo conteúdo, pela originalidade das ideias, pelas análises e pelas conclusões aqui apresentadas é inteiramente do autor humano, que supervisionou, validou criticamente e assumiu a autoria de todas as informações e textos gerados.

Referências

Brasil, E. (2023). Edelman trust barometer - relatório nacional. Acessado em 26 de novembro de 2024. Disponível em `https://www.edelman.com.br/sites/g/files/aatuss291/files/2023-04/2023%20Edelman%20Trust%20Barometer_Brazil%20Report_POR%20%281%29_0.pdf`.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Fuks, M. and Marques, P. H. (2022). Polarização e contexto: medindo e explicando a polarização política no brasil. *Opinião Pública*, 28(3):560–593.

Koppel, M., Schler, J., and Argamon, S. (2009). Computational methods in authorship attribution. In *Journal of the American Society for Information Science and Technology*, volume 60, pages 9–26. Wiley Online Library.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Porfírio, F. (2024). Ideologias: conceitos, tipos, exemplos. Acessado em 26 de novembro de 2024. Disponível em `https://brasilescola.uol.com.br/filosofia/ideologia.htm`.

Sunstein, C. R. (2001). *Echo Chambers: Bush v. Gore, Impeachment, and Beyond*. Princeton University Press, Princeton, NJ.

YouGov Profiles (2024). Brasil: Em 2024, a tv ainda é a top mídia de notícias. Acessado em 26 de novembro de 2024. Disponível em `https://business.yougov.com/pt/content/50226-brasil-em-2024-a-tv-ainda-e-a-top-midia-de-noticias`.

Zellers, R., Holtzman, A., Rashkin, H., Bisk, Y., Farhadi, A., Roesner, F., and Choi, Y. (2019). Defending against neural fake news. In *Advances in Neural Information Processing Systems*, volume 32.