UNIVERSIDADE FEDERAL DE GOIÁS (UFG)
INSTITUTO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

JULIANA RESPLANDE SANT'ANNA GOMES

# **Verificação Semi-Automática de Fatos em Português**: Enriquecimento de *Corpus* via Busca e Extração de Alegação

GOIÂNIA
2025

UFG

UNIVERSIDADE FEDERAL DE GOIÁS
INSTITUTO DE INFORMÁTICA

**TERMO DE CIÊNCIA E DE AUTORIZAÇÃO (TECA) PARA DISPONIBILIZAR VERSÕES ELETRÔNICAS DE TESES E DISSERTAÇÕES NA BIBLIOTECA DIGITAL DA UFG**

Na qualidade de titular dos direitos de autor, autorizo a Universidade Federal de Goiás (UFG) a disponibilizar, gratuitamente, por meio da Biblioteca Digital de Teses e Dissertações (BDTD/UFG), regulamentada pela Resolução CEPEC nº 832/2007, sem ressarcimento dos direitos autorais, de acordo com a Lei 9.610/98, o documento conforme permissões assinaladas abaixo, para fins de leitura, impressão e/ou download, a título de divulgação da produção científica brasileira, a partir desta data.

O conteúdo das Teses e Dissertações disponibilizado na BDTD/UFG é de responsabilidade exclusiva do autor. Ao encaminhar o produto final, o autor(a) e o(a) orientador(a) firmam o compromisso de que o trabalho não contém nenhuma violação de quaisquer direitos autorais ou outro direito de terceiros.

**1. Identificação do material bibliográfico**

[X] Dissertação [ ] Tese [ ] Outro*:_____

*No caso de mestrado/doutorado profissional, indique o formato do Trabalho de Conclusão de Curso, permitido no documento de área, correspondente ao programa de pós-graduação, orientado pela legislação vigente da CAPES.

**Exemplos**: Estudo de caso ou Revisão sistemática ou outros formatos.

**2. Nome completo do autor**

Juliana Resplande Sant'anna Gomes

**3. Título do trabalho**

Verificação Semi-Automática de Fatos em Português: Enriquecimento de Corpus via Busca e Extração de Alegação

**4. Informações de acesso ao documento (este campo deve ser preenchido pelo orientador)**

Concorda com a liberação total do documento [ X ] SIM [ ] NÃO[1]

**[1]** Neste caso o documento será embargado por até um ano a partir da data de defesa. Após esse período, a possível disponibilização ocorrerá apenas mediante:
**a)** consulta ao(à) autor(a) e ao(à) orientador(a);
**b)** novo Termo de Ciência e de Autorização (TECA) assinado e inserido no arquivo da tese ou dissertação.
O documento não será disponibilizado durante o período de embargo.
Casos de embargo:
- Solicitação de registro de patente;
- Submissão de artigo em revista científica;
- Publicação como capítulo de livro;
- Publicação da dissertação/tese em livro.

**Obs. Este termo deverá ser assinado no SEI pelo orientador e pelo autor.**

Documento assinado eletronicamente por **Arlindo Rodrigues Galvao Filho**, **Professor do Magistério Superior**, em 13/07/2025, às 09:11, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Juliana Resplande Sant'Anna Gomes**, **Discente**, em 21/07/2025, às 21:13, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no site https://sei.ufg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **5499635** e o código CRC **FD57804B**.

JULIANA RESPLANDE SANT'ANNA GOMES

# **Verificação Semi-Automática de Fatos em Português**: Enriquecimento de *Corpus* via Busca e Extração de Alegação

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação do Instituto de Informática da Universidade Federal de Goiás (UFG) como requisito para obtenção do título de Mestra em Ciência da Computação.
Área de concentração: Ciência da Computação
Linha de pesquisa: Sistemas Inteligentes e Aplicações

Orientador(a): Prof. Dr. Arlindo Rodrigues Galvão Filho

GOIÂNIA
2025

Ficha de identificação da obra elaborada pelo autor, através do
Programa de Geração Automática do Sistema de Bibliotecas da UFG.

Gomes, Juliana Resplande Sant'Anna
Verificação Semi-Automática de Fatos em Português [manuscrito] : Enriquecimento de Corpus via Busca e Extração de Alegação / Juliana Resplande Sant'Anna Gomes. - 2025.
119 f.

Orientador: Prof. Dr. Arlindo Rodrigues Galvão Filho.
Dissertação (Mestrado) - Universidade Federal de Goiás, Instituto de Informática (INF), Programa de Pós-Graduação em Ciência da Computação, Goiânia, 2025.
Bibliografia. Apêndice.
Inclui tabelas, lista de figuras, lista de tabelas.

1. Processamento de Linguagem Natural. 2. Fake News. 3. Verificação Semi-automática de Fatos. 4. Corpora em português. I. Filho, Arlindo Rodrigues Galvão, orient. II. Título.

CDU 004

Processo: 23070.025069/2025-84 Documento: 5522857

UNIVERSIDADE FEDERAL DE GOIÁS

INSTITUTO DE INFORMÁTICA

**ATA DE DEFESA DE DISSERTAÇÃO**

Ata nº **15** da sessão de Defesa de Dissertação de **Juliana Resplande Sant'anna Gomes**, que confere o título de Mestra em Ciência da Computação, na área de concentração em Ciência da Computação.

Aos dez dias do mês de junho de dois mil e vinte e cinco, a partir das quinze horas, via sistema de webconferência, realizou-se a sessão pública de Defesa de Dissertação intitulada **"Verificação Semi-Automática de Fatos em Português: Enriquecimento de Corpus via Busca e Extração de Alegação"**. Os trabalhos foram instalados pelo Orientador, Professor Doutor Arlindo Rodrigues Galvão Filho (INF/UFG) com a participação dos demais membros da Banca Examinadora: Professor Doutor Eliomar Araújo de Lima (INF/UFG), membro titular externo; Professora Doutora Telma Woerle de Lima Soares (INF/UFG) , membra titular interna. A realização da banca ocorreu por meio de videoconferência. Durante a arguição os membros da banca não fizeram sugestão de alteração do título do trabalho. A Banca Examinadora reuniu-se em sessão secreta a fim de concluir o julgamento da Dissertação, tendo sido a candidata **aprovada** pelos seus membros. Proclamados os resultados pelo Professor Doutor Arlindo Rodrigues Galvão Filho, Presidente da Banca Examinadora, foram encerrados os trabalhos e, para constar, lavrou-se a presente ata que é assinada pelos Membros da Banca Examinadora, aos dez dias do mês de junho de dois mil e vinte e cinco.

TÍTULO SUGERIDO PELA BANCA

Documento assinado eletronicamente por **Eliomar Araujo De Lima**, **Professor do Magistério Superior**, em 23/07/2025, às 15:03, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Juliana Resplande Sant'Anna Gomes**, **Discente**, em 23/07/2025, às 16:30, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Telma Woerle De Lima Soares**, **Professora do Magistério Superior**, em 24/07/2025, às 08:41, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Arlindo Rodrigues Galvao Filho**, **Professor do Magistério Superior**, em 25/07/2025, às 07:32, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no site https://sei.ufg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **5522857** e o código CRC **2AB40969**.

**Referência:** Processo nº 23070.025069/2025-84 SEI nº 5522857

JULIANA RESPLANDE SANT'ANNA GOMES

# **Verificação Semi-Automática de Fatos em Português**: Enriquecimento de *Corpus* via Busca e Extração de Alegação

Dissertação apresentada ao Programa de Pós–Graduação da Instituto de Informática da Universidade Federal de Goiás (UFG) como requisito parcial para obtenção do título de Mestre em Programa de Pós-Graduação em Ciência da Computação, aprovada em 10 de Abril de 2025, pela Banca Examinadora constituída pelos professores:

---

**Prof. Arlindo Rodrigues Galvão Filho**
Instituto de Informática – UFG
Presidente da Banca

---

**Profa. Telma Woerle de Lima Soares**
Instituto de Informática – UFG

---

**Prof. Eliomar Araújo de Lima**
Instituto de Informática – UFG

**Juliana Resplande Sant'Anna Gomes**

Mestra pela Universidade Federal de Goiás (UFG), filiada ao Centro de Excelência em Inteligência Artificial (CEIA). Meus interesses de pesquisa são obtenção e tradução de *corpora* ao português, verificação de *fake news*, sistemas de *Question Answering* e geração de texto.

Dedico este trabalho à minha mãe, que com imensa força e cumplicidade, esteve ao meu lado em cada desafio, formando comigo o pilar central da nossa jornada.

# Agradecimentos

A conclusão desta dissertação de mestrado marca o fim de uma jornada desafiadora e gratificante, e seria impossível alcançá-la sem o apoio e a contribuição de diversas pessoas e instituições a quem desejo expressar minha mais sincera gratidão.

Em primeiro lugar, agradeço profundamente ao meu orientador, Prof. Dr. Arlindo Rodrigues Galvão Filho, pela orientação segura, pelos valiosos ensinamentos, pela paciência e pela confiança depositada em meu trabalho ao longo desta caminhada. Sua expertise foi fundamental para o desenvolvimento desta pesquisa. Aos membros da banca examinadora, Prof. Dr. Eliomar Araújo de Lima e Prof.ª Dr.ª Telma Woerle de Lima Soares, minha gratidão pelas leituras atentas, críticas construtivas e sugestões perspicazes que enriqueceram significativamente este estudo e abriram novos horizontes para reflexão.

Meu reconhecimento se estende ao Centro de Competência EMBRAPII em Tecnologias Imersivas (Advanced Knowledge Center for Immersive Technologies - AKCIT) e ao Centro de Excelência em Inteligência Artificial (CEIA). O apoio institucional, as discussões estimulantes e a colaboração frutífera foram cruciais para o avanço e a concretização desta pesquisa. Sou igualmente grata ao Instituto de Informática da Universidade Federal de Goiás (INF-UFG) e à Universidade Federal de Goiás (UFG) como um todo, por fornecerem a infraestrutura, os recursos e o ambiente acadêmico propício para a realização deste mestrado.

No âmbito pessoal, palavras são insuficientes para expressar minha eterna gratidão à minha mãe, por seu amor incondicional, por ser a base da minha formação e por sempre me incentivar tanto na esfera profissional quanto pessoal. Seus sacrifícios e apoio foram meu alicerce.

Ao meu companheiro de mestrado e de vida, Eduardo Augusto Santos Garcia, meu mais profundo agradecimento. Sua compreensão, apoio constante e as valiosas revisões e discussões ao longo de todo o processo, foram essenciais. Sua presença tornou cada desafio mais leve.

Aos meus amigos, pela amizade leal, pelo incentivo vibrante, pelas palavras de conforto nos momentos de dificuldade e pela celebração de cada conquista. Vocês foram um refúgio e uma fonte de energia.

"Nós não estamos somente a pandemia, nós estamos combatendo uma infodemia."

**Tedros Adhanom Ghebreyesus - diretor geral da Organização Mundial da Saúde**,
*Conferência de Segurança de Munique (MSC) em 2020.*

# Resumo


Gomes, Juliana Resplande Sant'Anna. **Verificação Semi-Automática de Fatos em Português**: Enriquecimento de *Corpus* via Busca e Extração de Alegação. Goiânia, 2025. 119p. Dissertação de Mestrado. Instituto de Informática, Universidade Federal de Goiás (UFG).

A disseminação acelerada de desinformação excede a capacidade da verificação manual de fatos, evidenciando a necessidade de sistemas de Verificação Semi-Automática de Fatos (AFC). No contexto da língua portuguesa, constata-se uma carência de conjuntos de dados (*corpora*) publicamente disponíveis que integrem evidências externas, um componente essencial para o desenvolvimento de sistemas robustos de AFC, uma vez que muitos recursos existentes focam apenas na classificação baseada em características intrínsecas do texto.

Esta dissertação aborda essa lacuna desenvolvendo, aplicando e analisando uma metodologia para enriquecer *corpora* de notícias em português (Fake.Br, COVID19.BR, MuMiN-PT) com evidências externas. A abordagem simula o processo de verificação de um usuário, empregando Modelos de Linguagem Grandes (LLMs, especificamente Gemini 1.5 Flash) para extrair a alegação principal dos textos e APIs de mecanismos de busca (API de busca do Google, API de busca de alegações do Google FactCheck) para recuperar documentos externos relevantes (evidências). Adicionalmente, um processo de validação e pré-processamento de dados, incluindo detecção de quase duplicatas, é introduzido para aprimorar a qualidade dos *corpora* base.

Os principais resultados demonstram a viabilidade da metodologia, fornecendo *corpora* enriquecidos e análises que confirmam a utilidade da extração de alegações, a influência das características dos dados originais no processo, e o impacto positivo do enriquecimento no desempenho de modelos de classificação (Bertimbau e Gemini 1.5 Flash), especialmente com ajuste fino. Este trabalho contribui com recursos valiosos e *insights* para o avanço da AFC em português.



**Palavras–chave**

Processamento de Linguagem Natural, *Fake News*, Verificação Semi-Automática de Fatos, *Corpora* em português.

# Abstract

Gomes, Juliana Resplande Sant'Anna. Semi-automated Fact-checking in Portuguese: *Corpora* Enrichment using Retrieval with Claim extraction. Goiânia, 2025. 119p. MSc. Dissertation. Instituto de Informática, Universidade Federal de Goiás (UFG).

The accelerated dissemination of disinformation often outpaces the capacity for manual fact-checking, highlighting the urgent need for Semi-Automated Fact-Checking (SAFC) systems. Within the Portuguese language context, there is a noted scarcity of publicly available datasets (*corpora*) that integrate external evidence, an essential component for developing robust AFC systems, as many existing resources focus solely on classification based on intrinsic text features.

This dissertation addresses this gap by developing, applying, and analyzing a methodology to enrich Portuguese news *corpora* (Fake.Br, COVID19.BR, MuMiN-PT) with external evidence. The approach simulates a user's verification process, employing Large Language Models (LLMs, specifically Gemini 1.5 Flash) to extract the main claim from texts and search engine APIs (Google Search API, Google FactCheck Claims Search API) to retrieve relevant external documents (evidence). Additionally, a data validation and preprocessing framework, including near-duplicate detection, is introduced to enhance the quality of the base *corpora*.

The main results demonstrate the methodology's viability, providing enriched *corpora* and analyses that confirm the utility of claim extraction, the influence of original data characteristics on the process, and the positive impact of enrichment on the performance of classification models (Bertimbau and Gemini 1.5 Flash), especially with fine-tuning. This work contributes valuable resources and insights for advancing SAFC in Portuguese.

**Keywords**

Natural Language Processing, Fake news, Semi-Automated Fact-checking, Portuguese *Corpora*.

# Sumário

# Lista de Figuras

# Lista de Tabelas

CAPÍTULO 1

# Introdução

A era digital, especialmente a partir de 2015, redefiniu o consumo de informação, caracterizada pelo uso massivo das mídias sociais. Nesse cenário, observa-se que conteúdos circulam frequentemente sem os critérios de rigor e qualidade associados ao jornalismo tradicional. Em países como Reino Unido e Estados Unidos, constata-se uma mudança geracional, na qual os jovens adotam crescentemente as redes sociais digitais como principal fonte de notícias, em detrimento de meios consolidados como a televisão [4]. Contudo, essa mesma dinâmica digital facilitou a disseminação de desinformação e *fake news*, que se tornaram ferramentas potentes de manipulação, capazes de infligir danos significativos a reputações *corpora*tivas, governamentais e a grupos sociais [79, 109].

Diante desse desafio, agências de checagem de fatos (*fact-checking*), como a Agência Lupa e o Boatos.org no Brasil, desempenham um papel crucial ao investigar manualmente a veracidade das informações [44, 30]. No entanto, a velocidade viral com que a desinformação se propaga excede em muito a capacidade de verificação humana, um problema exacerbado durante a pandemia de COVID-19, período em que o confinamento intensificou o uso da internet e a circulação de conteúdos duvidosos [125].

Essa limitação intrínseca da verificação manual impulsionou a pesquisa e o desenvolvimento de ferramentas de verificação automática ou semi-automática de fatos, área conhecida como *Automated Fact-Checking* (AFC). Essas abordagens buscam analisar a veracidade de alegações comparando-as com fontes externas de conhecimento, integrando técnicas de Recuperação de Informação (IR) e Processamento de Linguagem Natural (PLN) [51].

Reconhece-se, no entanto, que a automação completa ainda enfrenta desafios significativos, especialmente na interpretação de nuances contextuais e na avaliação da credibilidade das fontes. Por essa razão, muitos sistemas operam de forma semi-automática, onde a tecnologia auxilia o especialista humano, que realiza a validação final [78, 127]. Adotou-se, nesta dissertação, o termo Verificação Semi-Automática de Fatos para refletir essa interação colaborativa.

No contexto da língua portuguesa, apesar da existência de pesquisas e recursos para a detecção de *fake news*, uma lacuna significativa persiste. Uma análise de 18 *corpora*

publicamente disponíveis (detalhados na Seção 3.3) revela que a maioria se concentra na classificação da notícia com base em suas características intrínsecas (estilo de escrita, parcialidade), como exemplificado pelo *corpus* Fake.br [83] e abordagens iniciais em inglês [51]. São escassos os conjuntos de dados em português que fornecem as evidências externas associadas às alegações, um componente crucial para treinar e avaliar sistemas semi-automáticos de AFC robustos que se baseiam na verificação factual contra fontes externas, em vez de apenas classificar o texto isoladamente.

Esta dissertação visa abordar diretamente essa lacuna. O objetivo central deste trabalho é desenvolver uma método para enriquecer conjuntos de dados de notícias em português já existentes, agregando a eles evidências contextuais relevantes recuperadas de fontes externas. Para este fim, foram selecionados três *corpora* proeminentes com distintas características de fonte, método de coleta e temporalidade: Fake.Br [83], notícias de páginas da web de domínio geral; COVID19.BR [76], mensagens do WhatsApp sobre saúde e MuMiN-PT [90], *tweets* de domínio geral.

A abordagem proposta simula o processo cognitivo de um usuário que busca informações adicionais para verificar a veracidade de uma notícia. Para tanto, foram empregados Modelos de Linguagem Grandes (LLMs), especificamente o Gemini 1.5 Flash, para extrair a alegação principal contida no texto original, especialmente quando a busca direta não encontrava correspondência forte. Essa alegação serve como consulta otimizada para mecanismos de busca, como a API de busca do Google (CSE) e a API de busca de alegações do Google FactCheck, na recuperação de documentos externos relevantes (evidências), que são então associados ao item original do *corpus*.

O resultado principal desta pesquisa é a criação de versões enriquecidas destes *corpora*, acompanhadas de uma análise detalhada do processo de coleta, das características dos dados obtidos (como a prevalência de quase duplicatas e a natureza das evidências recuperadas) e do impacto das características originais dos conjuntos de dados nas análises subsequentes. Esta dissertação, portanto, não se limita a apresentar os conjuntos de dados enriquecidos como produto final, mas detalha e analisa criticamente todo o fluxo de trabalho operacional — desde a validação dos dados brutos, passando pela extração de alegações, até a recuperação e avaliação das evidências —, oferecendo insights sobre os desafios e as decisões metodológicas em cada etapa.

Uma análise qualitativa dos dados enriquecidos identificou padrões recorrentes na corroboração ou refutação de alegações, incluindo o reconhecimento de exemplos dos *corpora* em publicações acadêmicas. Além isso, o processo de validação semi-automático dos dados, que incluiu a detecção de quase duplicatas e checagens de consistência de rótulos, mostrou-se fundamental para melhorar a qualidade dos dados base antes do enriquecimento.

