Levantamento e Análise Qualitativa de Bases de Dados de Fake News em Português

Juliana Karla de C. M. Baracho[1], **Lucas A. Lisboa**[2], **Roberta Vilhena V. Lopes**[1]

[1]Instituto de Computação – Universidade Federal de Alagoas (UFAL)
Maceió – AL – Brasil

[2]Núcleo de Tecnologia da Informação – Universidade Federal de Alagoas (UFAL)
Maceió – AL – Brasil

jkcm@ic.ufal.br, lucas.lisboa@nti.ufal.br, rvvl@ic.ufal.br

Abstract. *The spread of fake news on social media is an increasingly serious problem, directly influencing public opinion. Artificial intelligence algorithms are used to combat them, but their effectiveness depends on the quality of the databases. In this context, there are still a limited number of databases available in the Portuguese language. Therefore, this study conducted a survey of fake news datasets in Portuguese, focusing specifically on the Brazilian context. Among the findings, the limited number of instances in the identified datasets stands out when compared to those in English.*

Resumo. *A disseminação de fake news nas redes sociais é um problema cada vez maior, influenciando diretamente a opinião pública. Para combatê-las, algoritmos de inteligência artificial são usados, mas a eficácia deles depende da qualidade das bases de dados. Nesse sentido, há ainda um número baixo de bases em língua portuguesa. Assim, este trabalho realizou um levantamento das bases de notícias falsas em português, com enfoque no contexto brasileiro. Dentre os achados, destaca-se a quantidade reduzida de instâncias nas bases encontradas quando comparadas às de língua inglesa.*

1. Introdução

Nos últimos anos, as redes sociais e as mídias digitais se tornaram os principais meios de informação para a maior parte da população. No entanto, essa ampla disseminação de conteúdos também facilitou a propagação de *fake news*. Estudos como o de [Garcia et al. 2024] demonstram que notícias falsas, quando divulgadas intencionalmente, podem influenciar a opinião pública e até comprometer a estabilidade social e política de um país.

Para mitigar esse problema, algoritmos de inteligência artificial têm sido amplamente utilizados na identificação de desinformação. No entanto, a eficácia dessas soluções depende diretamente da qualidade dos dados utilizados no treinamento dos modelos. Se as bases de dados forem limitadas ou desatualizadas, os resultados obtidos podem ser imprecisos e pouco eficazes.

No contexto da língua portuguesa, as bases de dados disponíveis ainda apresentam desafios quando comparadas às suas equivalentes em inglês [Garcia et al. 2024],

[Santos et al. 2018]. Diante disso, este artigo tem como objetivo analisar as bases de dados existentes em português, compará-las com as internacionais e identificar pontos que podem ser aprimorados para fortalecer a detecção de *fake news* de maneira mais eficiente.

A importância desta análise se justifica pelo fato de que, embora o português seja um dos idiomas mais falados do mundo em número de nativos [Yibo 2024], sua presença como língua de comunicação científica global é restrita, o que limita o desenvolvimento de recursos tecnológicos, como bases de dados para detecção de conteúdos enganosos, especialmente em contextos críticos como processos eleitorais, políticas públicas e crises sanitárias, que são frequentemente alvos desse tipo de campanha.

Este estudo oferece contribuições relevantes ao realizar uma avaliação das bases de dados existentes em português, produzidas no Brasil, estabelecendo comparações com conjuntos internacionais que permitem identificar lacunas e oportunidades de aprimoramento, ao propor a ampliação de volume, diversidade temática e incorporação de elementos multimodais. Além disso, as reflexões apresentadas servem como base para pesquisas futuras, destacando a necessidade crítica de conjuntos de dados balanceados e abrangentes que possam sustentar modelos de detecção com maior precisão e capacidade de generalização.

Dessa forma, os resultados alcançados não apenas mapeiam o cenário atual de recursos para detecção de *fake news* em português, mas também apontam sugestões para a construção de bases de dados mais completas e representativas ou melhora das existentes. Esses avanços são fundamentais para fortalecer as iniciativas de combate à desinformação em contextos em que a língua portuguesa é predominante, contribuindo para a proteção do debate público e dos processos democráticos.

2. Fundamentação Teórica

2.1. Definição de Fake News

A definição utilizada aqui foi baseada em [Villela et al. 2023], que descreve *fake news* como conteúdos que contêm informações falsas ou manipuladas, com potencial de causar impactos negativos na sociedade. Nesse sentido, as bases de dados foram selecionadas com base em sua capacidade de representar diferentes tipos de desinformação, como notícias falsas sobre política, saúde e economia.