De forma a avaliar experimentalmente o impacto do processo, comparou-se

o desempenho do ajuste fino do Bertimbau base e de *few-shot prompting* do Gemini 1.5 Flash em diferentes configurações de dados (originais, validados, e validados e enriquecidos). Os resultados indicaram que, embora os dados apenas validados pudessem apresentar desempenho inferior aos originais devido ao aumento da complexidade da tarefa, o enriquecimento com conteúdo externo geralmente melhorou o desempenho sobre os dados apenas validados, especialmente para o Bertimbau e Gemini no COVID19.BR. Consistentemente, o ajuste fino do Bertimbau superou o *few-shot learning* com Gemini. Estes achados sugerem que o enriquecimento adiciona contexto valioso, mas sua eficácia depende da qualidade da busca, da extração de alegações, da cobertura das APIs e da temporalidade da informação, apontando para o potencial de abordagens híbridas.

Este trabalho foi motivado pela participação da autora em um projeto de pesquisa sobre *Fake News* em colaboração com a Agência Nacional de Telecomunicações (ANATEL) e Fundação de Amparo à Pesquisa do Estado de Goiás (FAPEG), em que se publicou uma revisão terciária rápida sobre a área de detecção de fake news [49].

A dissertação também conta com o apoio do Centro de Excelência em Inteligência Artificial (CEIA) e do Centro de Competência EMBRAPII em Tecnologias Imersivas (Advanced Knowledge Center for Immersive Technologies - AKCIT) do Instituto de Informática da Universidade Federal de Goiás (INF-UFG), no qual a autora participou da competição CLEF CheckThat! 2025 [5, 6]. O sistema proposto, utilizando o ajuste fino do modelo Mono-PTT5 [98], provou-se altamente competitivo ao alcançar a terceira posição para o português (METEOR 0.5290) [118].

A seguir, na Seção 1.1, são apresentadas as hipóteses que nortearam esta investigação. Os objetivos específicos decorrentes dessas hipóteses são detalhados na Seção 1.2, e as principais contribuições do trabalho estão listadas na Seção 1.3. A estrutura restante desta dissertação é organizada da seguinte forma:

**Capítulo 2** Apresenta a fundamentação teórica sobre *fake news*, Processamento de Linguagem Natural (com foco em Modelos de Linguagem Grandes - LLMs) e Recuperação de Informação (IR).

**Capítulo 3** Discute os trabalhos relacionados, abordando o uso de LLMs na verificação de fatos e pesquisas em PLN para detecção de *fake news* em português.

**Capítulo 4** Descreve detalhadamente os métodos propostos para a validação semi-automático, enriquecimento e avaliação experimental dos *corpora*.

**Capítulo 5** Apresenta o desenvolvimento prático do enriquecimento dos *corpora*, a análise exploratória dos dados originais, a análise quantitativa e qualitativa dos dados enriquecidos, e os resultados da avaliação experimental.

**Capítulo 6** Sumariza as conclusões do trabalho e aponta direções para pesquisas futuras.

## 1.1 Hipóteses

Com o intuito de analisar o cenário em língua portuguesa de verificação de *fake news* com conteúdo externo, foram elaboradas as seguintes hipóteses:

**H1: Há uma escassez de *Corpora* em português acessível para detecção de *fake news* usando conteúdo externo, técnica conhecida como verificação de fatos**. Dentre 18 conjuntos de dados identificados, poucos fornecem evidências associadas, e os que o fazem podem apresentar limitações de acesso ou escopo.

**H2: A partir dos conjuntos de dados em português de detecção de *fake news* que possuem somente alegações, é possível enriquecê-los com evidências**. Utilizando mecanismos de busca e, quando necessário, extração de alegação via LLM, pode-se obter material contextual para cada notícia, embora se reconheçam limitações temporais (relevância da evidência ao longo do tempo, mudança de veracidade de fatos) e de verificabilidade (alegações subjetivas, "meias-verdades") inerentes ao processo.

**H3: A extração de alegação via LLM auxilia no processo de verificação de *fake news* em português, otimizando a busca por evidências**. Acredita-se que esta abordagem melhora a precisão da busca por evidências, especialmente para textos menos diretos, embora sua necessidade varie conforme a clareza do texto original e a natureza do *corpus*.

**H4: A natureza da coleta dos dados (*top-down* vs. *bottom-up*) interfere no processo de enriquecimento e nas características dos *corpora***. Conjuntos de dados *bottom-up* (ex. MuMiN-PT), originados de verificações de agências, tendem a ter mais exemplos falsos e maior correspondência com resultados de busca de verificações, demandando menos extração de alegação.

**H5: O veículo de publicação impacta nas características distintivas entre *fake news* e textos verdadeiros.** Conforme observado na análise exploratória, o meio original de publicação da notícia (e.g., Twitter, sites de notícias, WhatsApp) influencia características textuais como tamanho, homogeneidade, presença de URLs e a natureza das quase duplicatas, que podem diferenciar notícias verdadeiras de falsas de maneiras específicas para cada plataforma.

**H6: O processo de validação e enriquecimento dos dados auxiliam o desempenho de modelos de detecção de *fake news*.** A validação, ao remover certos sinais (e.g., URLs), pode tornar a tarefa mais desafiadora. O enriquecimento com conteúdo externo, por outro lado, visa fornecer contexto adicional que pode melhorar a capacidade de generalização e o desempenho dos modelos, embora a qualidade e relevância da informação externa sejam cruciais.

## 1.2 Objetivos

Com base no problema identificado e nas hipóteses formuladas, o objetivo geral desta dissertação é **desenvolver uma metodologia para enriquecer conjuntos de dados de notícias em português com evidências externas recuperadas automaticamente**. Os objetivos específicos são:

1. Realizar um levantamento abrangente e uma análise comparativa aprofundada dos *corpora* existentes para detecção de *fake news* em português, focando em suas características relevantes para a verificação baseada em evidências (ex., fonte, método de coleta, disponibilidade de evidências, características textuais, prevalência de quase duplicatas).
2. Projetar e implementar um fluxo de trabalho para o enriquecimento de *corpora*, incluindo uma etapa de validação semi-automática dos dados base, investigando e selecionando técnicas de engenharia de *prompt* para extração de alegações via LLMs e utilizando APIs de mecanismos de busca para a recuperação de evidências.
3. Aplicar a metodologia de validação e enriquecimento em *corpora* selecionados em português (Fake.Br, COVID19.BR, MuMiN-PT), gerando conjuntos de dados validados e enriquecidos com referências externas.
4. Avaliar os resultados do processo de enriquecimento, considerando aspectos como a aplicabilidade da extração de alegação em diferentes *corpora*, a distribuição e natureza das fontes de evidência recuperadas (busca web geral vs. API de Fact Check, incluindo domínios governamentais, de mídia e de checagem), e as características dos dados gerados, relacionando-as às propriedades originais dos conjuntos de dados e identificando padrões qualitativos nas evidências.
5. Comparar experimentalmente a eficácia das etapas propostas de validação semi-automática dos dados e de enriquecimento com conteúdo externo no desempenho de modelos de PLN (Bertimbau e Gemini 1.5 Flash) na tarefa de classificação de veracidade.

## 1.3 Contribuição

As contribuições desta dissertação são:

1. Um levantamento e comparação *corpora* de *fake news* em português (18 conjuntos de dados), destacando características frequentemente negligenciadas como métodos de coleta (*top-down* vs. *bottom-up*) e a prevalência e impacto de exemplos quase duplicados, e a influência das fontes de informação primárias (notícias da internet, Twitter, WhatsApp).

2. Um processo de validação e pré-processamento de dados, incorporando a detecção de quase duplicatas (MinHash LSH) e checagens de consistência de rótulos, visando melhorar a confiabilidade dos *corpora* base antes do enriquecimento, com a remoção de vieses como URLs explícitas.
3. O desenvolvimento de uma metodologia para enriquecer conjuntos de dados de notícias em português com informações contextuais externas, utilizando LLMs (Gemini 1.5 Flash) para extração de alegações e APIs de mecanismos de busca (API de Busca do Google e API de Busca de alegações do Google FactCheck) para recuperação de evidências.
4. Uma análise aprofundada dos *corpora* enriquecidos sobre a eficácia da extração de alegações em diferentes contextos, a natureza e distribuição dos domínios das fontes de evidência recuperadas (incluindo fontes jornalísticas, governamentais, acadêmicas e de checagem), e como as características originais dos *corpora* (fonte, temporalidade, método de coleta) influenciam o processo de enriquecimento e a natureza das evidências. Inclui-se uma análise qualitativa dos padrões de evidências e a identificação de publicações acadêmicas que referenciam exemplos dos *corpora*.
5. Uma avaliação experimental do impacto das etapas de validação e enriquecimento no desempenho de modelos de detecção de *fake news* (Bertimbau e Gemini 1.5 Flash), demonstrando o potencial do enriquecimento contextual, mas também suas nuances e dependências da qualidade e relevância da informação externa.

CAPÍTULO 2

# Fundamentos

Este capítulo apresenta os conceitos essenciais para a compreensão do problema da *fake news* e das abordagens computacionais para seu combate. Inicialmente, explora-se o fenômeno das *fake news* na Seção 2.1, detalhando sua taxonomia e as técnicas empregadas para sua detecção, com ênfase na verificação de fatos (*fact-checking*). Em seguida, abordam-se os fundamentos das áreas de conhecimento que viabilizam a automação ou semi-automação desse processo: o Processamento de Linguagem Natural (PLN) na Seção 2.2, e a Recuperação de Informação (RI) em 2.4.

No âmbito do PLN, discute-se a evolução dos modelos de linguagem, desde abordagens estatísticas até os Modelos de Linguagem Pequenos (SLMs), como o BERT e o Bertimbau, e os Modelos de Linguagem Grandes (LLMs), como o Gemini. Destaca-se, neste contexto, a técnica de engenharia de *prompts*, fundamental para interagir com LLMs na tarefa de extração de alegações, como realizado neste trabalho. Finalmente, explora-se a Recuperação de Informação, definindo seus conceitos gerais e sua interconexão com o PLN, a verificação de fatos (especificamente na busca por evidências) e os mecanismos de busca, como o buscador Google, cujas APIs são frequentemente empregadas em sistemas de verificação.

## 2.1 *Fake news*

*Fake news* podem ser compreendidas como conteúdo fabricado que mimetiza o formato de notícias genuínas, com o intuito de enganar o leitor [132]. É relevante notar que o termo "*fake news*" carece de uma definição única e universalmente aceita, sendo sua interpretação sujeita a variações contextuais e disciplinares [4].

A Seção 2.1.1 aprofunda a terminologia associada. Posteriormente, a Seção 2.1.2 delineia as principais abordagens tecnológicas para a detecção e análise de *fake news*, com base em [52].

### 2.1.1 Taxonomia

Sharma et al. (2019) definem *fake news* como notícias ou mensagens publicadas e disseminadas pela mídia que contêm informações falsas, independentemente dos meios ou motivações por trás de sua propagação. Esta definição abrange diversos tipos de conteúdo problemático identificados na literatura [112].

Com base nessa perspectiva, é possível categorizar as *fake news* em subtipos como: conteúdo fabricado (totalmente falso), conteúdo enganoso (uso de informação para distorcer um fato ou indivíduo), conteúdo impostor (fontes genuínas falsamente atribuídas), conteúdo manipulado (informação ou imagem genuína alterada para enganar), conexão falsa (títulos, imagens ou legendas não condizentes com o conteúdo) e contexto falso (conteúdo genuíno compartilhado com informação contextual incorreta) [112].

Adicionalmente, a informação falsa pode ser classificada pela intenção. A *misinformation* (informação falsa, não intencional) refere-se à disseminação involuntária de falsidades, que pode advir de vieses cognitivos, falta de compreensão ou atenção. Em contraste, a *disinformation* (desinformação, intencional) envolve a criação e propagação deliberada de informações falsas com o objetivo explícito de enganar [4, 112].
Outras formas de conteúdo relacionadas incluem:

- **Paródia e Sátira:** Frequentemente associadas ao humor, onde sátiras podem imitar o estilo da mídia tradicional para criticar ou expor algo, sem a intenção primária de enganar, embora possam ser mal interpretadas [52].
- ***Clickbaits*:** Títulos ou chamadas sensacionalistas projetados para atrair cliques e direcionar tráfego para um *site*, muitas vezes com fins de monetização por publicidade, mesmo que o conteúdo não entregue o prometido [52].
- **Propaganda:** Informação, muitas vezes enviesada ou enganosa, usada para promover uma causa ou ponto de vista político específico, visando influenciar a opinião pública [4].
- **Teorias da Conspiração:** Explicações para eventos que invocam tramas secretas por parte de grupos poderosos, geralmente carecendo de evidências robustas e resistindo à refutação [4].

### 2.1.2 Técnicas de Detecção

Dada a escala e complexidade da disseminação de *fake news*, diversas abordagens tecnológicas têm sido desenvolvidas para sua detecção e análise. As principais se baseiam em aprendizado de máquina (ML), aprendizado profundo (DL), processamento de linguagem natural (PLN), verificação de fatos (*fact-checking* ou FC, incluindo a Verificação Automática de Fatos - AFC e a Verificação Semi-Automática de Fatos - SAFC), *crowdsourcing* (CDS), *blockchain* (BKC) e redes neurais de grafos (GNN) [4].

**Inteligência Artificial (IA):** É um termo guarda-chuva frequentemente empregado no contexto de ML ou DL [48, 52].

- **Aprendizado de Máquina (ML):** Refere-se a métodos clássicos onde características (*features*) relevantes, como contagem de palavras, comprimento de sentenças ou frequência de certos termos, são frequentemente extraídas ou projetadas manualmente (*feature engineering*) para treinar modelos preditivos.
- **Aprendizado Profundo (DL):** Utiliza redes neurais com múltiplas camadas para aprender representações hierárquicas dos dados. Geralmente, a extração de características é feita automaticamente pela própria rede durante o treinamento. Métodos de DL costumam exigir maior volume de dados e capacidade computacional, mas podem alcançar desempenho superior em tarefas complexas.

**Processamento de Linguagem Natural (PLN):** Engloba técnicas focadas na análise do conteúdo textual das notícias. Podem-se examinar aspectos:

- **Lexicais e Sintáticos:** Análise de frequência de palavras, uso de classes gramaticais (POS *tagging*), comprimento de sentenças, complexidade sintática, presença de erros ortográficos [105].
- **Semânticos:** Análise do significado do texto, incluindo a detecção de tópicos, análise de sentimentos, e a representação do texto através de *embeddings* (vetores numéricos) que capturam relações semânticas.
- **Psicolinguísticos:** Extração de características relacionadas a processos psicológicos e sociais, utilizando ferramentas como o LIWC (Linguistic Inquiry and Word Count) [122].

Classificadores tradicionais de ML, como Regressão Logística, podem ser combinados com representações textuais como TF-IDF (Frequência do Termo–Inverso da Frequência do Documento) ou com *embeddings* gerados por modelos de linguagem neural para a classificação de notícias.

**Verificação de Fatos (*Fact-Checking*, FC):** Consiste em avaliar a veracidade de alegações factuais comparando-as com fontes de informação externas e confiáveis. Este processo pode ser:

- **Manual:** Realizado por jornalistas especializados ou checadores de fatos, como os das agências listadas na Tabela 2.1. Também pode envolver ***Crowdsourcing*** **(CDS)**, onde tarefas de verificação são distribuídas a um grande número de pessoas [7, 62, 52, 4].
- **Automático ou Semi-Automático (AFC / SAFC):** Busca automatizar partes ou todo o processo de verificação usando técnicas computacionais, principalmente de PLN e RI. A abordagem semi-automática (SAFC), **adotada como foco conceitual**

**nesta dissertação**, reconhece a importância da intervenção humana em etapas críticas, como na validação final ou no tratamento de casos complexos [78, 89, 127].

**Blockchain (BKC):** É uma tecnologia de registro distribuído e imutável. Dados são armazenados em blocos encadeados criptograficamente em múltiplos servidores, sem uma autoridade central. Uma vez adicionado, um registro não pode ser alterado ou excluído. Propriedades como imutabilidade, descentralização, resistência à adulteração e transparência (controlada) tornam o blockchain uma ferramenta potencial para rastrear a proveniência e a integridade de conteúdos digitais, ajudando a verificar sua autenticidade [40, 3].

**Redes Neurais de Grafos (GNN):** São modelos de DL projetados para operar sobre dados estruturados em grafos. No contexto de *fake news*, podem modelar as relações entre usuários, notícias e fontes, ou a propagação de informação em redes sociais, para identificar padrões associados à *fake news* [4].

As técnicas de detecção podem ser agrupadas conforme seu foco principal [7, 62, 52], como ilustrado na Figura 2.1. A taxonomia divide as abordagens em quatro categorias principais: Baseadas em Características, Baseadas em Conhecimento e Baseadas em Aprendizado, complementadas por aspectos como idioma, granularidade da detecção, granularidade da verdade e plataforma.

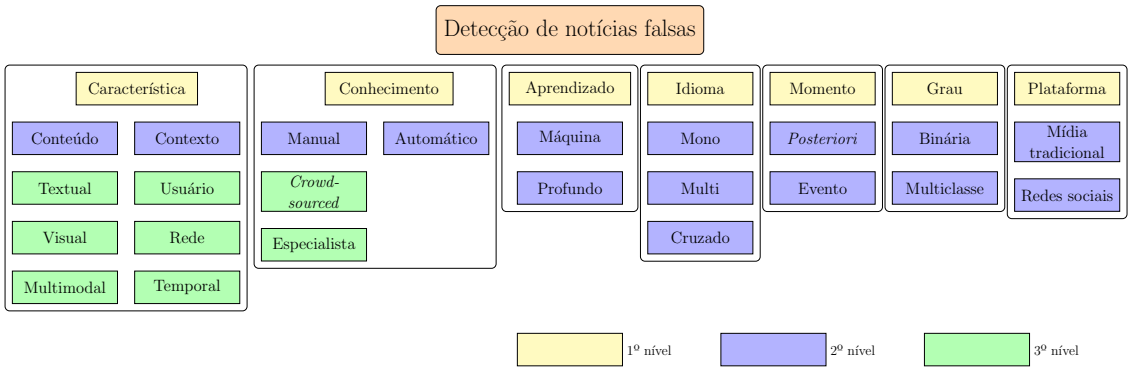


**Figura 2.1:** *Taxonomia de técnicas de detecção de fake news, adaptado de [52].*

**Abordagens Baseadas em Características:** Examinam propriedades intrínsecas ou extrínsecas da notícia para identificar padrões suspeitos.

- **Baseadas em Conteúdo:** Focam nos elementos da própria notícia: texto, imagem, vídeo (ou combinações multimodais). Análise textual (PLN), análise de sentimento (viés negativo), detecção de manipulação em imagens ou vídeos (*deepfakes*) são exemplos. Conteúdo textual pode ser extraído de áudio/vídeo via Reconhecimento Automático de Fala (ASR) ou de imagens via Reconhecimento Óptico de Caracteres (OCR).

- **Baseadas em Contexto Social:** Analisam o ambiente e a forma como a notícia se propaga. Inclui o estudo de padrões de disseminação em redes sociais, reações de usuários (comentários, compartilhamentos), características dos perfis que compartilham (detecção de bots), e a estrutura da rede de interações (usando GNNs ou análise de redes). Fatores temporais, como a recorrência de certos rumores, e a credibilidade atribuída à fonte da notícia também são considerados [4, 7, 62].