2.2. Métodos de Detecção de Fake News

Existem diferentes maneiras de identificar *fake news* com o apoio da inteligência artificial. Um dos caminhos mais utilizados é o uso de algoritmos que aprendem a distinguir conteúdos verdadeiros e falsos a partir de exemplos já classificados. Esse tipo de abordagem funciona bem quando há bases de dados com informações organizadas em duas categorias (verdadeiro ou falso) e com número equilibrado de exemplos em cada uma, como é o caso das bases Fake.BR e FakeRecogna [Garcia et al. 2024, Santos et al. 2018].

A qualidade dessas bases depende diretamente do processo de checagem das notícias, conhecido como *fact-checking*. Esse método consiste na verificação manual das informações por jornalistas especializados, que consultam fontes oficiais, especialistas e documentos confiáveis para determinar a veracidade de uma afirmação. As notícias classificadas como verdadeiras ou falsas por esses profissionais servem como referência para

treinar os modelos de inteligência artificial. No Brasil, agências como Lupa, Aos Fatos e Boatos.org desempenham papel central nesse processo [Macedo et al. 2022].

Nos últimos anos, métodos mais modernos têm sido usados para aprimorar esse processo. São técnicas que tentam compreender melhor o conteúdo das notícias, observando não só palavras isoladas, mas também o sentido das frases e a forma como as informações são organizadas. Esses modelos conseguem identificar padrões mais sutis e complexos [Farhangian et al. 2024, Villela et al. 2023], mas exigem grandes volumes de dados para funcionar bem — algo que ainda é limitado nas bases em português, conforme estudos recentes que analisam o cenário brasileiro [Garcia et al. 2024].

Além do conteúdo textual, há estratégias que buscam analisar outros elementos das notícias, como imagens, vídeos e até áudios. Esse tipo de abordagem, chamado de multimodal, é especialmente útil porque muitas *fake news* usam recursos visuais para enganar o leitor [Boididou et al. 2018, Macedo et al. 2022]. No entanto, no Brasil, a maioria das bases disponíveis traz apenas o texto das notícias, o que dificulta a aplicação dessas soluções mais completas [Macedo et al. 2022, Irís and da Silva 2024].

Assim, tanto os métodos mais simples, quanto os mais sofisticados podem ser aplicados às bases em português. No entanto, o tamanho reduzido dessas bases e a ausência de outros tipos de mídia ainda representam obstáculos importantes para o avanço na detecção de *fake news* no país [Garcia et al. 2024, Villela et al. 2023].

2.3. Definição das Categorias

Neste trabalho, as categorias utilizadas para classificar as notícias foram definidas com base nas práticas adotadas pelas principais bases de dados em português para detecção de *fake news*, sendo as mais relevantes:

- **Verdadeira:** Informação comprovadamente correta, conforme verificação realizada por agências de checagem de fatos ou extraída de fontes confiáveis. Essa categoria é amplamente utilizada na maioria das bases brasileiras, que adotam uma classificação binária para simplificar a tarefa de detecção.
- **Falsa:** Informação comprovadamente falsa, segundo as agências de checagem de fatos. Assim como a categoria Verdadeira, é parte do esquema binário predominante em bases como Fake.br e FakeRecogna.

Além do esquema binário, uma base em português adota uma classificação mais detalhada para capturar nuances na veracidade das notícias. A categoria intermediária identificada no estudo é a seguinte:

- **Meia-verdadeira:** Categoria presente em bases como o FACTCK.BR, que reúne informações que apresentam elementos verdadeiros e falsos, ou que mesclam fatos corretos com incorretos. Essa categoria possibilita uma análise mais granular da desinformação, refletindo a complexidade do conteúdo encontrado no ambiente digital brasileiro.

Dessa forma, a definição das categorias adotadas neste trabalho considera tanto o padrão binário amplamente utilizado nas principais bases de dados em português, quanto a introdução de categorias intermediárias, conforme observado na FACTCK.BR, permitindo uma análise mais detalhada da veracidade das notícias.