**Abordagens Baseadas em Conhecimento:** Correspondem essencialmente às técnicas de Verificação de Fatos (FC), que confrontam as alegações da notícia com fontes externas de conhecimento (bases de dados, artigos de checagem, fontes confiáveis). Um desafio central nesta abordagem é a coleta e organização das evidências relevantes, especialmente para idiomas com menos recursos, como o português, onde há carência de bases de dados que integrem alegações e suas respectivas evidências externas. Conforme mencionado, pode ser manual, automática (AFC) ou semi-automática (SAFC) [7, 62, 52].

**Abordagens Baseadas em Aprendizado:** Referem-se ao uso de IA (ML e DL). Métodos de ML clássico são geralmente mais interpretáveis e requerem menos dados, mas dependem de engenharia de características manual. Métodos de DL podem aprender representações complexas automaticamente a partir de grandes volumes de dados, frequentemente alcançando maior acurácia, mas são computacionalmente mais custosos e podem ser menos transparentes ("caixas pretas") [48].

Outros aspectos relevantes na detecção de *fake news* incluem:

- **Idioma:** A detecção pode ser monolíngue ou multilíngue. Técnicas de aprendizado translingual (*cross-lingual learning*) buscam transferir conhecimento de um idioma rico em recursos (como o inglês) para outro com menos recursos (como o português) [52].
- **Momento da Detecção:** A classificação pode ocorrer *a posteriori* (após a notícia circular) ou, idealmente, em estágios iniciais de sua disseminação (*early detection*). A maioria dos métodos atuais opera *a posteriori* [52].
- **Granularidade da Verdade:** A classificação pode ser binária (verdadeiro/falso) ou multiclasse, incorporando níveis intermediários como "enganoso", "fora de contexto", "impreciso", etc. [52]. O conjunto de dados LIAR, por exemplo, usa seis níveis. A Tabela 2.1 mostra a variedade de rótulos usados por agências brasileiras, refletindo diferentes graus de veracidade.
- **Plataforma:** A análise pode focar em uma plataforma específica (Twitter, Facebook, WhatsApp) ou em fontes de notícias tradicionais (web). As características da *fake news* e sua propagação podem variar significativamente entre plataformas [116].

| Agência | Rótulos |
|---|---|
| Projeto Comprova | enganoso (183), **falso (159)**, comprovado (9), evidência comprovada (6), contexto errado (4) |
| Estadão Verifica | **falso (299)**, enganoso (160), fora de contexto (134) |
| Aos Fatos | **falso (2522)**, verdadeiro (637), impreciso (394), exagerado (239), distorcido (125), insustentável (120), contraditório (65) |
| Boatos.org | **boato (5523)** |
| G1: Fato ou Fake | **fake (1098)**, fato (394), não é bem assim (346) |
| Agência Lupa | **falso (3209)**, verdadeiro (1469), exagerado (866), verdadeiro mas (723), de olho (222), contraditório (189), subestimado (108), ainda é cedo para dizer (108), insustentável (97) |

**Tabela 2.1:** *Agências brasileiras de verificação de fatos e os rótulos associados no conjunto de dados Central de Fatos [30] (rótulo mais comum destacado em negrito).*

A Tabela 2.1 indica os rótulos usados por agências brasileiras de checagem de fatos do conjunto de dados Central de Fatos [30]. O mais comum é "falso/fake/boato".

Somente a agência Boatos.org possui uma categoria que seria falsa: "boato". As outras cinco agências categorizam meias-verdades, em que quatro possuem mais de uma classe associada a meia-verdade. O Projeto Comprova, por exemplo, usa duas categorias para indicar meias-verdades, "enganoso" e "contexto errado". O Estadão Verifica, de forma análoga, utiliza as categorias correspondentes "enganoso" e "fora de contexto". A agência Lupa e Aos Fatos usam maior granularidade de "meias verdades", como exagerado, "verdadeiro mas", "ainda é cedo para dizer", subestimado e exagerado, no caso da agência Lupa, e distorcido, exagerado e impreciso, no caso do Aos Fatos.

### Verificação de Fatos (*Fact-Checking*)

Como introduzido, a Verificação de Fatos (FC) é uma técnica baseada em conhecimento que avalia a veracidade de uma alegação textual comparando-a com evidências externas provenientes de fontes confiáveis [51].

Historicamente realizada manualmente por jornalistas e organizações especializadas como mostrado na Figura 2.2, a escala da desinformação online impulsionou a pesquisa em Verificação Automática de Fatos (AFC). Entretanto, dada a complexidade da linguagem natural, nuances contextuais e a necessidade de julgamento crítico, muitos sistemas atuais operam de forma semi-automática (SAFC), onde ferramentas computacionais auxiliam o verificador humano, que retém um papel fundamental no processo [78, 127]. O *crowdsourcing* também pode ser empregado para escalar partes do processo manual ou para anotar dados para sistemas automáticos [62].

Será utilizado, nesta dissertação, o conceito de verificação semi-automática. Isso

porque esses processos não são por sua totalidade automatizados, necessitando do papel fundamental de especialistas da área [78, 127]. Quando realizada de forma manual, a técnica pode ser realizada por *crowd-sourcing* ou por especialistas da área. *Crowd-sourcing* pode ser realizada com o apoio de plataformas como o Amazon Mechanical Turk [1], em que se contratam pessoas de forma remota para a anotação dos dados.


**Figura 2.2:** *Exemplo de verificação manual de fatos realizada pela agência G1: Fato ou Fake, destacando a alegação (azul), o rótulo de veracidade (amarelo) e a evidência utilizada (vermelho)* [2].

O processo de AFC/SAFC é tipicamente modelado como um *pipeline* composto pelas seguintes etapas [51]:

1. **Identificação/Extração de Alegação (*Claim Detection/Extraction*):** Identificar sentenças ou trechos dentro de um texto (e.g., artigo de notícia, postagem em rede social) que contenham alegações factuais verificáveis, distinguindo-as de opiniões ou conteúdo não factual. Esta etapa é crucial e, como explorado nesta dissertação, pode ser abordada com o uso de LLMs.
2. **Recuperação de Evidências (*Evidence Retrieval*):** Dada uma alegação identificada, buscar em um vasto repositório de informações (como a web, bases de conhecimento ou arquivos de notícias) por documentos ou trechos de texto que possam servir como evidência para confirmar ou refutar a alegação. Esta etapa depende fundamentalmente de técnicas de Recuperação de Informação (RI), muitas vezes utilizando APIs de mecanismos de busca.
3. **Predição de Veredito (*Verdict Prediction*):** Com base na alegação e nas evidências recuperadas, determinar a veracidade da alegação. Isso envolve analisar a relação entre a alegação e cada evidência (suporta, refuta, não relacionado) e, em seguida, agregar essas análises para chegar a um veredito final (e.g., verdadeiro, falso,

[1] https://www.mturk.com/

[2] https://g1.globo.com/fato-ou-fake/noticia/2024/05/22/e-fake-que-a-receita-federal-apreendeu-avioes-com-doacoes-para-o-rs.ghtml

enganoso). Modelos de PLN são usados para avaliar a relação semântica entre alegação e evidência.

4. **Geração de Justificativa (*Justification Generation*):** Explicar o porquê de um determinado veredito ter sido atribuído, idealmente selecionando os trechos de evidência mais relevantes que o suportam. Esta etapa visa aumentar a transparência e a interpretabilidade do processo de verificação.

A construção de *corpora* anotados com alegações, evidências e vereditos é fundamental para treinar e avaliar modelos para cada uma dessas etapas, especialmente para a predição de veredito e a análise da relação alegação-evidência. Com isso, *pipeline* canônico serve como a inspiração fundamental para a metodologia de enriquecimento proposta neste trabalho, conforme detalhado no Capítulo 4. As etapas de Identificação/Extração de Alegação e Recuperação de Evidências são o foco central do nosso fluxo, apresentado na Figura 4.1.

## 2.2 Processamento de Linguagem Natural (PLN)

O Processamento de Linguagem Natural (PLN) é um campo interdisciplinar, envolvendo Ciência da Computação, Inteligência Artificial e Linguística, que se dedica a capacitar computadores a processar, analisar, compreender e gerar linguagem humana (texto e fala) de forma significativa [19]. O PLN é fundamental para diversas aplicações, incluindo tradução automática, análise de sentimentos, sistemas de diálogo, sumarização de texto e, para este trabalho, a análise e verificação de informações.

Sistemas de PLN são frequentemente desenvolvidos e avaliados com base em **tarefas** específicas, como classificação de texto, reconhecimento de entidades nomeadas, resposta a perguntas (*question answering*) ou geração de texto. As tarefas centrais abordadas nesta dissertação, no contexto da verificação de fatos, são a extração de alegações e a futura análise da relação alegação-evidência.

O desenvolvimento e a avaliação de modelos de PLN dependem de conjuntos de dados textuais, conhecidos como ***corpus*** (singular) ou ***corpora*** (plural). A qualidade e as características dos *corpora* (e.g., tamanho, diversidade, tipo de anotação) influenciam diretamente o desempenho dos modelos treinados sobre eles. Os *corpora* utilizados e enriquecidos neste trabalho são detalhados posteriormente (ver Seção 5.1).

### 2.2.1 Modelo de Linguagem (LM)

Um Modelo de Linguagem (LM) é um modelo computacional probabilístico ou neural treinado para compreender e/ou gerar linguagem natural. Fundamentalmente, um LM atribui probabilidades a sequências de palavras ou aprende representações vetoriais

(*embeddings*) de palavras, sentenças ou documentos, capturando propriedades sintáticas e semânticas da língua [94]. Dada uma entrada textual $x$, o LM a transforma em uma representação numérica $y$ (um vetor ou conjunto de vetores) que pode ser utilizada por algoritmos de aprendizado de máquina para realizar tarefas de PLN.

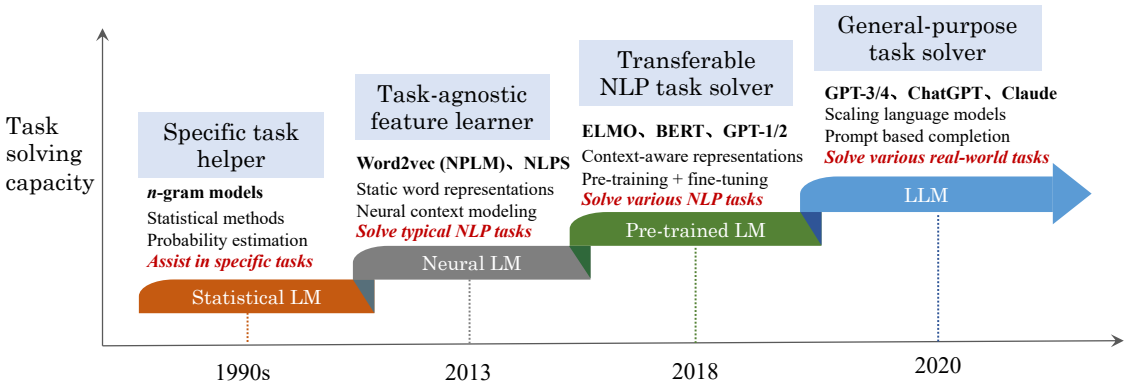


**Figura 2.3:** *Linha do tempo ilustrando a evolução dos modelos de linguagem e sua crescente capacidade de resolver tarefas complexas [137].*

A Figura 2.3 ilustra a evolução dos modelos de linguagem. Em cada geração de arquitetura, aumentou-se a habilidade dos sistemas de realizarem tarefas complexas, geralmente à custa de maior necessidade de dados e recursos computacionais. As principais famílias de LMs incluem:

**LMs Estatísticos:** Modelos baseados em contagem de sequências de palavras ($n$-gramas) ou frequências de termos (e.g., TF-IDF, BM25). São computacionalmente eficientes e não requerem (ou requerem pouco) treinamento supervisionado, mas têm dificuldade em capturar o significado semântico profundo ou lidar com a ordem das palavras e dependências de longo alcance. Motores de busca tradicionais, como o Lucene[3], frequentemente utilizam LMs estatísticos em seus índices [84].

**LMs (Neurais) Estáticos:** Primeiros modelos de linguagem baseados em redes neurais que aprenderam representações vetoriais densas (*embeddings*) para palavras, como o Word2vec [81], GloVe [96], FastText [13]). Esses modelos capturam relações semânticas e sintáticas (como a analogia "rei - homem + mulher = rainha", ilustrada na Figura 2.4), mas atribuem um único vetor a cada palavra, independentemente do contexto em que ela aparece, limitando sua capacidade de lidar com polissemia (múltiplos significados) e ambiguidades. Em português, o NILC disponibilizou *embeddings* estáticos treinados em grandes *corpora* [54].

[3] https://lucene.apache.org/
[4] https://ai.engin.umich.edu/2018/07/23/word-embeddings-and-how-they-vary/

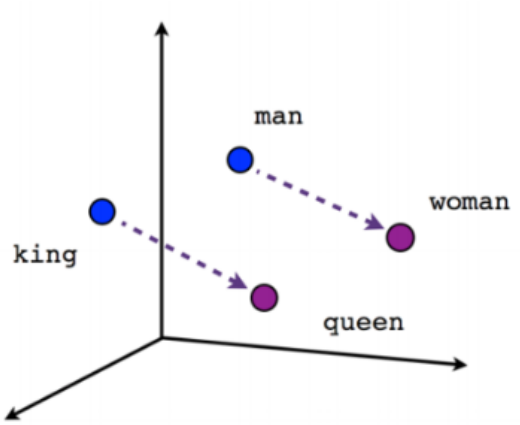


**Figura 2.4:** *Analogia na representação de LMs estáticos* [4]

**LMs (Pré-treinados) Contextuais:** Modelos baseados na arquitetura Transformer [126], como BERT [39] e RoBERTa [73], constituem uma classe importante de modelos de linguagem. Estes são pré-treinados em vastas quantidades de texto não rotulado da ordem de terabytes utilizando tarefas auto-supervisionadas, sendo a mais comum a predição de palavras mascaradas (*Masked Language Modeling* - MLM).

O mecanismo de auto-atenção, intrínseco aos Transformers, permite que a representação vetorial (*embedding*) de cada *token* (palavra ou sub-palavra) seja computada dinamicamente com base em seu contexto textual. Isso possibilita a geração de *embeddings* contextuais que capturam nuances semânticas, distinguindo, por exemplo, entre diferentes acepções de uma palavra polissêmica (como "banco" enquanto assento ou instituição financeira). Tipicamente compostos por centenas de milhões de parâmetros, esses modelos geralmente necessitam de uma etapa subsequente de ajuste fino (*fine-tuning*) sobre um conjunto de dados rotulados específico da tarefa alvo (frequentemente na ordem de milhares de exemplos) para otimizar seu desempenho em aplicações específicas [39].

No domínio da língua portuguesa, destaca-se o BERTimbau [114], uma adaptação do BERT otimizada para o português brasileiro. Desenvolvido a partir do modelo BERT multilíngue (*base*), o BERTimbau foi submetido a pré-treinamento adicional utilizando o *corpus* BrWAC [128], composto por aproximadamente 3,5 milhões de páginas web em português. Na presente pesquisa, a versão *base* do BERTimbau é empregada como um modelo representativo da abordagem de *fine-tuning*, conforme detalhado nas Seções 4.5 e 5.2.

Na literatura mais recente, modelos com a escala de parâmetros do BERT e seus derivados são frequentemente classificados como Modelos de Linguagem Pequenos (*Small Language Models* - SLMs), em contraste com os Modelos de Linguagem Grandes (*Large Language Models* - LLMs) [137]. Apesar da denominação "pequenos", os SLMs continuam a representar o estado da arte em diversas tarefas de Processamento de

Linguagem Natural, especialmente na classificação de textos, quando se dispõe de um volume suficiente de dados rotulados para o ajuste fino [99, 133, 88].

**Modelos de Linguagem Grandes (LLMs):** Os Modelos de Linguagem Grandes (*Large Language Models* - LLMs) constituem uma evolução dos LMs contextuais, distinguindo-se principalmente pela sua escala massiva, com um número de parâmetros na ordem de bilhões ou mesmo trilhões, e por serem treinados em volumes de dados textuais ainda mais extensos [137]. Exemplos proeminentes incluem a família de modelos GPT [16].

Uma característica marcante dos LLMs são suas habilidades emergentes, notadamente a capacidade de realizar tarefas novas com pouquíssima ou nenhuma demonstração prévia (*zero-shot* ou *few-shot learning*), respondendo diretamente a instruções fornecidas em linguagem natural (*prompts*). Diferentemente dos SLMs, os LLMs não requerem, necessariamente, um ajuste fino específico para cada nova tarefa, o que os torna modelos mais generalistas e flexíveis. Nesta dissertação, o LLM Gemini 1.5 Flash [123, 103] é empregado especificamente para duas finalidades: a extração de alegações factuais dos *corpora* e a avaliação do impacto do enriquecimento de dados na classificação de notícias, utilizando a abordagem de *few-shot learning*, conforme detalhado nas Seções 4.5 e 5.2.

## 2.3 Modelos de linguagens grandes (LLM)

A Figura 2.5 mostra a hierarquia de customização de um LLM. Inicialmente, o modelo é pré-treinado em trilhões de palavras, o que equivaleria a um humano lendo sem parar por 8 mil anos, de forma a aprender um ou mais idiomas. São tarefas auto-supervisionadas, em que o próprio processamento do texto é diretamente uma resposta, como predizer a próxima palavra. Essa etapa representa 90% do custo computacional, o que representa a eletricidade consumida por 200 brasileiros em um ano [92].

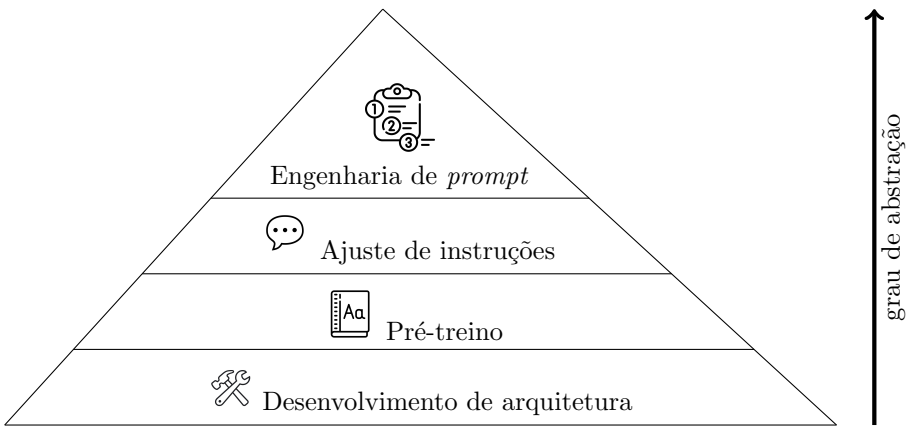


**Figura 2.5:** *Hierarquia da customização de um LLM*[5].

Posteriormente, o modelo é feito o ajuste fino para tarefas conversacionais para se tornar um assistente com 1 milhão de exemplos. A última etapa seria como adaptar a instrução do modelo no uso em inferência para gerar a resposta, conhecido como engenharia de *prompt* [92].