2.4. Bases de Dados em Inglês

A partir de 2016, houve um aumento significativo nos estudos acerca do fenômeno das notícias falsas [Farhangian et al. 2024]. Com isso, diversas bases surgiram para possibilitar a análise das notícias falsas, bem como serem usadas para treinamento e validação de modelos de detecção de *fake news*. Dentre os trabalhos que realizaram o levantamento dessas bases, há destaque para o de [D'ulizia et al. 2021], em que 27 *datasets* são analisados. Dentre eles, alguns são amplamente utilizados devido à sua diversidade e tamanho, incluindo:

- **CREDBANK**: Possui mais de 60 milhões de tweets, é uma das maiores bases disponíveis para análises temporais e estudos sobre a propagação social de *fake news*. Ela aborda temas como política internacional, eventos globais e desastres naturais, sendo amplamente utilizada para entender como a desinformação se espalha nas redes sociais [Mitra and Gilbert 2015];
- **FEVER**: Focada na verificação de fatos, abrangendo declarações públicas e notícias. Ela é balanceada e contém três classes, o que facilita a análise detalhada da desinformação em diferentes contextos sociais [Thorne et al. 2018];
- **LIAR**: Reconhecida por sua granularidade nas classificações. Contém declarações políticas rotuladas em seis categorias, permitindo uma análise profunda da desinformação no contexto político [Wang 2017];
- **NELA-GT-2020**: Uma das maiores para verificação de fatos, cobrindo uma ampla gama de tópicos como política, saúde e tecnologia. Ela contém uma vasta coleção de artigos e notícias rotulados por veracidade [Gruppi et al. 2021];
- **Verification Corpus**: Inclui texto, imagem e vídeo, o que é essencial para a detecção de *deepfakes* e notícias falsas multimodais. Ela é amplamente utilizado em estudos que buscam entender como a desinformação se propaga em diferentes formatos de mídia nos tópicos sociais [Boididou et al. 2018];
- **YELP**: Extraída da plataforma Yelp e é amplamente utilizada para estudos acadêmicos e aplicações práticas, como análise de sentimentos, classificação de texto e estudos sobre comportamento do consumidor. Está disponível em diferentes versões, a depender do propósito da pesquisa [Barbado et al. 2019].

Essas bases apresentam características únicas e capacidade de representar diferentes tipos de desinformação, o que é crucial para o desenvolvimento de modelos eficazes de detecção de *fake news*. A comparação com bases em português pode ajudar a identificar áreas de melhoria para as bases nacionais. Nesse sentido, para este trabalho, tais bases foram escolhidas para efeitos comparativos com as bases em português, sendo adotados quatro critérios principais, alinhados com os objetivos do estudo:

1. **Diversidade metodológica:** Cada base selecionada representa uma abordagem distinta para a detecção de desinformação, desde a análise de credibilidade em redes sociais (CREDBANK) até a verificação factual estruturada (FEVER) e a classificação multi-nível de veracidade (LIAR). Essa variedade permite uma avaliação abrangente das diferentes estratégias de detecção;
2. **Amplitude de formatos e plataformas:** Os conjuntos abrangem desde tweets curtos (CREDBANK) até artigos completos (NELA-GT-2020), incluindo também avaliações de consumidores (YELP) e conteúdo multimídia (Verification Corpus). Essa diversidade reflete os múltiplos formatos em que a desinformação se manifesta atualmente;

3. **Reconhecimento acadêmico:** Todos os *datasets* selecionados possuem ampla adoção na comunidade científica, garantindo que nossa análise se baseie em recursos consagrados e metodologicamente validados.

3. Metodologia

A metodologia utilizada neste trabalho consistiu em uma seleção cuidadosa de bases de dados em português para detecção de *fake news*, considerando fatores como tamanho, balanceamento entre classes, multimodalidade e diversidade temática. Em seguida, realizou-se uma análise comparativa dessas bases com as principais bases em inglês, utilizando os levantamentos existentes na literatura. Por fim, discutiu-se as implicações dos resultados para identificar áreas de melhoria nas bases nacionais em português.

3.1. Identificação das Bases em Português Brasileiro

Foram adotadas as seguintes *strings* de busca: *"base de dados fake news português"* e *"dataset desinformação Brasil"*. Com isso, tais *strings* foram aplicadas em repositórios de artigos científicos, tais como Google Scholar e Scopus, em repositórios de códigos, como Github e Kaggle, e demais indexadores similares. Para cada base encontrada, realizamos uma busca específica nos indexadores de artigos utilizando o nome do conjunto de dados (ex.: "Fake.BR *dataset*") para localizar artigos científicos que tivessem utilizado essas bases, o que nos permitiu validar sua relevância acadêmica.

3.2. Critérios de Análise

Para a análise comparativa, foram utilizados os seguintes critérios:

- **Tamanho:** Número de amostras;
- **Balanceamento entre Classes:** Proporção entre notícias reais e falsas;
- **Diversidade Temática:** Variedade de tópicos abordados;
- **Multimodalidade:** Inclusão de texto, imagens, vídeos ou áudios.

Esses critérios permitem verificar a abrangência das bases e a capacidade delas de generalização. Assim, com base nos resultados da análise comparativa, foram propostos pontos de melhoria para o desenvolvimento de novas bases.

4. Resultados e Discussões

O levantamento realizado identificou as principais bases de dados em português voltadas à detecção de *fake news*. Esta seção apresenta uma análise integrada dessas bases, estruturada em torno dos seguintes eixos: panorama geral e tabela-resumo; características das bases; balanceamento das classes; temas cobertos; e multimodalidade.

4.1. Panorama Geral e Tabela-Resumo das Bases

As bases de dados em português para detecção de *fake news* variam significativamente em volume, metodologia de coleta, temas abordados e formatos disponíveis. A Tabela 1 sintetiza as principais características das bases analisadas.

Tabela 1. Bases em Português Identificadas

Nome	Temas abordados	Classes	Número de instâncias	Fonte	Tipo de conteúdo
Base COVID-19	COVID-19	Falsa	2.808 (21,46%)	Boatos, Lupa	Texto
		Verdadeira	10.279 (78,54%)	G1	
Desinfopedia	Política Brasileira, COVID-19, Boatos e outros	Falsa	754 (100%)	Lupa, Aos Fatos e outros	Texto, imagens, vídeos
Factck.BR	Política, economia e saúde	Falsa	943 (72,04%)	Lupa, Aos Fatos e Truco	Texto
		Meia-Verdadeira	246 (18,79%)		
		Verdadeira	120 (9,17%)		
Fake.BR	Política, economia, saúde, TV, celebridades, sociedade, ciência, tecnologia, entre outros	Falsa	3600 (50%)	Diário do Brasil, A Folha do Brasil, The Jornal Brasil e Top Five TV	Texto
		Verdadeira	3600 (50%)	G1, Folha de São Paulo e Estadão	
FakeRecogna	Política, saúde, entretenimento, Brasil, entre outros	Falsa	5951 (50%)	Boatos, Fato ou Fake, E-farsas e outros	Texto
		Verdadeira	5951 (50%)	G1, UOL e Extra	
FakeTrueBR	Política, saúde, economia e cultura popular no Brasil contemporâneo	Falsa	1791 (50%)	Boatos	Texto
		Verdadeira	1791 (50%)	G1 e Folha	

4.2. Descrição das Bases em Português

As bases de dados em português utilizadas para treinar e avaliar modelos de detecção de *fake news* são compostas por textos noticiosos previamente classificados quanto à sua veracidade. Essa classificação é, em grande parte, realizada por meio de verificação manual, um processo conduzido por profissionais de agências de *fact-checking* como Lupa, Aos

Fatos e outras. Nesse procedimento, cada notícia é analisada individualmente com base em fontes confiáveis, documentos oficiais, evidências públicas e, em alguns casos, entrevistas com especialistas, a fim de determinar se o conteúdo é verdadeiro ou falso. Esse trabalho é essencial para o treinamento de modelos supervisionados.

- **Base COVID-19:** Possui 13.087 registros binários relacionados exclusivamente à pandemia de COVID-19. Os dados são textuais, coletados de portais de notícias e agências como Lupa e G1, com coleta por meio de *crawler* [TIGRE et al. 2023].
- **Desinfopedia:** Contém 754 notícias com classificação única, focando em política e coronavírus. Os dados são textuais e foram coletadas e rotuladas manualmente por especialistas de agências de verificação, como Lupa e Aos Fatos [Garcia et al. 2024].
- **Factck.BR:** Tem 1.309 instâncias e apresenta uma classificação multiclasse com 3 categorias, incluindo "meia-verdadeira". Os temas principais são política, economia e saúde. Cada notícia foi previamente checada por jornalistas especializados em verificação, com base em evidências públicas, documentos oficiais e fontes confiáveis [Moreno and Bressan 2019].
- **Fake.BR:** Possui 7.200 instâncias com classificação binária (verdadeiro/falso), focando em temas de política e saúde. Os textos foram coletados de agências como a Lupa, em que cada notícia foi avaliada por verificadores humanos com base em documentos, dados oficiais e reportagens confiáveis. Essa base é amplamente utilizada para treinamento de modelos supervisionados em português [Santos et al. 2018].
- **FakeRecogna:** Conta com 11.902 exemplos balanceados entre notícias verdadeiras e falsas, abrangendo temas como política, saúde e economia. As notícias foram extraídas de diversas fontes jornalísticas e plataformas digitais. A veracidade do conteúdo foi determinada por profissionais de *fact-checking*, que utilizaram processos rigorosos de checagem baseados em evidências, contexto e fontes primárias [Garcia et al. 2024].
- **FakeTrueBR:** Apresenta 3.582 exemplos balanceados entre notícias verdadeiras e falsas, abordando temas como política, saúde e cultura. Os dados foram coletados de portais de notícias e redes sociais, e cada item foi verificado manualmente por especialistas com base em análise documental, registros públicos e fontes confiáveis [Chavarro et al. 2023].