### 2.3.1 Engenharia de *Prompt*

A Engenharia de *Prompt* (*Prompt Engineering*) é a prática de projetar cuidadosamente as entradas (instruções ou *prompts*) fornecidas a um LLM para obter as saídas desejadas [137]. Como LLMs são sensíveis à forma como a tarefa é descrita, um bom *prompt* pode melhorar significativamente o desempenho do modelo em uma tarefa específica, sem a necessidade de re-treinamento. A formulação de um *prompt* eficaz pode ser vista como análoga aos elementos da comunicação descritos por Jakobson [57], como ilustrado na Figura 2.6.

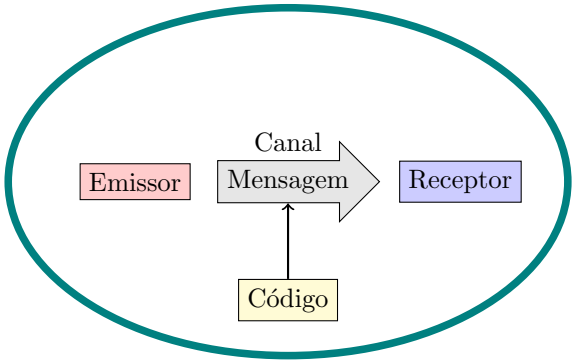


**Figura 2.6:** *Os elementos básicos da comunicação segundo Jakobson [57], adaptado de [82]. Estes elementos podem inspirar a construção de prompts eficazes para LLMs.*

**Emissor :** A entidade que formula a instrução para o LLM.
**Mensagem :** O conteúdo do *prompt*, incluindo a tarefa, o contexto e os exemplos.
**Canal :** A interface de entrada do LLM.
**Receptor :** O LLM que processa o *prompt*.
**Código :** A linguagem natural (e possivelmente formatos específicos) compreendida pelo LLM.
**Contexto :** Informações adicionais ou restrições que guiam a resposta do LLM.

A Figura 2.6 demonstra um exemplo de *prompt* aplicando técnicas como a definição de um papel (*role* - Emissor) e a especificação do formato e tom da resposta (Mensagem, Código, Contexto) direcionada a um público específico (Receptor). A clareza, especificidade e, por vezes, a inclusão de exemplos (*few-shot prompting*) são cruciais

para guiar LLMs eficazmente [80], especialmente em tarefas como a extração precisa de alegações factuais, central neste trabalho.

Sou o CEO de uma empresa de médio porte.

Escreva um e-mail curto, bem-humorado e profissional para meu gerente regional.

Peça a ele que:

- Me envie uma lista atualizada do nosso estoque de papel.
- Organize uma reunião esta semana com outros gerentes regionais.
- Me informe como foi o *workshop* de IA em toda a empresa em seu escritório.

**Figura 2.7:** *Exemplo de um prompt para um LLM, destacando diferentes componentes inspirados nos elementos da comunicação. Adaptado de* `https://learnprompting.org/docs/intro`.

A clareza, especificidade e, por vezes, a inclusão de exemplos são cruciais para guiar LLMs eficazmente [80], especialmente em tarefas complexas como a extração precisa de alegações factuais de textos noticiosos.

## 2.4 Recuperação de Informação (RI)

A Recuperação de Informação (RI), ou *Information Retrieval* (IR), é a área de estudo focada em encontrar material (geralmente documentos) de natureza não estruturada (geralmente texto) que satisfaça uma necessidade de informação a partir de grandes coleções (geralmente armazenadas em computadores) [75]. A RI situa-se na interseção entre ciência da computação, banco de dados, PLN e ciência da informação [84].

No contexto da verificação de fatos, a RI é a tecnologia central para a etapa de Recuperação de Evidências [51]. Dada uma alegação, um sistema de RI é usado para buscar e ranquear documentos (e.g., artigos de notícias, páginas web, posts de blog) de um índice que possam conter informações relevantes para verificar essa alegação.

Quando a busca é textual, o mecanismo de busca está diretamente atrelado aos modelos de linguagem. A consulta (*query*) é transformada em uma representação vetorial pelo modelo de linguagem e comparada via métricas de similaridade com as representações vetoriais dos documentos candidatos [59]. No contexto de LLMs, a integração de RI para fornecer informações externas ao modelo antes da geração de uma resposta é conhecida como *Retrieval Augmented Generation* (RAG) [66, 104, 84].

Mecanismos de busca web, como o buscador Google, são implementações de sistemas de RI em larga escala. Além da relevância textual (calculada usando técnicas de RI), eles consideram muitos outros fatores para ranquear os resultados, como a popularidade e autoridade da página (historicamente medida por algoritmos como o PageRank [95]), a localização e o histórico do usuário, o frescor do conteúdo, entre outros [6]. APIs fornecidas por esses mecanismos de busca (como a API de Busca do Google e a API de busca de alegações do Google FactCheck, utilizadas neste trabalho), conforme mecionado na Seção 4.4, são ferramentas valiosas para implementar a etapa de recuperação de evidências em sistemas de AFC/SAFC, permitindo acesso programático a vastos índices da web e a repositórios de verificações de fatos existentes.

[6] https://developers.google.com/search/docs/fundamentals/how-search-works#ranking

CAPÍTULO 3

# Trabalhos Relacionados

Neste capítulo, são introduzidos os trabalhos relacionados à classificação de *fake news* utilizando técnicas de PLN. De forma geral, apresentamos as técnicas na Seção 3.1. Na Seção 3.2, são explanadas pesquisas de detecção de notícias falsas com o uso de IA generativa em texto e, na Seção 3.3 são explicados trabalhos em português para checagem de texto.

## 3.1 PLN para detecção de notícias falsas

De forma geral, as técnicas em PLN para detecção de notícias falsas podem ser agrupadas pelas seguintes abordagens de análise:

**Linguística e estilo.** As técnicas que analisam a linguística e o estilo examinam dados sintáticos, léxicos, psico-linguísticos e semânticos. No domínio sintático, é examinado o uso de categorias de palavras como substantivos, verbos e adjetivos pela técnica de *POS-tagging*. No domínio léxico, são extraídas diversas características como número de palavras únicas e suas frequências, número de frases e erros gráficos para detectar textos suspeitos. As informações psico-linguísticas podem ser obtidas por meio do sistema estatístico LIWC [105].

**Análise de sentimentos.** Esta abordagem é semântica e procura detectar vieses, como discurso de ódio e toxicidade [2]. O estado da arte na subárea de análise de sentimentos, assim como na grande área de PLN, envolve o uso de modelos contextuais baseados em Transformers [129].

**Verificação de fatos (*Fact-Checking*).** O processo de *fact-checking* automatizado, definido na Seção 2.1.2, é, em alguns aspectos, análogo a sistemas de perguntas e respostas, nos quais, dada uma pergunta (ou alegação), procuram-se referências para gerar uma resposta (ou veredito) [59]. Dentre as abordagens utilizadas no processo de *fact-checking* automatizado, algumas técnicas notáveis são discutidas abaixo.

**Busca estatística.** A busca por evidências pode utilizar métodos baseados em palavras-chave, como o TF-IDF, ou procedimentos análogos ao sistema de busca PageRank [45].

**Busca contextual.** A busca por contra-evidências é frequentemente realizada por modelos contextuais do tipo Encoder, como os Sentence-Transformers, adaptados para tarefas de busca e recuperação de informação semanticamente relevante [45, 38, 104, 59].
**Gerador de resposta extrativo.** A resposta final é gerada pela extração das partes relevantes da contra-evidências pelos modelos Transformers Encoder treinados em tarefas análogas ao conjunto de dados SQuAD [38, 39].
**Modelos generativos.** Como será detalhado na próxima seção, modelos generativos podem ser utilizados em diversas etapas da cadeia de PLN associada a *fake news*, como gerar explicações para as contra-evidências ou gerar uma pergunta que explicite a alegação feita pela notícia, facilitando o processo de busca e verificação [24, 93].

## 3.2 Geração na verificação de fatos

Modelos de Linguagem Grandes (LLMs) têm se tornado peças centrais em tarefas de verificação de fatos. Atuam como uma alternativa mais econômica em comparação a anotadores humanos especializados, como jornalistas ou especialistas da área. Em muitos casos, o desempenho dos LLMs é comparável ao de anotações realizadas via *crowd-sourcing* [74, 89]. Além disso, LLMs permitem a geração de textos explicativos e, frequentemente, não necessitam de treinamento supervisionado específico para a tarefa (abordagem *zero-shot*) [46, 61, 135, 22].

Contudo, modelos como o GPT-3.5 podem apresentar resultados inferiores a modelos de linguagem menores (SLMs) treinados especificamente para a tarefa, como o BERT, na predição de *fake news* sem o uso de evidências externas [55]. Isso indica que, para tarefas de classificação de texto, o estado da arte ainda pode favorecer SLMs treinados com exemplos suficientes (tipicamente entre 100 e 1000 exemplos) [99, 133].

Ao utilizar um LLM para processar alegações com base em seu conhecimento de mundo intrínseco (sem evidência externa) e, subsequentemente, usar um modelo como o BERT para a verificação final, os resultados podem superar o uso isolado do BERT [55]. Incluir no *prompt* instruções para o LLM considerar o estilo de escrita e o conhecimento de mundo também demonstrou melhorar os resultados [72].

No processo semi-automatizado de verificação de fatos [51], LLMs podem ser utilizados em diversas etapas: identificação de alegações [63, 89], recuperação de evidências (muitas vezes delegada a mecanismos de busca externos como Google ou Bing [131, 67, 119]), verificação da alegação (a detecção em si) [119, 26, 64, 131] e justificativa da verificação [61, 135, 64, 131]. A seguir, algumas etapas são detalhadas:
**Identificação de alegação**. Nesta etapa, o modelo de linguagem extrai as informações que devem ser verificadas (alegações) do texto. Essa etapa será detalhada na Seção 3.2.1.

**Verificação da Alegação**. Com a alegação e as evidências relevantes, o LLM pode ser utilizado para detectar se o que foi alegado é falso ou não [119, 26, 120]. A detecção pode considerar também o raciocínio lógico associado, como o uso de operações lógicas [64, 130] e técnicas de *prompting* ligadas ao raciocínio, como a cadeia de pensamento (*chain-of-thought*) [26, 130] e o LLM como um juiz (*LLM-as-a-judge*) [89].
**Justificativa de verificação**. A conclusão da verificação pode ser justificada com o uso de LLMs, que interpretam os insumos (alegação, evidências, veredito) e geram uma explicação para o usuário [61, 135]. Para textos complexos, como na área médica, também são sugeridos processos automáticos de simplificação de textos complexos para torná-los acessíveis à população leiga [109].

### 3.2.1 Extração de Alegação via LLM

Na identificação de alegação, também conhecida como extração de alegação, podem ser extraídas uma alegação principal ou múltiplas alegações. Alternativamente, como a próxima etapa é buscar pela alegação, essa etapa também pode ser vista como geração de consulta (*query generation*) ou geração de perguntas (*question generation*) [25, 93].

Uma técnica intimamente relacionada à extração de alegações é a normalização de alegações (*claim normalization*), que visa simplificar textos, especialmente de redes sociais, removendo redundâncias e ambiguidades para obter a essência da afirmação [117]. A relevância desta tarefa foi destacada na conferência CheckThat! 2025, que organizou uma tarefa compartilhada sobre normalização de alegações em 20 idiomas, incluindo o português. Como produto paralelo desta dissertação de mestrado, a equipe AKCIT-FN participou desta competição, alcançando o pódio em quinze idiomas e obtendo o terceiro lugar em português [5, 6, 118].

A extração de uma única alegação é geralmente denominada detecção de alegação (*claim detection*) [63, 89]. No sentido de extrair múltiplas, chama-se decomposição de alegação (*claim decomposition*) ou quebra de alegações (*claim split*) [121, 58, 110]. Na quebra em múltiplas contextualizações, uma abordagem é remover a dependência semântica entre as alegações, processo conhecido como descontextualização (*decontextualization*) [28, 131].

Pode-se gerar diferentes alegações focando em diferentes entidades no texto, formando distintos pontos focais para gerar mais informações relevantes [93]. Outra forma de realizar a quebra seria por meio de *chain-of-thought* ou lógica de primeira ordem para decompor a alegação em múltiplas perguntas [130].

Único conjunto de dados em português que possui múltiplas alegações é FactNews, em que cada notícia é analisada ao nível de frase. As alegações são extraídas com

o uso de anotadores humanos e os dados são avaliados com o modelo BERT [124].

Ademais, Ni et al. argumentam que existem discrepâncias sobre o que constitui uma alegação: alguns pesquisadores consideram apenas afirmações factuais, enquanto outros incluem também opiniões com impacto social. A maioria dos estudos considera uma alegação como algo digno de verificação (*check-worthy*), mas a definição do que é relevante para checar também é considerada subjetiva [89].

Nesse contexto, Ni et al. listam explicitamente os tipos de fatos a serem considerados como alegações em discursos políticos e instruem o LLM a pensar passo a passo na extração de alegação, usando *chain-of-thought* na extração da alegação [89]. A primeira etapa do raciocínio seria extrair a alegação em si e depois indicar qual tipo de fato a alegação extraída contém. Na Seção 4.3, são explicitados os principais modelos de *prompts* encontrados na literatura.

## 3.3 Trabalhos em português

Em setembro de 2023, foram identificados 18 conjuntos de dados de *fake news* em língua portuguesa mencionados em artigos científicos e disponíveis publicamente, conforme apresentado na Tabela 3.1. Esses conjuntos de dados exibem uma variedade de domínios, fontes e tarefas, embora o ano de publicação esteja concentrado entre 2018 e 2020.

| Publicação | Nome | Domínio | Veículo | Alegação | Evidência | Ano dos dados |
|---|---|---|---|---|---|---|
| 12/2018 | Fake.Br [83] | Geral | Notícias | | ✓ | 2016 - 2018 |
| 10/2019 | Factck.BR [85] | - | Notícias | | ✓ | - |
| 10/2019 | FakeTweet.Br [29] | - | Twitter | ✓ | ✓ | - |
| 12/2019 | Bracis2019FakeNews [44] | Política | WhatsApp, Twitter | | ✓ | 2018 - |
| 07/2020 | fake news Multilabel [35] | Eleições | Notícias | | ✓ | - |
| 11/2020 | FakeNewsSetGen [31] | - | Twitter | ✓ | ✓ | 2017 - 2020 |
| 11/2020 | MM-COVID-PT [68] | COVID-19 | Twitter | ✓ | ✓ | - |
| 06/2020 | FakeCovid-PT [111] | COVID-19 | Redes sociais Digitais | ✓ | ✓ | 2020 |
| 04/2021 | FakeWhatsApp.Br [17] | Eleições | WhatsApp | ✓ | | 2018 |
| 07/2021 | Dataset-fake-news [8] | Política | Notícias | ✓ | | - |
| 10/2021 | COVID19.BR [76] | COVID-19 | WhatsApp | ✓ | | 2020 |
| 10/2021 | Central de Fatos [30] | Geral | Notícias | | ✓ | 2013 - 2021 |
| 02/2022 | MuMiN-PT [90] | Geral | Twitter | ✓ | ✓ | 2020-2022 |
| 03/2022 | FakeRecogna [47] | Geral | Notícias | ✓ | | 2019 - 2021 |
| 03/2022 | Fakepedia [20] | Geral | Notícias | | ✓ | 2013 - 2021 |
| 06/2022 | Fact-check_tweet-PT [60] | - | Twitter | ✓ | ✓ | - |
| 06/2022 | SIRENE-news [11] | Geral | Notícias | | ✓ | 2019 |
| 09/2023 | FactNews [124] | Gerak | Notícias | ✓ | | 2006 - 2007 e 2021 - 2022 |

**Tabela 3.1:** *Conjuntos de dados de fake news em português identificados (até Set/2023).*

Esses conjuntos de dados em língua portuguesa exibem uma variedade de domínios, fontes e tarefas, embora o ano de publicação esteja concentrado entre 2018 e 2020. A análise revela que aproximadamente metade dos trabalhos verifica alegações provenientes de notícias de fontes estabelecidas, enquanto os demais focam em contextos de redes sociais digitais, notadamente Twitter e WhatsApp. Quando mencionados, os

domínios dos dados distribuem-se amplamente, com ênfase em tópicos como saúde (especialmente COVID-19), política/eleições, entretenimento, e assuntos gerais nacionais e internacionais.

No que tange às tarefas de PLN e aos dados fornecidos, a maioria dos *corpora* (12 dos listados) fornece apenas as alegações a serem verificadas, sem evidências associadas. Dentre estes, somente o FactNews decompõe cada exemplo em múltiplas alegações ao nível de frase [124].

Por outro lado, alguns conjuntos [29, 31, 111, 90] oferecem as alegações e as respectivas evidências associadas, geralmente provenientes de agências de verificação de fatos ou fontes confiáveis que validam ou refutam as alegações. No entanto, mesmo os conjuntos que frequentemente fornecem apenas identificadores de tweets ou referências indiretas, tornando o acesso direto às evidências desafiador ou inviável, especialmente após as mudanças na API do Twitter/X em 2023.

| Agência | Região | Menções |
|---|---|---|
| Lupa[1] | Brasil | 8 |
| Boatos.org[2] | Brasil | 7 |
| Aos fatos[3] | Brasil | 5 |
| Projeto Comprova[4] | Brasil | 4 |
| AFP: Checamos[5] | Brasil | 3 |
| G1: Fato ou Fake[6] | Brasil | 2 |
| Estadão Verifica[7] | Brasil | 1 |
| UOL Confere[8] | Brasil | 1 |
| Pública: Truco[9] | Brasil | 1 |
| Observador: Fact-checks[10] | Portugal | 1 |
| E-farsas[11] | Brasil | 1 |

**Tabela 3.2:** *Agências e menções nos corpora.*

As agências verificadoras de *fake news* mencionadas explicitamente nos trabalhos são listadas na Tabela 3.2. Dentre elas, a Lupa e a Boatos.org são as mais citadas, sendo empregadas diretamente em sete e oito trabalhos, respectivamente. Três agências são das maiores mídias jornalísticas tradicionais brasileiras, como sendo G1: Fato ou Fake, Estadão Verifica e UOL Confere.

Três são de portuguesas como Observador: *Fact-checks*, Polígrafo e Viral. AFP: Checamos é uma filial brasileira da agência francesa AFP. Além disso, o Viral é o única agência jornalística de domínio específico, no qual é saúde.

É importante ressaltar que os conjuntos de dados MM-COVID, FakeCovid, MuMinLarge e Fact-check_tweet são multilíngues e incorporam agências de verificação de fatos internacionais provenientes de listas confiáveis, como a Polynter[12] e o Google

[12] https://www.poynter.org/ifcn/

Fact Check[13] [68, 111, 90, 60]. Por exemplo, a Polynter lista agências como Lupa, Observador, AFP e Estadão Verifica. Portanto, embora não mencionem explicitamente as agências lusófonas, esses conjuntos de dados as utilizam indiretamente.

Entre os conjuntos de dados monolíngues em português, o Fake-Recogna extrai informações de agências confiáveis de uma lista internacional chamada Duke Reporter's Labs[14] [30]. O conjunto de dados Central de Fatos menciona a lista Polynter, mas também utiliza verificadores não presentes na listagem [30].

A dificuldade em acessar evidências consistentes e diretamente utilizáveis na maioria dos conjuntos de dados em português existentes motiva a abordagem deste trabalho, que visa não apenas identificar alegações, mas também recuperar e associar novas evidências da web a essas alegações, contribuindo para o enriquecimento dos recursos disponíveis para a pesquisa em verificação de fatos na língua portuguesa.