4.3. Características Gerais das Bases

O volume das bases nacionais voltadas à detecção de desinformação é consideravelmente inferior ao observado em bases internacionais, com a maioria apresentando menos de 10 mil registros, conforme é possível observar na Figura 1, na qual estão ilustrados os volumes das bases em português. Entre as iniciativas brasileiras, destacam-se FakeRecogna (11.902 amostras) e Fake.BR (7.200 registros), que apresentam um número maior de instâncias em comparação às demais, ainda assim distantes de bases como o CREDBANK (60 milhões de tweets) e a NELA-GT-2020 (mais de 1 milhão de artigos) [Mitra and Gilbert 2015, Gruppi et al. 2021]. Essa limitação é crítica, pois 89% dos brasileiros já acreditaram em *fake news*, com 73,7% dos casos circulando pelo WhatsApp [Agência Brasil 2024, Fiocruz 2024]. Por outro lado, bases como Factck.BR e Desinfopedia possuem menos dados disponíveis, o que impacta diretamente na capacidade de uso em larga escala [Garcia et al. 2024, Moreno and Bressan 2019].

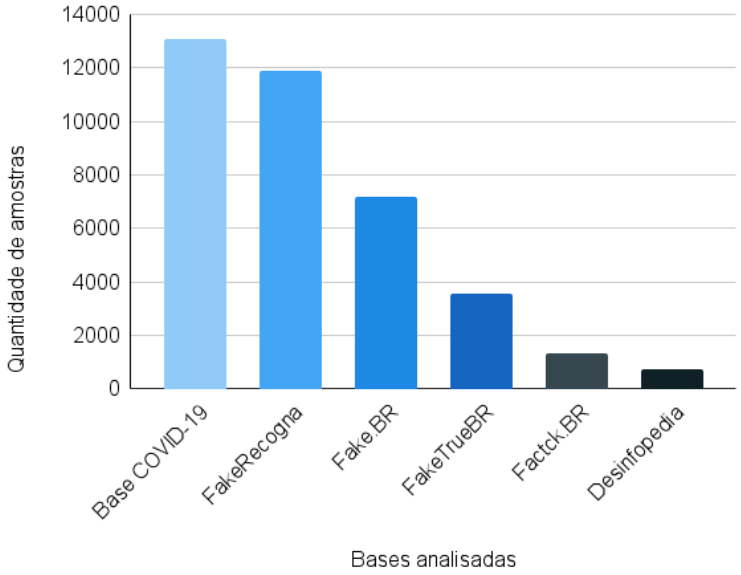

Figura 1. Gráfico comparativo do volume das bases de dados brasileiras utilizadas para detecção de *fake news*.

4.4. Balanceamento das Classes

A distribuição entre notícias falsas e verdadeiras varia consideravelmente entre as bases de dados nacionais voltadas à detecção de desinformação, o que influencia diretamente o desempenho e a imparcialidade dos modelos de aprendizado. Bases como Fake.BR, FakeRecogna e FakeTrueBR apresentam conjuntos equilibrados — por exemplo, FakeRecogna possui 5.951 notícias falsas e 5.951 verdadeiras, o que contribui para o treinamento de modelos mais robustos e com menor propensão a vieses [Garcia et al. 2024, Chavarro et al. 2023].

Por outro lado, bases como Factck.BR e Base COVID-19 são significativamente desbalanceadas, com uma forte predominância de uma das classes. Esse desbalanceamento pode prejudicar a capacidade dos modelos de identificar corretamente os exemplos minoritários, comprometendo a eficácia da classificação em cenários reais [TIGRE et al. 2023, Moreno and Bressan 2019].

Entre as bases internacionais, observa-se cenário semelhante: enquanto o LIAR Dataset mantém uma distribuição relativamente equilibrada entre suas categorias, o CREDBANK apresenta desbalanceamento mesmo com um grande volume de dados, o que impõe desafios adicionais no treinamento de modelos confiáveis [Mitra and Gilbert 2015, Wang 2017].

Essas diferenças reforçam a importância de considerar a distribuição das classes como um critério central no desenvolvimento e na escolha de bases para estudos em detecção de desinformação, especialmente quando se busca generalização e imparcialidade nos resultados.

4.5. Temas Cobertos

A análise temática das bases nacionais revela uma forte concentração em tópicos como política, saúde — com destaque para a COVID-19 — e economia. Por exemplo, a Desinfopedia foca principalmente em Política Brasileira, Coronavírus e Eleições, enquanto a FakeTrueBR inclui temas relacionados à cultura popular, embora de forma ainda limitada [Garcia et al. 2024, Chavarro et al. 2023]. Essa concentração reflete o alinhamento das bases com crises recentes que impactaram o Brasil, como a pandemia e os processos eleitorais [Santos et al. 2018, Moreno and Bressan 2019].

No entanto, temas importantes como ciência, tecnologia, meio ambiente e segurança pública permanecem sub-representados, o que restringe a capacidade dos modelos treinados nessas bases de generalizar para diferentes domínios da desinformação [Farhangian et al. 2024, Villela et al. 2023]. Estudos indicam que segurança pública e meio ambiente correspondem a cerca de 10,5% das notícias falsas que circulam no país, demonstrando a relevância dessas áreas [D'ulizia et al. 2021, Agência Brasil 2024], mas elas ainda não são suficientemente contempladas nas bases existentes.

Em contraste, bases internacionais em inglês, como a NELA-GT-2020 e a FEVER Dataset, apresentam uma variedade temática muito maior, abrangendo diversos setores e assuntos, o que favorece o desenvolvimento de modelos mais generalizáveis e aplicáveis a múltiplos contextos [Gruppi et al. 2021, Thorne et al. 2018]. A atual restrição temática das bases nacionais limita a eficácia das ferramentas brasileiras no combate à desinformação em áreas emergentes e relevantes, como ciência, tecnologia e cultura, que são especialmente importantes para o contexto social e informacional do país [Farhangian et al. 2024, Villela et al. 2023].

4.6. Multimodalidade

A multimodalidade representa um desafio significativo para as bases nacionais voltadas à detecção de desinformação. A grande maioria dessas bases em português é composta exclusivamente por dados textuais, sem a incorporação de imagens, vídeos ou áudios. Essa limitação restringe a capacidade de identificar *fake news* desinformações disseminadas por meio de memes, *deepfakes* e outros formatos visuais ou audiovisuais, que têm ganhado crescente relevância e impacto na sociedade [Macedo et al. 2022, Irís and da Silva 2024].

Esse aspecto é especialmente preocupante considerando que aproximadamente 35% das mensagens falsas analisadas pela Fiocruz continham fotos ou vídeos manipulados, evidenciando a necessidade de bases que abranjam múltiplos formatos [Fiocruz 2024]. Em contraste, bases internacionais como a Verification Corpus já incorporam diferentes modalidades, incluindo texto, imagem e vídeo, o que possibilita o desenvolvimento de modelos multimodais mais robustos e eficazes na detecção de *deepfakes* e memes enganosos [Boididou et al. 2018].

A ausência da multimodalidade nas bases nacionais dificulta o avanço de soluções capazes de lidar com a complexidade atual da desinformação, que frequentemente se manifesta em múltiplos formatos. Reconhecer a importância do conteúdo visual e audiovisual para a percepção pública reforça a necessidade de investir na criação e expansão de bases multimodais no contexto brasileiro [Macedo et al. 2022, Irís and da Silva 2024].

4.7. Limitações Verificadas e Implicações Sociais e Políticas

A coleta dos dados nas bases em português que foram estudadas neste artigo ocorre de diferentes formas, incluindo a extração automatizada de conteúdos de portais de notícias e a seleção baseada em informações verificadas por agências de *fact-checking*. No cenário nacional, observa-se uma forte concentração em torno de um número restrito de fontes, como as agências Lupa, Aos Fatos e Boatos.org. A predominância dessas fontes na construção das bases nacionais pode introduzir vieses, principalmente relacionados à abrangência limitada de verificação. Cada agência possui seus próprios critérios editoriais e tende a priorizar determinados temas ou figuras públicas, o que reduz a diversidade dos dados reunidos. Como consequência, muitos conteúdos verificados acabam sendo repetidos entre as bases, limitando o espectro de tipos de desinformação representados e afetando a capacidade de generalização dos modelos treinados, que aprendem a partir de exemplos que não necessariamente refletem toda a complexidade do cenário informativo brasileiro.