[13] https://toolbox.google.com/factcheck/explorer
[14] https://reporterslab.org/fact-checking/

CAPÍTULO 4

# Métodos

Este capítulo detalha os procedimentos metodológicos empregados para o enriquecimento de *corpora* destinados à detecção de notícias falsas (*fake news*) em língua portuguesa. O enriquecimento é realizado por meio da incorporação de informações contextuais externas, provenientes de Modelos de Linguagem Grandes (LLMs) e mecanismos de busca.

A Seção 4.1 apresenta o fluxo geral do processo de enriquecimento dos dados. Subsequentemente, a Seção 4.2 descreve os critérios para a seleção dos conjuntos de dados, o pré-processamento aplicado e o processo de validação semi-automática realizado. A Seção 4.3 define o método de extração de alegações centrais dos textos. Adiante, a Seção 4.4 elucida a estratégia adotada para a recuperação de evidências, englobando a busca inicial na *web*, a extração condicional de alegações e a utilização de Interfaces de Programação de Aplicações (APIs) específicas para verificação de fatos. Por fim, a Seção 4.5 explicita a metodologia de avaliação experimental, concebida para mensurar a eficácia das etapas propostas de validação e de enriquecimento com conteúdo externo.

## 4.1 Fluxo de Enriquecimento

O fluxo de enriquecimento proposto é fundamentalmente inspirado no *pipeline* canônico de Verificação Semi-Automática de Fatos (SAFC), conforme detalhado na Seção 2.1.2. O processo de SAFC trenvolve tradicionalmente astapas de Extração de Alegação, Recuperação de Evidências, Predição de Veredito e Geração de Justificativa. O método foca nas duas primeiras etapas — Extração de Alegação e Recuperação de Evidências — essenciais para coletar o contexto externo necessário para a verificação.

A principal inovação metodológica desta dissertação reside na implementação de um fluxo adaptativo, que busca otimizar a eficiência e os custos computacionais. A premissa central é que a etapa de Extração de Alegação, embora crucial para textos longos ou coloquiais, é redundante para textos que já são, em sua essência, alegações concisas e verificáveis. Por exemplo, uma postagem em rede social repleta de opiniões e ruído necessita que sua alegação factual seja isolada para uma busca eficaz [118, 117].

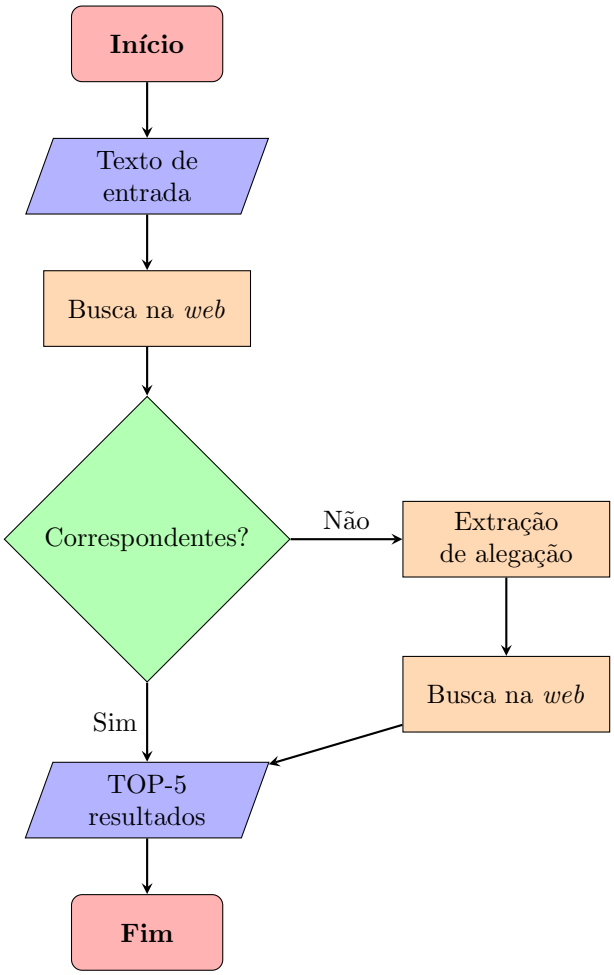


**Figura 4.1:** *Diagrama de fluxo geral do processo de enriquecimento dos corpora.*

Em contrapartida, um texto curto de um *corpus* de verificação de fatos, como "O novo coronavírus pode ser transmitido através de encomendas enviadas da China" da Figura B.1, já funciona como uma consulta de alta qualidade para um mecanismo de busca.

Para operacionalizar essa lógica, o fluxo geral do processo de enriquecimento é ilustrado na Figura 4.1. Inicialmente, uma consulta derivada do texto de entrada é submetida a um mecanismo de busca na *web*. Avalia-se a correspondência entre os resultados obtidos (considerando os cinco primeiros) e o texto original. Se for identificada uma correspondência significativa, assume-se que a busca inicial recuperou informações diretamente relevantes, e os resultados são armazenados. Caso contrário, procede-se à etapa de extração de alegação, na qual um LLM é utilizado para identificar a afirmação central do texto de entrada. Esta alegação extraída é, então, empregada como uma nova consulta para uma segunda busca na *web*. Adicionalmente, realiza-se uma busca específica em uma API de verificação de fatos.

**Busca na web.** Optou-se pelo uso da API oficial de busca do Google, Google Custom Search Engine (CSE), por conta dos créditos gratuitos iniciais. Adicionalmente,

os textos foram buscados na API de busca por alegações do Google, Google FactCheck Claim Search, [1] de forma gratuita. Caso o texto inicial a ser buscado no Google FactCheck não retornar resultados e houver alegação previamente extraída na busca principal, a alegação será buscada no Google FactCheck.

**Correspondência.** Avalia-se a presença dos termos da consulta com exceção de *stopwords*) nos fragmentos destacados pelos marcadores HTML `<b>` e `</b>` dos resultados da API do Google CSE. Na prática, se 80% ou mais desses termos estiverem presentes nos fragmento em destaque, o processo de extração de afirmações não é executado.

**Extração de alegação.** Para a extração de alegação, são extraídas até as 75 primeiras palavras. Dos cinco primeiros resultados obtidos. Optou-se pelo uso do Gemini 1.5 Flash por conta dos créditos gratuitos do Google Cloud. Foi utilizado também o framework Langchain[21] para a engenharia de *prompt*. Para a extração de alegação, são extraídas até as 75 primeiras palavras.

A metodologia foi desenhada para extrair uma única alegação central do texto, simplificando o processo a uma única busca subsequente por evidências. Consequentemente, a abordagem não contempla a separação de múltiplas alegações (*claim splitting*), uma simplificação intencional do fluxo que será discutida como limitação na Seção 6.1. Adicionalmente, a reprodutibilidade e a generalização dos resultados são afetadas pela dependência de tecnologias específicas (Gemini 1.5 Flash, APIs do Google), um desafio agravado pela ausência de uma semente determinística (*seed*) na API do Gemini e pela natureza personalizada da busca do Google.

## 4.2 Conjunto de dados

Selecionaram-se conjuntos de dados de detecção de *fake news* em textos com os seguintes critérios: (i) públicos e acessíveis, (ii) possuam pelo menos um mil exemplos, (iii) possuam artigos publicados. Foram excluídos conjuntos de dados que consistiam somente em reportagens de agências verificadoras sobre *fake news*, em vez das notícias ou alegações originais.

Com base nesses critérios, três *corpora* em língua portuguesa foram selecionados, cujas características são resumidas na Tabela 4.1:

1. **Fake.br** [83]: Composto por notícias de páginas web, abrangendo domínios gerais, coletadas entre 2016 e 2018. Utiliza uma abordagem *bottom-up*, coletando diretamente o conteúdo das páginas. Para cada *fake news* obtida, foi extraídas uma notícia

[1] https://developers.google.com/fact-check/tools/api#the-google-factcheck-claim-search-api

| Conjunto de dados | Domínio | Veículo | Abordagem [52] | Tamanho por rótulo | Ano dos dados | Agências mencionadas |
|---|---|---|---|---|---|---|
| Fake.br [83] | Geral | Páginas da web | *bottom-up* | 3600 `true`<br>3,600 `fake` | 2016 – 2018 | Estadão<br>Folha<br>G1 |
| COVID19.BR [76] | Saúde | WhatsApp | *bottom-up* | 1987 `true`<br>912 `fake` | 2020 | Boatos.org<br>Lupa |
| MuMiN-PT [90] | Geral | Twitter (atual X) | *top-down* | 1339 `fake`<br>65 `true` | 2020 – 2022 | AFP<br>Aos Fatos<br>Comprova<br>Observador<br>O Globo<br>Piauí<br>Uol |

**Tabela 4.1:** *Conjuntos de dados selecionados para o estudo. As abordagens de coleta seguem a classificação por [52], em que a **top-down** envolve a coleta de publicações para fake news conhecidos e de longa data (frequentemente a partir de sites de checagem de fatos), e a **bottom-up** consiste em reunir todas as publicações relevantes de um determinado período para identificar a fake news emergente. As publicações originais listam fontes de verificação, mas não fornecem os excertos específicos de evidência usados para checar cada item.*

verdadeira correspondente. Com isso, o *corpus* consiste em 3600 pares de *fake news* e a notícia verdadeira associada.

2. **COVID19.BR** [76, 37]: Contém mensagens da plataforma WhatsApp de 236 grupos públicos focadas no tema da saúde (COVID-19) entre junho e abril de 2020. Também segue uma abordagem *bottom-up*. Possui o campo de fonte da agência mas é nulo em 98% dos dados.
3. **MuMiN-PT** [90]: Formado por *tweets* (da plataforma X, anteriormente Twitter) de domínio geral, coletados entre 2020 e 2022. Ele representa um contraste metodológico crucial com os outros, pois foi coletado por meio de uma abordagem top-down [52], encontrando publicações correspondentes a alegações verificadas por checadores de fatos. Trata-se de um subconjunto em língua portuguesa do *corpus* multilíngue MuMiN, extraído utilizando a ferramenta Lingua [115].

A diversidade desses conjuntos de dados – em termos de fonte (notícias web, WhatsApp, X/Twitter), domínio temático (geral, saúde), método de coleta (*top-down*, *bottom-up*) e período temporal – foi considerada vantajosa para avaliar a generalização da abordagem de enriquecimento proposta.

### 4.2.1 Limpeza e Validação Semi-automática dos Dados

A qualidade e a confiabilidade dos dados foram garantidas por um pipeline de validação semiautomático. Essa metodologia integrou rotinas automatizadas a uma

curadoria manual detalhada, realizada por um dos autores, visando identificar e retificar inconsistências. Exemplos práticos que ilustram cada etapa do tratamento são detalhados no Apêndice A. As fases do processo foram:

1. **Filtragem Inicial Automatizada:** Remoção de duplicatas exatas, exemplos compostos inteiramente por URLs e textos excessivamente curtos [2].
2. **Filtragem de Idioma:** Identificação automática do idioma português com a biblioteca Lingua [115], seguida de verificação manual e exclusão de exemplos em outros idiomas.
3. **Remoção de Rótulos Conflitantes em Pares Relacionados:** Revisão manual de instâncias com alta similaridade textual, mas com rótulos de veracidade divergentes. A detecção desses pares foi feita por dois métodos:
   i. **Quase-duplicatas:** Identificadas via algoritmo MinHash, como explicado em detalhe na Seção 5.1.3.
   ii. **Referências de URL:** Identificadas quando textos distintos mencionavam a mesma URL.
4. **Verificação de Rótulos com Fonte Externa:** Revisão manual de exemplos cujos rótulos conflitavam com informações do Google FactCheck.
5. **Inspeção de Subconjunto Aleatório:** Verificação manual de um subconjunto aleatório de cada conjunto de dados para avaliar a qualidade geral.
6. **Tratamento Específico para o Fake.br:**
   (a) **Remoção de Quase-Duplicatas da Mesma Fonte:** Exclusão de textos quase-idênticos que provinham da mesma URL de origem, uma vez que, neste *corpus*, a fonte primária de cada notícia é uma URL
   (b) **Correção de Exemplos Incompletos:** Remoção de exemplos da versão normalizada do *corpus* que não continham o trecho original completo da notícia [3].
   (c) Remoção de textos quase duplicados provenientes da mesma URL de origem, uma vez que, neste *corpus*, a fonte primária de cada notícia é uma URL.
   (d) **Remoção de par correspondente:** Dado que o Fake.br é estruturado em pares de notícias (uma falsa e uma verdadeira sobre o mesmo evento), exemplos cujo par correspondente havia sido removido em etapas anteriores do pré-processamento também foram excluídos para manter a integridade da estrutura pareada.

   Como etapa final de pré-processamento, todas as URLs mencionadas explicitamente nos textos originais foram removidas. Esta medida visou mitigar potenciais

[2] Com menos de 15 tokens removendo *emojis*, URLs, *stopwords* e pontuações ao utilizar o tokenizador `cl100k_base` da biblioteca tiktoken https://github.com/openai/tiktoken

[3] https://github.com/roneysco/Fake.br-Corpus/issues/7

vieses introduzidos pelos domínios das URLs, que poderiam ser aprendidos pelos modelos de forma espúria. Uma análise preliminar no *corpus* COVID19.BR, por exemplo, indicou que a simples presença de certos domínios estava fortemente correlacionada com o rótulo de veracidade (conforme Tabela 5.2).

A Tabela 4.2 detalha a quantidade de exemplos que foram corrigidos ou removidos em cada etapa do fluxo de validação semi-automático descrito:

| Etapa de Validação | COVID19.BR | Fake.br | MuMiN-PT |
|---|---|---|---|
| Filtragem Inicial Automatizada | 804 | 1 | 0 |
| Filtragem de idioma | 8 | 0 | 11801 |
| Resolução de Contradições | 20 | 0 | 0 |
| Verificação Externa de Rótulos | 23 | 0 | 4 |
| Inspeção de Subconjunto | 88 | 0 | 0 |
| Tratamento específico Fake.br | 0 | 61 | 0 |

**Tabela 4.2:** *Quantidade de exemplos corrigidos ou removidos nos corpora durante o fluxo de validação semi-automático.*

## 4.3 Extração de alegação

Foram pesquisados quais *prompts* estão reportados na literatura na área de extração de alegação e de extração da consulta (*query extraction*), definidos na Seção 3.2.1. Buscou-se no Google Scholar em meados de maio de 2024 por palavras-chave `claim detection`, `claim split`, `claim decomposition`, `query generation` e `query extraction`. Foram encontrados seis padrões principais de *prompts*:

**Detecção de alegação [63]** O que o texto alega no geral?

**Separação de alegação [121, 131, 58, 110]** Definimos uma afirmação como uma "unidade elementar de informação em uma sentença, que não precisa ser dividida ainda mais." Segmente os fatos do texto a seguir.

**Uso de perspectiva (*role*) [63]** Você é um assistente que ajuda os jornalistas verificarem *fake news*.

**Explicitar as categorias de fatos [89]** Categorias de fatos: C1. Mencionar que alguém (incluindo ele(a) próprio(a)) fez ou está fazendo algo. C2. Citando quantidades e estatísticas. C3. Alegando correlação ou relação causal. C4. Afirmando leis existentes ou regras de operação. C5. Prometer um plano específico para o futuro ou fazer previsões específicas sobre o futuro.

**Extração de consulta [1]** Imagine que você está navegando na internet e topa com uma notícia e dá uma lida rápida. O que você buscaria no Google para ver se o que você

entendeu no geral da notícia é verdadeira? Extraia o que você buscaria do texto a seguir.

***Few-shot* [110]** ENTRADA: O rover Perseverance da NASA descobriu vida microbiana antiga em Marte, de acordo com um estudo recente publicado na revista Science. SAÍDA: 'afirmações': ["O rover Perseverance da NASA descobriu vida microbiana antiga.", "Vida microbiana antiga foi descoberta em Marte.", "A descoberta foi feita de acordo com um estudo recente.", "O estudo foi publicado na revista Science."]

Para cada padrão de *prompt* foram testadas algumas variações em poucos exemplos dos conjuntos de dados. Não foi utilizado o cenário de múltiplas alegações, o qual envolve a separação de alegação (*claim split*), definido em 4.3, pois geraria mais de uma busca para ser verificada. Também não foi testado o impacto de explicitar a categoria de fatos.

Qual principal fato exposto no texto?

1. Extraia um trecho de até 20 palavras do texto a seguir.
2. Retorne somente a alegação e sem título

Texto: `{TEXTO DE ENTRADA}`
Alegação:

**Figura 4.2:** *Prompt final para a extração de alegação.*

A Figura 4.2 mostra o *prompt* desenvolvido para a extração de alegação. Verificou-se que não foi necessário acrescentar o uso de perspectiva (*role*) ou exemplos de *few-shot* para que o modelo extraísse corretamente as alegações em um conjunto de 10 exemplos de teste.

Para simular uma leitura rápida e fornecer contexto ao LLM, utilizou-se como `{TEXTO DE ENTRADA}` até os três primeiros parágrafos do texto original ou, aproximadamente, até 50 palavras, caso o texto fosse mais longo. Para o retorno, exigiu-se que a saída do modelo (*Alegação*) contivesse até 20 palavras, visando uma consulta de busca concisa e focada.

Ao utilizar o Gemini 1.5 Flash para extração de alegação, a configuração de segurança foi desativada[4] para evitar filtragem excessiva de conteúdo. Adicionalmente, observou-se que a documentação da API do LLM (consultada em 29/06/2024) não oferecia opção para fixar uma semente (*seed*) de geração, o que impede o determinismo completo dos resultados para fins de reprodutibilidade[5].

[4] https://ai.google.dev/gemini-api/docs/safety-settings
[5] https://cloud.google.com/vertex-ai/generative-ai/docs/model-reference/gemini?hl=pt-br

## 4.4 Mecanismos de Busca

A recuperação de evidências externas é realizada por meio de dois mecanismos de busca principais: a API Google Custom Search Engine (CSE) para buscas genéricas na *web* e a API Google FactCheck Claims Search, especializada na busca por alegações já verificadas. Ambas as APIs são consultadas com o texto pré-processado (ou a alegação extraída) em português, solicitando-se os cinco primeiros resultados.

### 4.4.1 Pré-processamento da Consulta de Busca

Para simular a criação de uma consulta única a partir de texto puro, foi desenvolvido um pipeline de pré-processamento baseado em heurísticas. Este processo é implementado pela função `preprocess_query` em Python, apresentada na Figura 4.3.

Em textos com até 20 palavras, o conteúdo integral foi usado como consulta. Para textos mais longos, a primeira sentença era extraída. Se ela fosse longa o suficiente para conter a alegação principal (7 ou mais palavras), tornava-se a consulta. Contudo, se a primeira sentença fosse muito curta (menos de 7 palavras), a consulta era formada pelo que fosse mais longo: o primeiro parágrafo completo ou as 20 primeiras palavras do texto.

Na frase de busca, foram removidos os caracteres UNICODE de aspas simples e duplas, pois alteram o comportamento da busca do Google, forçando correspondência exata de frases[6]. Também foram retirados emojis[7], pois experimentalmente notou-se que esses símbolos pioraram alguns resultados de buscas como o da Figura 5.7.

**Google Custom Search Engine (CSE).** A API Google CSE é utilizada para buscas genéricas na *web*. Além da consulta pré-processada e da solicitação dos cinco primeiros resultados (`num=5`), especificam-se os seguintes parâmetros para refinar a busca: geolocalização do usuário como Brasil (`gl=pt-BR`) e idioma preferencial dos resultados como português (`lr=lang_pt`)[8]. A Figura 4.4 ilustra um exemplo da estrutura de retorno da API CSE. Desta estrutura, são armazenados o título (`title`), o *link* (`link`) e o trecho de texto relevante (`snippet`) de cada resultado.

**Google FactCheck Claims Search** A API Google FactCheck Claims Search é empregada para buscar alegações específicas que já foram verificadas por organizações de checagem de fatos. Os parâmetros utilizados na consulta são o idioma (`languageCode=pt-BR`) e o número de resultados (`pageSize=5`). Diferentemente da CSE, esta API não oferece argumentos de localidade adicionais para o português. O pré-processamento do texto de busca é idêntico ao utilizado para a CSE. A Figura 4.5 apresenta um exemplo da estrutura de retorno desta API. Para cada alegação encontrada,

[6] https://blog.google/products/search/how-were-improving-search-results-when-you-use-quotes/
[7] https://home.unicode.org/emoji/about-emoji/
[8] https://developers.google.com/custom-search/v1/reference/rest/v1/cse/list

```python
from nltk import sent_tokenize
import re

SPACES = re.compile(r"\s+")
QUOTES = #todas as representações unicode de aspas

def preprocess_query(text: str) -> str:
    text = text.strip()
    for quote_char in QUOTES:
        text = text.replace(quote_char, "")

    words = re.split(SPACES, text)

    if len(words) <= 20:
        query = text
    else:
        fst_sent = sent_tokenize(text, language='portuguese')[0]

        if len(re.split(SPACES, fst_sent.strip())) >= 7:
            query = fst_sent
        else:
            fst_paragraph = re.split("\n+", text)[0]

            if len(re.split(SPACES, fst_paragraph.strip())) < 20:
                query = " ".join(words[:20])
            else:
                query = fst_paragraph

    return query
```

**Figura 4.3:** *Código de pré-processamento da frase de busca inicial.*

armazena-se apenas o primeiro objeto dentro da lista `claimReview`, que contém a avaliação da veracidade e o texto da alegação verificada[9].

## 4.5 Avaliação dos dados

Com o objetivo de mensurar a eficácia das etapas propostas de validação semi-automática dos dados e de enriquecimento com conteúdo externo, delineou-se um conjunto de configurações experimentais, aplicadas ao COVID19.BR e ao Fake.br, para permitir uma avaliação comparativa e incremental do impacto de cada etapa. O *corpus* MuMiN-PT foi excluído desta fase experimental devido ao severo desbalanceamento de classes resultante após a filtragem para o idioma português, onde restaram apenas 65

[9] https://developers.google.com/fact-check/tools/api#the-google-factcheck-claim-search-api

exemplos da classe minoritária (notícias verdadeiras), conforme detalhado na análise exploratória (ver Figura 5.1)

As configurações avaliadas foram:

1. **Dados Originais (Linha de Base)**: Representa os conjuntos de dados com o mínimo pré-processamento necessário para a modelagem, servindo como ponto de partida comparativo. O processamento aplicado limitou-se à remoção de menções a URLs nos textos, justificada pela análise de viés de domínio, discutida na Seção 4.2.1 e exemplificada na Tabela 5.2. Especificamente para o *corpus* Fake.br, foi empregado o texto normalizado em tamanho, em conformidade com a metodologia de classificação adotada pelos autores originais [83].
2. **Dados Validados:** Corresponde aos conjuntos de dados resultantes da aplicação integral do fluxo de validação semi-automático detalhado na Seção 4.2.1. Esta configuração visa isolar o efeito da melhoria da qualidade intrínseca dos dados (remoção de ruídos, inconsistências, etc.) antes do enriquecimento.
3. **Dados Validados e Enriquecidos com Conteúdo Externo**: Corresponde aos conjuntos de dados obtidos após a execução do fluxo completo ilustrado na 4.1, incorporando os resultados de busca relacionada como informação contextual adicional. Essa configuração foi avaliada em dois cenários distintos para investigar o impacto de diferentes tipos de fontes externas:
   (a) **Enriquecimento Completo:** Considera o primeiro resultado de busca obtido através das APIs do Google (CSE e FactCheck).
   (b) **Enriquecimento Filtrado (Sem Redes Sociais):** Exclui, dos resultados de busca da CSE, aqueles identificados como provenientes de plataformas de redes sociais (e.g., Twitter/X, Facebook). Esta variação permite avaliar se o conteúdo de redes sociais, frequentemente menos curado, introduz ruído ou benefício ao processo de classificação.

Para assegurar a comparabilidade entre as configurações, adotou-se uma divisão padronizada dos dados em conjuntos de treino (80%), validação (10%) e teste (10%) para o COVID19.BR e o Fake.br. Essa padronização foi necessária porque os autores de COVID19.BR e Fake.br não disponibilizaram suas divisões originais, reportando apenas o uso de validação cruzada *5-fold* [77, 83]. A re-divisão garante proporções consistentes para avaliação.

Uma etapa adicional foi implementada para o *corpus* Fake.br: durante a divisão, garantiu-se que os pares de notícias (a *fake news* e sua correspondente notícia verdadeira sobre o mesmo evento) fossem mantidos juntos na mesma partição (treino, validação ou teste). Esta medida é crucial para prevenir o vazamento de informação (*data leakage*)

entre os conjuntos e evitar uma avaliação excessivamente otimista do desempenho do modelo.

O desempenho dos modelos em cada configuração de dados foi aferido por meio de duas abordagens distintas de aprendizado de máquina, descritas na Seção 2.2.1. Estas abordagens foram escolhidas para representar diferentes paradigmas de treinamento e capacidades de modelos de linguagem:

1. Treinamento supervisionado completo (*fine-tuning*) do modelo de linguagem Bertimbau (versão *base*), um representante de Modelos de Linguagem Pequenos (SLMs) adaptados ao português.
2. Aprendizado com poucos exemplos (*few-shot learning*) utilizando o LLM Gemini 1.5 Flash.

## 4.6 Resumo dos métodos

Este capítulo detalhou a metodologia empregada para o enriquecimento de *corpora* de detecção de notícias falsas em língua portuguesa, utilizando contexto externo proveniente de mecanismos de busca e Modelos de Linguagem Grandes (LLMs). O objetivo central foi avaliar o impacto tanto da melhoria da qualidade dos dados quanto da incorporação de evidências externas na tarefa de classificação de veracidade.

O processo iniciou-se com a seleção e validação de três conjuntos de dados distintos — **Fake.br**, **COVID19.BR** e **MuMiN-PT** —, escolhidos por sua diversidade de fontes (notícias web, WhatsApp, X/Twitter), domínios temáticos (geral, saúde), abordagens de coleta (*top-down* e *bottom-up*) e períodos temporais. Foi aplicado um rigoroso fluxo de validação semi-automático para garantir a qualidade dos dados, envolvendo filtragem inicial automatizada, filtragem de idiomas, resolução de rótulos conflitantes em pares relacionados, verificação externa de rótulos e tratamento específico para cada *corpus*. Como medida final, URLs explícitas foram removidas dos textos para mitigar vieses de domínio.

O núcleo metodológico consiste em um **fluxo de enriquecimento adaptativo** inspirado no *pipeline* canônico de Verificação Semi-Automática de Fatos (SAFC). A principal inovação reside na otimização da eficiência computacional: para textos que já funcionam como consultas de alta qualidade, procede-se diretamente à busca na *web*; caso contrário, aciona-se uma etapa de **extração de alegação** via LLM Gemini 1.5 Flash para identificar a afirmação factual central. A correspondência entre consulta e resultados é avaliada através da presença de termos da consulta nos fragmentos destacados pelos marcadores HTML dos resultados do Google CSE. A recuperação de evidências foi realizada por meio de duas fontes complementares: a API **Google Custom Search**

**Engine** para buscas gerais na *web* e a API **Google FactCheck Claims Search** para alegações previamente verificadas por organizações de checagem de fatos.

Para a extração de alegações, foi desenvolvido um *prompt* otimizado que solicita a identificação do principal fato em até 20 palavras, utilizando até os três primeiros parágrafos do texto original como contexto. O pré-processamento das consultas de busca empregou heurísticas específicas baseadas no comprimento do texto, privilegiando a primeira sentença para textos longos quando esta contém informação suficiente (7 ou mais palavras), ou utilizando o texto integral para textos curtos (até 20 palavras).

Por fim, foi delineada uma estratégia de avaliação experimental para mensurar o impacto incremental de cada etapa metodológica. Foram definidas configurações comparativas de dados: (1) dados originais como linha de base, (2) dados validados através do fluxo semi-automático, e (3) dados validados e enriquecidos com contexto externo, incluindo variações com e sem conteúdo de redes sociais. A avaliação foi conduzida nos *corpora* COVID19.BR e Fake.br através de duas abordagens distintas de aprendizado de máquina: *fine-tuning* supervisionado do modelo Bertimbau e *few-shot learning* com Gemini 1.5 Flash, garantindo uma avaliação abrangente da eficácia das metodologias propostas sob diferentes paradigmas de modelagem.

Tendo detalhado a metodologia de validação e enriquecimento, o Capítulo 5 apresentará a aplicação prática deste fluxo. Serão exploradas as características dos dados originais, os resultados quantitativos e qualitativos do processo de enriquecimento e, por fim, a avaliação experimental do impacto dessas etapas.

```
{
    "htmlTitle": "Cadastro em aplicativo do SUS é recomendado, mas não é ...",
    "link":
    ↪ "https://lupa.uol.com.br/jornalismo/2021/01/12/verificamos-cadastro-sus",
    "htmlSnippet": "Jan 12, 2021 <b>...</b> ... todos os brasileiros precisam
    ↪ se cadastrar no aplicativo Conecte ... “<b>Pessoal</b>, <b>todo mundo
    ↪ precisa se cadastrar no conectesus para vacinar</b>.",
    "pagemap": {
        "metatags": [
        {
            "og:type": "article",
            "og:title": "Cadastro em aplicativo do SUS é recomendado, mas não
            ↪ é obrigatório para a vacinação",
            "og:description": "Circula pelo WhatsApp que todos os brasileiros
            ↪ precisam se cadastrar no aplicativo Conecte SUS para receber a
            ↪ vacina contra o novo coronavírus -- caso contrário, não serão
            ↪ imunizados. Por WhatsApp, leitores da Lupa sugeriram que esse
            ↪ conteúdo fosse analisado. Confira a seguir o trabalho de
            ↪ verificação : “Pessoal, todo mundo precisa se cadastrar no
            ↪ conectesus para [...]",
        }
        ],
        "ClaimReview": [
        {
            "datePublished": "2021-01-12",
            "claimReviewed": "Pessoal, todo mundo precisa se cadastrar no",
        }
        ]
    }
}
```


**Figura 4.4:** *Exemplo de um item da lista de resultados da API Google CSE. Alguns campos foram omitidos para concisão. Consulta realizada em 09/07/2024 para a frase “Pessoal, todo mundo precisa se cadastrar no conectesus para vacinar.”.*

```
{
    "claims": [
        {
            "claimDate": "2021-04-20T16:52:25Z",
            "claimReview": [
                {
                    "languageCode": "pt",
                    "publisher": {
                        "name": "Observador",
                        "site": "observador.pt"
                    },
                    "reviewDate": "2021-06-29T09:21:38Z",
                    "textualRating": "Errado",
                    "title": "Fact Check. Vacinas "não criam imunidade contra
                    ↪  a Covid-19"?",
                    "url": "https://observador.pt/factchecks/..../"
                }
            ],
            "claimant": "Utilizador de Facebook",
            "text": "\"Vacinas contra a Covid-19 não criam imunidade\""
        }
    ]
}
```


**Figura 4.5:** *Exemplo de um item da lista de resultados da API Google FactCheck Claims Search. Alguns campos foram omitidos para concisão. Consulta realizada em 08/01/2025 para a frase "Vacinas contra a Covid-19 não criam imunidade.".*

CAPÍTULO 5

# Desenvolvimento

Este capítulo detalha o processo de desenvolvimento e apresenta os resultados obtidos a partir da aplicação da metodologia descrita no Capítulo 4. Inicia-se com uma análise exploratória dos *corpora* originais (Seção 5.1), destacando suas características e limitações intrínsecas. Em seguida, a Seção 5.3 descreve quantitativamente os dados enriquecidos com informações externas via LLMs e mecanismos de busca e a seção 5.4 aprofunda a análise através de um estudo qualitativo dos padrões encontrados nos dados enriquecidos. Os resultados da avaliação dos dados, comparando diferentes configurações de processamento, são detalhados na Seção 5.4. Por fim, a Seção 5.6 oferece uma síntese e discussão integrada dos achados.

## 5.1 Análise Exploratória dos Dados

Antes de proceder ao enriquecimento, realizou-se uma análise exploratória detalhada dos três conjuntos de dados selecionados (Fake.br, COVID19.BR, MuMiN-PT) após a etapa de pré-processamento e validação descrita na Seção 4.2.1.

### 5.1.1 Balanceamento e Características Textuais

A Figura 5.1 mostra o número de exemplos verdadeiros e falsos de cada conjunto de dados processado. Observa-se que o Fake.br apresenta um balanceamento razoável, enquanto MuMiN-PT e COVID19.BR exibem desbalanceamentos distintos: MuMiN-PT possui consideravelmente mais exemplos falsos, e COVID19.BR, mais exemplos verdadeiros.

Uma hipótese para esses desbalanceamentos seria da natureza da obtenção dos dados. Como o MuMiN-PT é uma abordagem *bottom-up*, ou seja, parte das agências que verificam fatos, é muito provável que as agências foquem em notícias falsas e, portanto, possuam mais exemplos falsos. Isso pode ser reforçado pela Tabela 2.1, em que a contagem de rótulos utilizada pelas entidades jornalísticas era, em sua maioria, falsos.

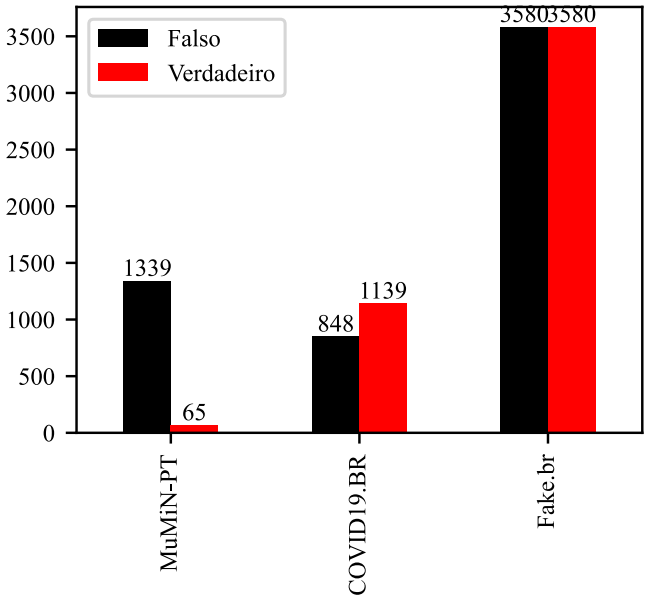


**Figura 5.1:** *Balanceamento dos dados após pré-processamento.*

Já o COVID19.BR, embora coletado via abordagem *top-down* em um contexto de alta circulação de mensagens sobre a pandemia via WhatsApp, resultou em uma maior quantidade de exemplos verdadeiros. Isso sugere que a coleta capturou não apenas *fake news*, mas também um volume significativo de mensagens informativas, alertas oficiais ou discussões factuais que foram classificadas como verdadeiras, ou que o processo de rotulação original teve um critério específico que levou a este resultado.

| Estatística | **MuMiN-PT** | | **Fake.br** | | **COVID19.BR** | |
|---|---|---|---|---|---|---|
| | `fake` | `true` | `fake` | `true` | `fake` | `true` |
| Média núm. de palavras | 18,9 | 16,3 | 181,4 | 183,1 | 167,7 | 111,1 |
| Média. tam. das palavras (em caracteres) | 5,0 | 4,9 | 4,8 | 5,0 | 4,9 | 6,6 |
| Média de núm. de frases | 1,4 | 1,4 | 10,4 | 9,0 | 10,9 | 5,8 |
| Média núm de palavras por frase | 14,5 | 12,3 | 18,6 | 22,1 | 19,2 | 22,9 |
| Presença de URLs | 0,3% | 0,0% | 1,0% | 0,7% | 28,9% | 56,9% |

**Tabela 5.1:** *Estatísticas textuais por conjunto de dados e rótulo.*

A Tabela 5.1 resume as estatísticas textuais. O tamanho médio dos textos varia significativamente, refletindo a plataforma de origem: *tweets* (MuMiN-PT) são os mais curtos, seguidos por mensagens de WhatsApp (COVID19.BR) e notícias web (Fake.br). O tamanho médio das palavras, contudo, mostrou-se notavelmente estável entre os conjuntos e rótulos, pairando em torno de 4,8-5,0 caracteres.

Em termos de rótulos, o MuMiN-PT não varia as estatísticas das palavras e

frases entre os rótulos verdadeiro e falso, possivelmente por conta do limite baixo de caracteres imposto no X. Já no Fake.br, a notícia verdadeira, vinda de sites confiáveis, tende originalmente a ser maior. Para garantir uma classificação justa, o artigo original do Fake.br normaliza o tamanho dos textos em número de palavras, truncando os textos maiores em relação aos menores [83]. Conforme explicado na Seção 4.2, a versão normalizada é utilizada neste trabalho.

No COVID19.BR, por sua vez, tem a tendência inversa do Fake.br, uma hipótese seria por conta dos "textões" do WhatsApp comuns em notícias falsas, além do típico de mensagens menores do WhatsApp. Os autores do conjunto de dados mencionam que os exemplos *fake* tendem a possuir tamanho maior e mais variável, quanto aos diferentes estilos de escrita [76].

O COVID19.BR é o conjunto de dados que mais possui links nos textos, tanto de forma absoluta em 889 exemplos ou relativa (44,7%). Uma possível explicação seria que no ambiente do WhatsApp se compartilhem mais links em textos do que sites e *tweets*. Tanto no COVID19.BR quanto no Fake.br as notícias verdadeiras possuem mais links associados. O MuMiN-PT não possui praticamente links associados nos exemplos, o que é esperado para *tweets*.

| Domínios de URLs mencionadas | `fake` (%) | `true` (%) | Total |
|---|---|---|---|
| **`gazetabrasil`** | 0 (0,0%) | 259 (**100,0%**) | 259 |
| **`bit.ly`** | 6 (6,8%) | 82 (**93,2%**) | 88 |
| `youtube` | 47 (67,1%) | 23 (32,9%) | 70 |
| **`globo`** | 6 (11,3%) | 47 (**88,7%**) | 53 |
| `facebook` | 19 (38,0%) | 31 (62,0%) | 50 |
| `dunapress` | 0 (0,0%) | 46 (**100,0%**) | 46 |
| `twitter` | 25 (56,8%) | 19 (43,2%) | 44 |
| **`whatsapp`** | 1 (3,1%) | 31 (**96,9%**) | 32 |
| `uol` | 11 (37,9%) | 18 (62,1%) | 29 |
| `conexaopolitica` | 16 (59,3%) | 11 (40,7%) | 27 |
| `gov.br` | 6 (26,1%) | 17 (73,9%) | 23 |
| `instagram` | 5 (21,7%) | 18 (78,3%) | 23 |
| `jornaldacidadeonline` | 14 (63,6%) | 8 (36,4%) | 22 |
| **`japinaweb`** | 0 (0,0%) | 21 (**100,0%**) | 21 |
| `atrombetanews` | 7 (35,0%) | 13 (65,0%) | 20 |

**Tabela 5.2:** *Distribuição dos 15 domínios de URL mais frequentes no corpus COVID19.BR por rótulo. A tabela apresenta a contagem absoluta de menções para cada rótulo (`fake`/`true`) e o total por domínio. As porcentagens indicam a proporção de cada rótulo dentro do total de menções daquele domínio. Domínios em negrito são aqueles em que mais de 80% das menções ocorrem em textos com rótulo `true`.*

No contexto de *links* no conjunto de dados COVID19.BR, a Tabela 5.2 mostra os domínios de URLs mais mencionados nos exemplos. Os domínios da Gazeta Brasil, do bit.ly, da Globo, do WhatsApp e o JapinaWeb representam quase sempre exemplos verdadeiros. No entanto, a presença de um domínio governamental não é, por si só, garantia de veracidade. Domínios do governo brasileiro, como `gov.br`, podem ser instrumentalizados em contextos enganosos, como ilustrado por um exemplo do *corpus* MuMiN-PT na Figura 5.2.