Para ampliar a representatividade e reduzir possíveis enviesamentos, recomenda-se que futuras bases incorporem dados provenientes de múltiplas agências de checagem e veículos de comunicação. Além da Lupa, Aos Fatos e Boatos.org, destacam-se outras iniciativas relevantes, como Agência Pública, Fato ou Fake (Agência Brasil), Projeto Comprova, Agência Truco, Factcheck.org.br, Checamos, É Isso Mesmo?, UOL Confere e E-farsas. Essas fontes atuam em diferentes regiões e com variados focos temáticos, o que contribui para uma cobertura mais ampla e heterogênea [Villela et al. 2023, Chavarro et al. 2023]. Nesse contexto, adotar abordagens mais diversificadas na construção de bases representa uma oportunidade importante para o desenvolvimento de pesquisas mais robustas e representativas no combate à desinformação.

Ademais, as bases nacionais para detecção de *fake news* apresentam limitações que impactam diretamente a eficácia dos modelos desenvolvidos a partir delas. Entre os principais desafios estão o volume reduzido de dados (como pode ser observado na Figura 1), a concentração temática em poucos assuntos, a predominância da seleção manual dos dados — ou seja, a escolha e verificação feitas principalmente por pessoas, o que limita o volume e a diversidade das informações —, e a ausência de multimodalidade. Esses aspectos comprometem a generalização dos modelos e a capacidade de detectar desinformação em formatos diversos. Socialmente, essa limitação temática pode contribuir para o fortalecimento de bolhas de desinformação em áreas pouco representadas, restringindo o alcance das ações de combate. Politicamente, reduz a capacidade de resposta a campanhas coordenadas, especialmente em períodos críticos como eleições.

Em comparação com bases internacionais, as brasileiras ainda enfrentam obstáculos significativos relacionados à escassez de dados em português, à baixa diversidade temática e à falta de inclusão de conteúdos multimodais, como imagens e vídeos, que têm papel crescente na propagação de notícias falsas [Farhangian et al. 2024, Villela et al. 2023]. Para avançar, é fundamental promover esforços colaborativos que ampliem o volume de dados, integrem múltiplas fontes — incluindo redes sociais e portais de *fact-checking* —, diversifiquem os temas abordados e incorporem diferentes formatos de mídia. Essa expansão colaborativa pode impulsionar o desenvolvimento de modelos mais representativos e eficazes no combate à desinformação no contexto brasileiro [Revista Pesquisa FAPESP 2024, FEBRACE 2023].

5. Conclusão

Este trabalho realizou uma análise abrangente das bases de dados em português para detecção de *fake news*, evidenciando lacunas estruturais que impactam diretamente o desenvolvimento de soluções tecnológicas robustas para o contexto brasileiro. Ao comparar essas bases com *datasets* internacionais, identificou-se que as iniciativas nacionais ainda enfrentam desafios significativos em termos de volume, diversidade temática, multimodalidade e, em alguns casos, balanceamento de classes.

A principal contribuição deste estudo reside em sistematizar e categorizar essas limitações, oferecendo um panorama detalhado dos obstáculos que dificultam a criação de modelos de detecção de *fake news* realmente eficazes para a língua portuguesa. Ao destacar que a maioria das bases nacionais é pequena, restrita a poucos temas (especialmente política e saúde) e predominantemente textuais, o trabalho evidencia como essas características limitam a aplicabilidade dos modelos em cenários reais, onde a desinformação circula em múltiplos formatos e sobre os mais variados assuntos.

Além disso, a análise mostra que, embora algumas bases nacionais sejam balanceadas, a ausência de dados multimodais e a baixa diversidade temática ainda comprometem a capacidade dos modelos de atender às necessidades da sociedade brasileira, que é fortemente impactada pela desinformação em redes sociais e aplicativos de mensagens [Agência Brasil 2024, Fiocruz 2024]. Assim, este artigo contribui ao campo ao mapear não apenas o estado da arte, mas também os pontos críticos que precisam ser superados para que a tecnologia possa servir como aliada no combate à desinformação.