**Exemplo de Uso Enganoso de URL Oficial (MuMiN-PT)**

Pessoal, todo mundo precisa se cadastrar no conectesus para vacinar. Sugiro fazer já. Provavelmente o site nao aguentará os acessos quando for o momento. https://conectesus-paciente.saude.gov.br/ É um cadastro no SUS. Quem tomou a vacina da Febre Amarela em 2018 já tem. Ou quem usou o SUS nos últimos anos. O aplicativo funciona mais ou menos como esses apps de carteira de motorista ou título de eleitor.

**Figura 5.2:** *Exemplo do MuMiN-PT onde uma URL oficial é usada em contexto de fake news.* [1]

### 5.1.2 Tópicos Predominantes e Temporalidade

A análise dos termos mais frequentes (após remoção de *stopwords*), ilustrada na Figura 5.3, revela a forte dependência temporal dos conjuntos de dados. O Fake.br (coletado entre 2016-2018) reflete o cenário político pré-pandêmico brasileiro, com menções frequentes ao então presidente Temer.

Por outro lado, COVID19.BR e MuMiN-PT, coletados durante 2020-2022, compartilham tópicos centrados na pandemia de COVID-19, como “covid-19”, “vacina”, “máscaras”) e no cenário político da época (menções ao então presidente Bolsonaro). Embora o MuMiN-PT se declare de domínio geral, a predominância de tópicos relacionados à COVID-19 sugere que eventos de grande impacto global dominam a conversação online, mesmo em conjuntos de dados que não são explicitamente focados neles.

Analisando os termos mais mencionados nos gráficos da Figura 5.3, percebe-se que os conjuntos de dados de detecção de notícias falsas são altamente dependentes do cenário temporal. Isso é característico em tarefas de análise de sentimentos de redes sociais digitais, em que os termos em alta são voláteis, o que gera um vocabulário mais dinâmico e temporal.

O Fake.br é o único que possui data associada a cada exemplo. A distribuição da data dos exemplos é mostrada na Figura 5.4. A metade dos dados é de 2017, um quarto é de 2016, um pouco menos de um quinto de 2018 e menos de 5,0% é entre 2009 a 2015.

[1] https://lupa.uol.com.br/jornalismo/2021/01/12/verificamos-cadastro-sus

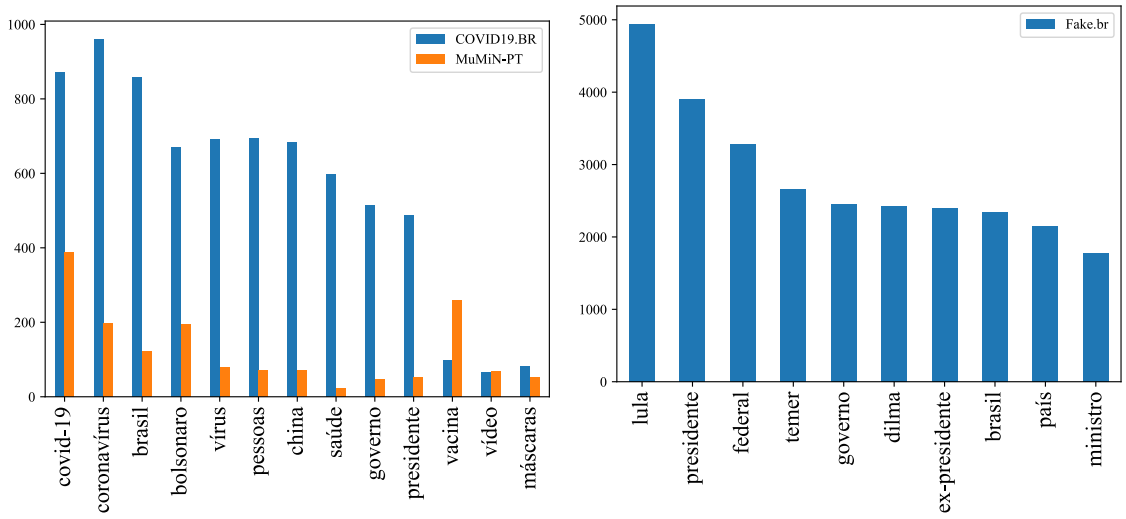


**Figura 5.3:** *10 Termos mais comuns (sem stopwords) nos corpora COVID19.BR/MuMiN-PT (esquerda) e Fake.br (direita).*

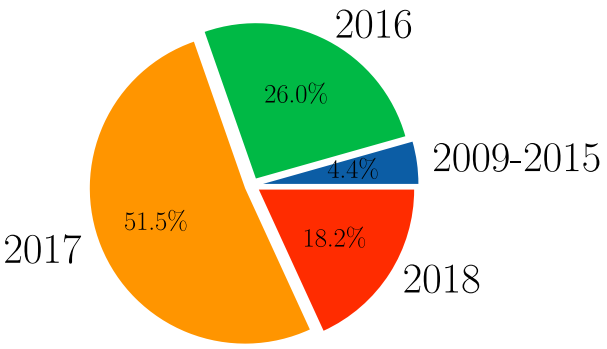


**Figura 5.4:** *Tempo de publicação dos dados do Fake.br*

### 5.1.3 Análise de duplicatas

A presença de textos quase idênticos (quase duplicatas) foi investigada utilizando a técnica MinHash LSH, uma técnica em que trata a semelhança textual como um problema de interseção de caracteres ou palavras, e estima o tamanho relativo das interseções utilizando amostragem aleatória [15, 53].Foi utilizada a biblioteca Akin[2] na versão 0.1.0. Os algoritmos de MinHash com LSH tiveram os seguintes parâmetros: semente aleatória `seed=3`, tipo de n-grama como sendo carácter, tamanho do n-grama `n_gram = 5`, bits do HASH `hash_bits = 128`, número de bandas `no_of_bands = 50` e distância mínima de Jaccard `near_duplicates` de 0,7.

A Figura 5.5 mostra um gráfico de explosão solar dos exemplos que foram

[2] https://github.com/justinbt1/Akin/tree/v0.1.0

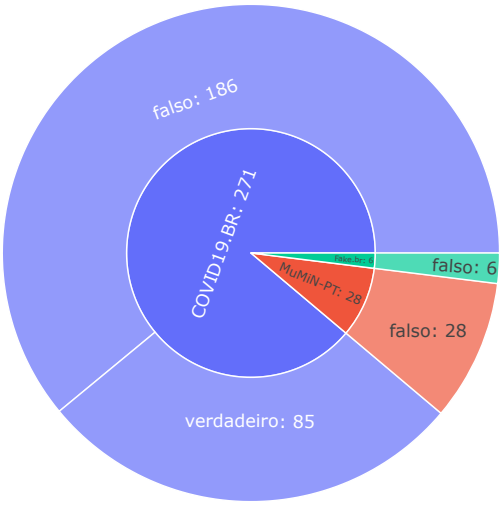


**Figura 5.5:** *Contagem de quase duplicatas entre os corpora e rótulos.*

encontrados outros exemplos quase duplicatas do mesmo *corpus*. O COVID19.BR possui mais exemplos quase duplicados, 271 exemplos (13,6%). Uma hipótese seria o domínio do WhatsApp, em que as conversas curtas podem ser parecidas e redundantes e o compartilhamento de mensagens é alto e fácil. No Fake.br são 6 exemplos, o que corresponde a 0,08% dos dados e o MuMiN-PT são 28 exemplos, correspondendo a 2,00%.

Analisando manualmente os exemplos, notou-se que a mudança entre quase duplicatas seria no contexto de: (1) espaçamento, (2) acréscimos de caracteres como traços "-" ou asteriscos "*" e (3) trocas ou acréscimos de poucas palavras. Na Figura 5.6, são destacados dois exemplos correspondentes do *corpus* COVID19.BR, a diferença entre os dois textos é destacada em amarelo. À esquerda o texto possui duas quebras de linha a mais e à direita possui as palavras "– Conexão Política -".

**Exemplo 1 (COVID19.BR)**

China força países atingidos por vírus chinês a se ajoelharem diante da Huawei: "Nós lhe daremos máscaras se aceitar a Huawei 5G"

https://conexaopolitica.com.br/ultimas/china-forca-paises-atingidos-por-virus-chines-a-se-ajoelharem-diante-da-huawei-nos-lhe-daremos-mascaras-se-aceitar-a-huawei-5g/

**Exemplo 2 (COVID19.BR)**

China força países atingidos por vírus chinês a se ajoelharem diante da Huawei: "Nós lhe daremos máscaras se aceitar a Huawei 5G" – Conexão Política - https://conexaopolitica.com.br/ultimas/china-forca-paises-atingidos-por-virus-chines-a-se-ajoelharem-diante-da-huawei-nos-lhe-daremos-mascaras-se-aceitar-a-huawei-5g/

**Figura 5.6:** *Fake news quase duplicatas no COVID19.BR (diferenças destacadas).*

**Exemplo 1 (Clickbait)**

🔵 Internet gratuita de 500 GB 4G 🔵 Por causa da epidemia de COVID-19, oferecemos 10 GB de Internet grátis, válidos por 90 dias para ajudá-lo a ficar em casa! **Obtenha acesso gratuito à Internet e fique em casa** 👉
https://bit.ly/10gbytes

**Exemplo 2 (Clickbait)**

🔵 Internet gratuita de 500 GB 4G 🔵 Por causa da epidemia de COVID-19, oferecemos 500 GB de Internet grátis, válidos por 90 dias para ajudá-lo a ficar em casa! **Obtenha acesso gratuito à Internet e fique em casa** 👉
http://hu5k.com/Covid

**Exemplo 3 (Clickbait)**

🔵 Internet gratuita de 500 GB 4G 🔵 Por causa da epidemia de COVID-19, oferecemos 500 GB de Internet grátis, válidos por 90 dias para ajudá-lo a ficar em casa! **Obtenha acesso gratuito à Internet e fique em casa** 👉
https://bit.ly/500gbytes

**Figura 5.7:** *Exemplos de clickbait quase duplicados no COVID19.BR (diferenças destacadas).*

Na Figura 5.7, é um exemplo no COVID19.BR de *clickbait*, sites usados para atrair cliques, termo definido na Seção 2.1.1. É divulgada uma internet gratuita 4G de gigas grátis, e varia de 10GB e 50 GB no tweet e o link. O texto é desmentido pela agência Boatos.org[3].

A análise dos tamanhos dos agrupamentos de duplicatas (Figura 5.8) mostra que a maioria forma pares (tamanho 2), mas o COVID19.BR apresenta agrupamentos maiores (até tamanho 12), reforçando a ideia de replicação massiva de certas mensagens falsas nessa plataforma.

A detecção e análise de quase duplicatas, incluindo os casos contraditórios removidos durante a validação (Seção 4.2), foram fundamentais. Elas não apenas ajudaram a limpar os dados, mas também revelaram padrões de propagação de informação (atualizações vs. mutações virais) e destacaram a sensibilidade da classificação a pequenas variações textuais, um desafio importante para modelos de detecção.

### 5.1.4 Limitações Identificadas nos Dados Originais

A análise exploratória expôs limitações inerentes aos conjuntos de dados, que motivam a busca por enriquecimento com informações externas:

[3] https://www.boatos.org/tecnologia/site-da-internet-gratuita-500-gb-4g-compartilhar-link-whatsapp.html

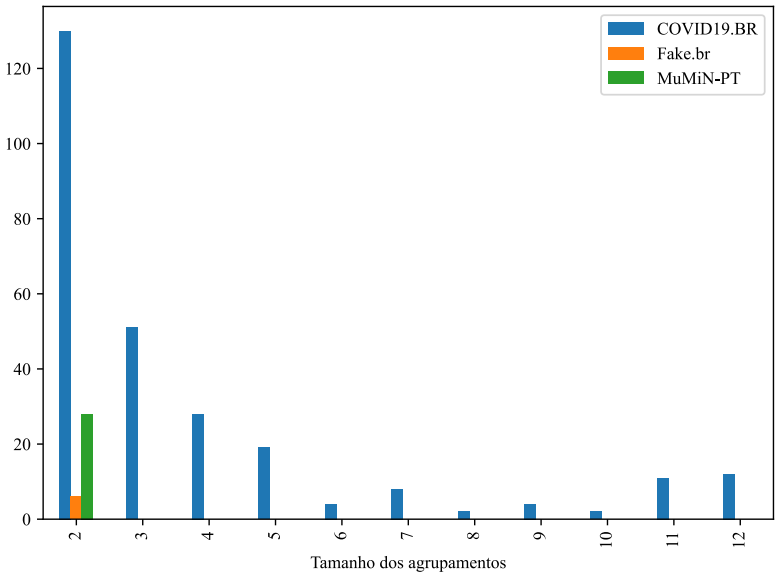


**Figura 5.8:** *Histogramas de tamanho de agrupamentos*

- **Temporalidade dos Tópicos e Rótulos:** Como visto (5.1.2), os assuntos são voláteis. Além disso, a veracidade de uma alegação pode mudar com o tempo (e.g., obrigatoriedade de máscaras). Modelos estáticos enfrentam dificuldades com essa dinâmica.
- **Presença de Quase Duplicatas:** A redundância (5.1.3) pode enviesar o treinamento e a avaliação. As pequenas variações entre duplicatas falsas também sugerem táticas de evasão de detecção.
- **Granularidade dos Rótulos:** A classificação binária (verdadeiro/falso) simplifica a complexidade da *fake news*, omitindo nuances como "enganoso" ou "fora de contexto" usadas por agências (ver Tabela 5.6).
- **Verificabilidade da Alegação:** Alguns textos contêm opiniões ou relatos pessoais não passíveis de verificação factual externa (e.g., "Hj faleceu um senhor com corona vírus aqui").
- **Dependência de URLs:** Muitos exemplos contêm URLs (especialmente no COVID19.BR). A análise do texto isolado ignora o conteúdo dessas páginas vinculadas, que pode ser crucial para a verificação. Um exemplo notório é o uso de um URL oficial do governo (Conecte SUS) para disseminar *fake news*, como mostrado na Figura 5.2. Além disso, alguns domínios estão enviesados no conjunto de dados COVID19.BR como mostra a Tabela 5.2, de forma que o domínio por si só poderia fornecer uma classificação de veracidade.

Estas limitações, em conjunto, reforçam a necessidade de ir além do texto original, buscando evidências externas e contexto atualizado para uma avaliação mais robusta da veracidade, objetivo central do enriquecimento proposto.

## 5.2 Ambiente de Avaliação

Nesta seção, detalha-se o ambiente computacional e as ferramentas empregadas na condução das avaliações experimentais, seguindo a metodologia delineada na Seção 4.5. Os experimentos foram implementados em `Python`, com o auxílio da biblioteca `LiteLLM`[4]. Esta biblioteca foi utilizada para realizar as inferências com o modelo `Gemini 1.5 Flash`, facilitando a interação com a respectiva API.

Para o *few-shot learning*, todos os exemplos recebem os mesmos 15 exemplos aleatórios de entrada. A Figura 5.9 apresenta o *prompt* base utilizado para a inferência pelo modelo.

> A seguir são apresentados textos de mensagens e notícias em português. Sua tarefa é classificar cada texto como contendo uma *Fake News* ou como sendo VERDADEIRO.
> Para auxiliar na classificação, será também fornecido um contexto extra, correspondente a uma busca no Google pelos termos do texto a ser classificado.
> Responda **apenas** com uma das seguintes *tags*: "`FAKE NEWS`" ou "`VERDADEIRO`".

**Figura 5.9:** *Prompt base para a detecção de fake news. A seção sublinhada é incluída quando o contexto extra (oriundo da busca externa) é considerado na análise.*

Os experimentos de ajuste fino (*fine-tuning*) foram executados em um ambiente com acesso a GPUs NVIDIA V100 (32 GB de VRAM) ou NVIDIA A100 (80 GB de VRAM). O *framework* PyTorch foi utilizado, abstraído pela biblioteca SimpleTransformers [102]. Cada execução experimental de ajuste fino demandou, tipicamente, entre 8 GB e 12 GB de memória VRAM da GPU, variando conforme a configuração de dados e o tamanho do lote (*batch size*) utilizado.

A Tabela 5.3 apresenta o espaço de busca de hiperparâmetros configurado para o processo de ajuste fino do modelo Bertimbau base. Esta busca em grade (*grid search*) foi aplicada a cada uma das configurações de dados de treinamento especificadas na Seção 4.5. A seleção da melhor combinação de hiperparâmetros para cada configuração de dados e *corpus* foi realizada com base no maior valor de F1-Score obtido no conjunto de validação.

A definição do número máximo de épocas de treinamento em 10 baseou-se na observação de trabalhos anteriores com modelos BERT e RoBERTa em tarefas de classificação de texto com volumes de dados ou domínios análogos [39, 73, 41, 91, 42, 114]. O valor de *weight decay* (0.01) adotado é consistente com as configurações padrão sugeridas para os modelos BERT e RoBERTa [39, 114, 73].

[4] https://github.com/BerriAI/litellm

| **Hiperparâmetros** | **Espaço de busca / Valor** |
|---|---|
| Batch size | {8, 16} |
| Learning rate | {1e-6, 5e-6, 1e-5,2.5e-5, 5e-5, 1e-4, 2.5e-5, 5e-5, 1e-4} |
| Dropout of task layer | {0.1, 0.2, 0.3} |
| Seed | 2025 |
| Weight decay | 0.01 |
| Maximum training epochs | 10 |
| Maximum Sequence Length | 512 |
| Learning rate scheduler | Linear com 6% de *warmup* |
| Optimizer | AdamW |
| AdamW ε | 1e-8 |
| AdamW $\beta_1$ | 0.9 |
| AdamW $\beta_2$ | 0.999 |
| Early stopping patience | 3 épocas |
| Early stopping threshold (F1-score) | 0.001 |

**Tabela 5.3:** *Espaço de busca de hiperparâmetros para o ajuste fino do modelo Bertimbau base.*

O otimizador AdamW foi escolhido por ser o padrão para modelos da família BERT, utilizando-se os valores canônicos para AdamW $\beta_1$ (0.9) e AdamW $\beta_2$ (0.999) [39, 114, 102]. O parâmetro AdamW ε (1e-8) corresponde ao valor padrão implementado na biblioteca SimpleTransformers [102]. O comprimento máximo da sequência (*Maximum Sequence Length*) foi definido em 512 *tokens*, que é o limite padrão para modelos Bertimbau base e BERT base.