Para caminhos futuros, destaca-se a necessidade de construção de bases de dados mais amplas, diversificadas e multimodais, incorporando texto, imagens, vídeos e áudios. A ampliação temática é também fundamental, permitindo que modelos de detecção sejam treinados para reconhecer *fake news* em diferentes domínios, como ciência, meio ambiente, cultura e segurança pública. Tais iniciativas podem, inclusive, inspirar políticas públicas e práticas de letramento digital, essenciais para mitigar o impacto social das notícias falsas. Por fim, recomenda-se que pesquisas futuras avaliem sistematicamente o desempenho de modelos de classificação treinados em bases em português e em inglês, de modo a identificar adaptações necessárias para o contexto local.

Referências

Agência Brasil (2024). Quase 90% dos brasileiros admitem ter acreditado em fake news.

Barbado, R., Araque, O., and Iglesias, C. A. (2019). A framework for fake review detection in online consumer electronics retailers. *Information Processing & Management*, 56(4):1234–1244.

Boididou, C., Papadopoulos, S., Zampoglou, M., Apostolidis, L., Papadopoulou, O., and Kompatsiaris, Y. (2018). Detection and visualization of misleading content on twitter. *International Journal of Multimedia Information Retrieval*, 7(1):71–86.

Chavarro, J. P., Carvalho, J. T., Portela, T. T., and Silva, J. C. (2023). Faketruebr: Um corpus brasileiro de notícias falsas. In *Escola Regional de Banco de Dados (ERBD)*, pages 108–117. SBC.

D'ulizia, A., Caschera, M. C., Ferri, F., and Grifoni, P. (2021). Fake news detection: a survey of evaluation datasets. *PeerJ Computer Science*, 7:e518.

Farhangian, F., Cruz, R. M., and Cavalcanti, G. D. (2024). Fake news detection: Taxonomy and comparative study. *Information Fusion*, 103:102140.

FEBRACE (2023). Poster - soc 1845. https://virtual.febrace.org.br/2023/SOC/1845/poster/.

Fiocruz (2024). Pesquisa revela dados sobre fake news relacionadas à covid-19.

Garcia, G. L., Paiola, P. H., Jodas, D. S., Sugi, L. A., and Papa, J. P. (2024). Text summarization and temporal learning models applied to portuguese fake news detection in a novel brazilian corpus dataset. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese*, pages 86–96.

Gruppi, M., Horne, B. D., and Adalı, S. (2021). Nela-gt-2020: A large multi-labelled news dataset for the study of misinformation in news articles. *arXiv preprint arXiv:2102.04567*.

Irís, A. and da Silva, W. M. (2024). (des) montagem de uma fake news exibida em vídeo: A multimodalidade em enunciados de leitura. *SAPIENS-Revista de divulgação Científica*, 6(1).

Macedo, L. B. B., de Sousa Oliveira, I., and de Lima, L. M. (2022). Multimodalidade e fake news: investigando os significados visuais nas postagens do facebook contendo notícias falsas. *Entrepalavras*, 11(3):526–549.

Mitra, T. and Gilbert, E. (2015). Credbank: A large-scale social media corpus with associated credibility annotations. In *Proceedings of the international AAAI conference on web and social media*, volume 9, pages 258–267.

Moreno, J. and Bressan, G. (2019). Factck. br: a new dataset to study fake news. In *Proceedings of the 25th Brazillian Symposium on Multimedia and the Web*, pages 525–527.

Revista Pesquisa FAPESP (2024). Ferramenta on-line tenta identificar fake news.

Santos, R. L., Monteiro, R. A., and Pardo, T. A. (2018). The fake. br corpus-a corpus of fake news for brazilian portuguese. In *Latin American and Iberian Languages Open Corpora Forum (OpenCor)*, pages 1–2.

Thorne, J., Vlachos, A., Christodoulopoulos, C., and Mittal, A. (2018). Fever: a large-scale dataset for fact extraction and verification. *arXiv preprint arXiv:1803.05355*.

TIGRE, M. F. F. d. S. et al. (2023). Utilizando modelos de machine learning para classificar fake news de covid-19.

Villela, H. F., Corrêa, F., Ribeiro, J. S. d. A. N., Rabelo, A., and Carvalho, D. B. F. (2023). Fake news detection: a systematic literature review of machine learning algorithms and datasets. *Journal on Interactive Systems*, 14(1):47–58.

Wang, W. Y. (2017). "liar, liar pants on fire": A new benchmark dataset for fake news detection. *arXiv preprint arXiv:1705.00648*.

Yibo, Z. (2024). Desenvolvimento da interação escrita em português língua não materna: uma experiência no nível a1. 2.