A faixa de valores para a taxa de abandono (*dropout*) na camada de tarefa (0.1, 0.2, 0.3) foi selecionada considerando-se que, em cenários com volumes de dados limitados, taxas de *dropout* ligeiramente mais elevadas podem contribuir para a regularização do modelo e mitigar o sobreajuste (*overfitting*) [43, 50]. Os tamanhos de lote (*batch sizes*) de 8 e 16 foram explorados, alinhando-se com as práticas de trabalhos que aplicaram modelos BERT, RoBERTa e Bertimbau base em *corpora* de dimensão similar ou para detecção de notícias falsas [39, 73, 114, 41, 91, 42].

Por fim, o escalonador da taxa de aprendizado (*learning rate scheduler*) com decaimento linear precedido por uma fase de aquecimento (*warmup*) de 6% das etapas de treinamento foi adotado em conformidade com as recomendações de [73]. A estratégia de parada antecipada (*early stopping*) foi configurada para monitorar o F1-Score no conjunto de validação, interrompendo o treinamento caso não houvesse melhoria superior a 0.001 por 3 épocas consecutivas, prevenindo o sobreajuste e otimizando o tempo de treinamento.

## 5.3 Conjunto de dados enriquecido

Esta seção apresenta os resultados da aplicação do fluxo de enriquecimento (Figura 4.1) aos três *corpora* pré-processados.

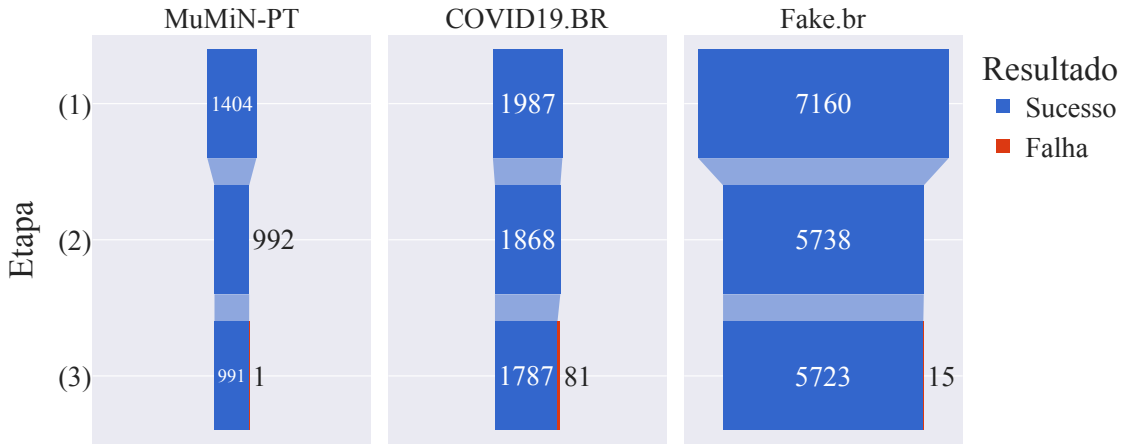


**Figura 5.10:** *Funil do processo de enriquecimento para cada corpus, mostrando o número de exemplos em cada etapa principal (Busca Inicial, Extração de Alegação, Busca pela Alegação).*

A Figura 5.10 sumariza o fluxo de processamento e a quantidade de exemplos em cada etapa para os três *corpora*. A etapa inicial envolve a busca na web pelo texto pré-processado (conforme Seção 4.4). Se não houver correspondência forte (critério de 80% de sobreposição de palavras, desconsiderando *stopwords*), prossegue-se para a extração de alegação via LLM (Gemini 1.5 Flash, com o prompt da Figura 4.2) e uma nova busca na web com a alegação extraída. Adicionalmente, realiza-se uma busca pela alegação na API Google FactCheck.

A necessidade de extração de alegação variou consideravelmente: 94,0% para COVID19.BR, 80,1% para Fake.br e 70,7% para MuMiN-PT. Uma hipótese para a menor necessidade no MuMiN-PT é pela sua característica *bottom-up*, ou seja, a coleta do *corpus* iniciou a partir de notícias verificadas por agências e, com isso, a correspondência da *fake news* original é mais alta com os resultados e não precisa da extração de alegação.Em contrapartida, a alta taxa no COVID19.BR pode refletir a natureza mais informal e fragmentada das mensagens de WhatsApp, que exigiriam sumarização (extração de alegação) para uma busca focada.

O processo foi robusto, com apenas 97 erros em mais de 10 mil exemplos processados, ou seja, a taxa de erro foi menor que 0,97%. Todos os erros ocorreram na

etapa de busca pela alegação extraída, onde a API não retornou resultados. Os dados enriquecidos resultantes são detalhados no Apêndice B.

Os resultados da análise qualitativa dos dados enriquecidos são apresentados na Seção 5.4, divididos entre os casos em que houve correspondência direta na busca inicial (Subseção 5.4.1) e aqueles que necessitaram de extração de alegação (Subseção 5.4.2).

### 5.3.1 Análise da Correspondência Direta na Busca Inicial

Quando a busca inicial na web (usando a API do CSE) retornava um trecho com alta similaridade lexical com a consulta (texto original pré-processado), a extração de alegação era dispensada. A Figura 5.11 mostra em qual posição (do 1º ao 5º resultado) ocorreu a primeira correspondência forte.

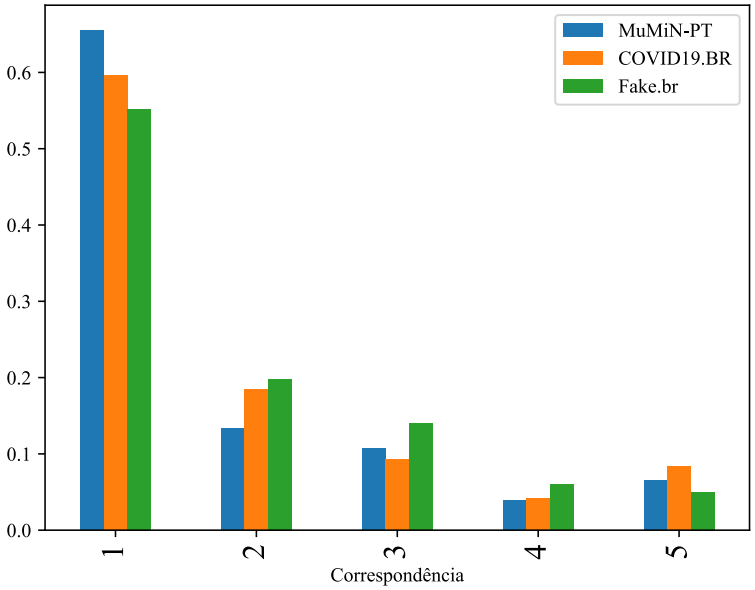


**Figura 5.11:** *Distribuição da posição do primeiro resultado de busca (Google CSE) com correspondência forte com a consulta inicial.*

Observa-se que, na vasta maioria dos casos de correspondência (cerca de 60%), ela ocorre já no primeiro resultado da busca. As taxas diminuem progressivamente para as posições seguintes. A distribuição é relativamente consistente entre os *corpora*, com o MuMiN-PT apresentando uma taxa ligeiramente maior de correspondência no primeiro resultado.

### 5.3.2 Análise da Extração de Alegações

Nos casos sem correspondência direta, procedeu-se à extração da alegação principal usando o LLM Gemini 1.5 Flash. Em média, as alegações extraídas continham

entre 11 e 12 palavras, formando uma única frase, com tamanho médio de palavra em torno de 4.9 caracteres.

A Figura 5.12 apresenta os termos mais comuns nas alegações extraídas. Comparando com os termos mais comuns nos textos originais (Figura 5.3), nota-se uma manutenção geral dos tópicos, mas com algumas substituições.

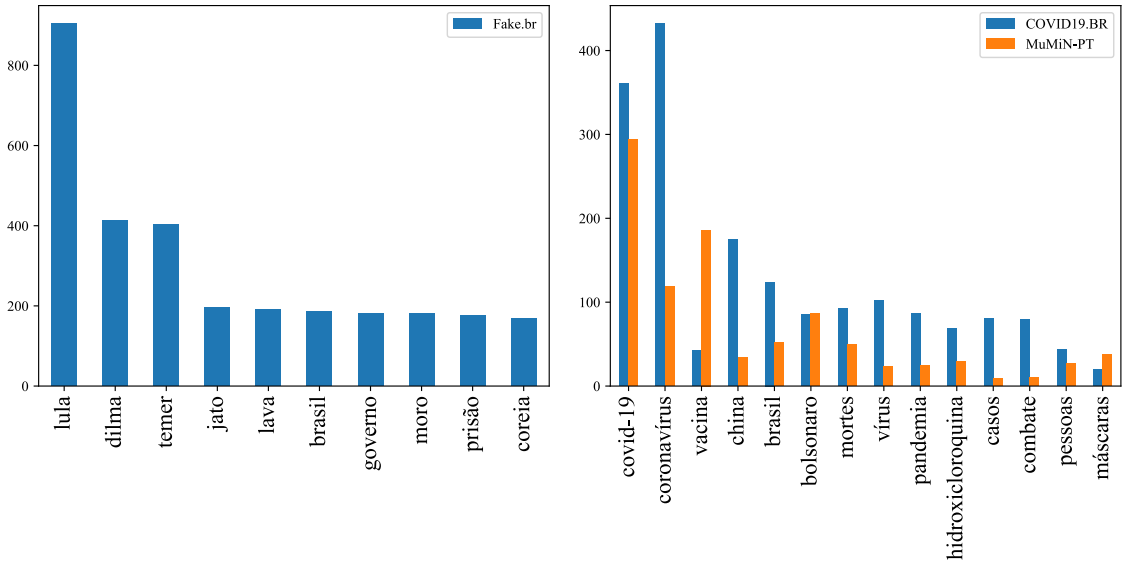


**Figura 5.12:** *10 Termos mais comuns (sem stopwords) nas alegações extraídas para Fake.br (esquerda) e COVID19.BR/MuMiN-PT (direita).*

No Fake.br, termos como "justiça" e "ex-presidente" deram lugar a "lava", "jato", "Moro" e "Dilma", refletindo um foco maior do LLM em entidades e eventos específicos (Operação Lava Jato, ex-presidenta Dilma Rousseff) ao sumarizar. No COVID19.BR/MuMiN-PT, "presidente", "saúde", "vídeo" foram parcialmente substituídos por "Bolsonaro", "hidroxicloroquina" e "casos", indicando uma sumarização focada em aspectos mais factuais ou controversos da pandemia.

Três hipóteses para mudança das palavras mais comuns das alegações podem ser a mudança temporal, o encurtamento do texto e o estilo de escrita. A passagem do tempo, por exemplo, no Fake.Br seria a palavra "Dilma" na alegação ser mais utilizada que ex-presidente no texto original, porque na coleta dos dados a Dilma era a última ex-presidente enquanto no treino do Gemini não é mais. O encurtamento do texto original para gerar a alegação pode ter contribuído para a remoção de algumas palavras mais frequentes e o estilo de escrita do Gemini pode ter contribuído para a troca de alguns termos equivalentes.

### 5.3.3 Análise dos Resultados da Busca Web (Google CSE)

Analisaram-se os domínios das URLs retornadas pela API do Google CSE, tanto na busca inicial quanto na busca pela alegação extraída. A Figura 5.13 mostra os domínios mais frequentes para cada *corpus*, os domínios governamentais brasileiros foram aglutinados em "gov.br" e os do governo americano em ".gov".

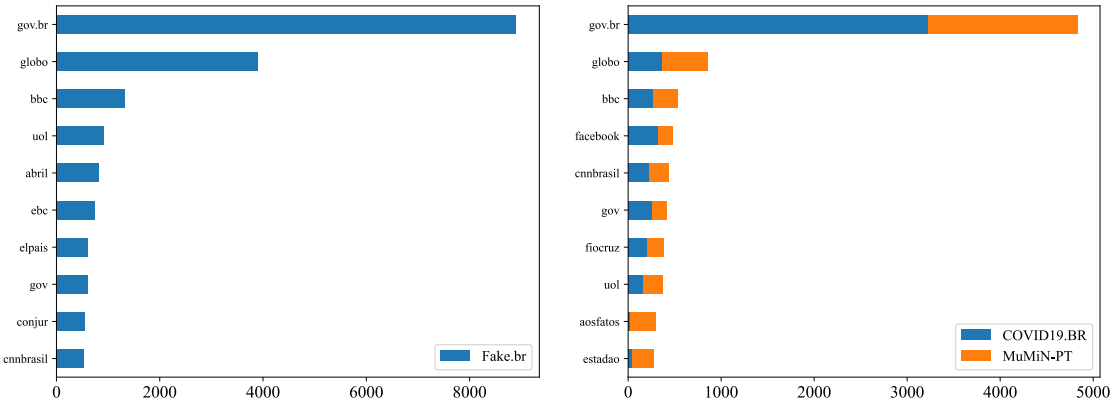


**Figura 5.13:** *10 Domínios mais frequentes nas URLs dos resultados da Google CSE para Fake.br (esquerda) e COVID19.BR/MuMiN-PT (direita).*

Os principais domínios (gov.br, globo.com, bbc.com, uol.com.br) são consistentes entre os *corpora*, indicando a proeminência de fontes governamentais e grandes veículos de mídia nos resultados de busca. Domínios como facebook.com aparecem com mais destaque para COVID19.BR e MuMiN-PT, refletindo a origem desses dados em redes sociais.

Similarmente, a presença de fiocruz.br e cnnbrasil.com.br nos resultados para os *corpora* relacionados à pandemia (COVID19.BR, MuMiN-PT) e sua ausência no Fake.br (pré-pandemia e pré-CNN Brasil) demonstra a sensibilidade da busca ao contexto temporal e temático dos dados originais.

Uma parcela dos resultados da CSE remetia a páginas indexadas pela ferramenta Google FactCheck, contendo verificações jornalísticas. Essa categoria foi especialmente relevante para o MuMiN-PT (21,3% dos retornos), reforçando a conexão deste *corpus* com alegações já verificadas, em comparação com COVID19.BR (1,0%) e Fake.br (0,6%). A Tabela 5.4 lista os domínios dessas agências identificadas indiretamente via CSE.

Os domínios mais frequentes foram AFP, UOL, Observador e Estadão. É importante notar que, embora essas páginas fossem indexadas como contendo verificações, a API da CSE por si só não fornece o rótulo de veracidade atribuído pela agência.

A análise temporal das datas de publicação dessas páginas de verificação (Figura 5.14) mostra alinhamento com os períodos de coleta dos dados originais para CO-

| Domínio da Agência (via CSE) | COVID19.BR | Fake.br | MuMiN-PT |
|---|---|---|---|
| afp.com | 8 | 10 | 154 |
| uol.com.br | 5 | 11 | 95 |
| observador.pt | 4 | 4 | 80 |
| estadao.com.br | 4 | 17 | 11 |
| e-farsas.com | 0 | 1 | 8 |
| sbt.com.br | 3 | 5 | 0 |
| globo.com | 0 | 0 | 7 |
| projetocomprova.com.br | 0 | 0 | 7 |
| sapo.pt | 0 | 1 | 0 |

**Tabela 5.4:** *Contagem de domínios de agências de checagem identificados nos resultados da Google CSE que continham marcação 'ClaimReview' (domínios principais aglutinados).*

VID19.BR (pico em 2020) e MuMiN-PT (2020-2022). Para o Fake.br, as datas dos resultados de verificação não mapeiam diretamente com a distribuição temporal dos dados originais (Figura 5.4), sugerindo que as buscas atuais recuperam verificações mais recentes ou que as verificações da época (2016-2018) têm menor visibilidade hoje. Isso reforça a sugestão de realizar buscas com restrição temporal para simular melhor a verificação em tempo real.

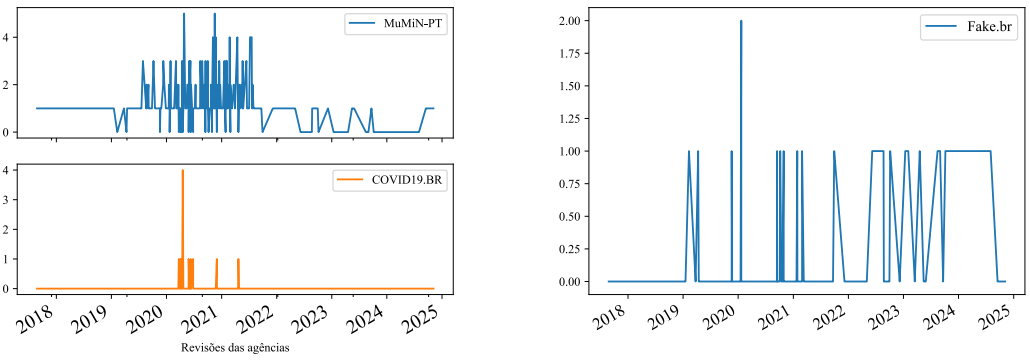


**Figura 5.14:** *Distribuição das datas de publicação das páginas de agências encontradas via Google CSE para COVID19.BR/MuMiN-PT (esquerda) e Fake.br (direita).*

Por fim, observou-se que, apesar dos parâmetros para restringir a busca ao português do Brasil (`gl=pt-BR`, `lr=lang_pt`), um pequeno número de resultados (<0.05%) ainda retornou páginas em outros idiomas (principalmente espanhol). Tentativas de restringir ainda mais pela localização do servidor (`cr=countryBR`) levaram à omissão de retornos relevantes, indicando um desafio na filtragem geográfica/linguística perfeita via API.

### 5.3.4 Análise dos Resultados da Busca de Alegações (Google Fact-Check API)

Paralelamente à busca geral (CSE), utilizou-se a API de busca de alegação do Google FactCheck, especificamente projetada para recuperar alegações verificadas. A taxa de sucesso (exemplos que retornaram pelo menos uma verificação) variou drasticamente: 58,8% para MuMiN-PT, 4,8% para COVID19.BR e apenas 0,7% para Fake.br. A Tabela 5.5 mostra os domínios das agências cujas verificações foram recuperadas diretamente por esta API.

| Domínio da Agência (via Fact Check API) | COVID19.BR | Fake.br | MuMiN-PT |
|---|---|---|---|
| aosfatos.org | 25 | 15 | 243 |
| uol.com.br | 27 | 24 | 219 |
| observador.pt | 4 | 3 | 140 |
| boatos.org | 20 | 2 | 77 |
| afp.com | 2 | 5 | 64 |
| estadao.com.br | 6 | 7 | 34 |
| projetocomprova.com.br | 3 | 0 | 42 |
| globo.com | 1 | 0 | 10 |

**Tabela 5.5:** *Contagem de domínios das agências fontes dos resultados da busca de alegações do Google FactCheck (domínios principais aglutinados).*

Comparando com a Tabela 5.4 (agências via CSE), observa-se uma mudança na proeminência: 'Aos Fatos' emerge como a principal fonte direta na Fact Check API, enquanto 'AFP' era mais visível indiretamente na CSE. 'UOL' e 'Observador' mantêm-se relevantes em ambas. Notavelmente, 'Boatos.org', importante fonte para o COVID19.BR (Tabela 4.1), aparece aqui, mas não estava entre os principais da CSE com marcação 'ClaimReview'. A distribuição temporal das verificações recuperadas (Figura 5.15) é similar à observada na Figura 5.14, alinhada aos períodos dos *corpora*.

A busca de alegações do Google FactCheck, como explicado na Figura 4.5, mostra o rótulo que a entidade jornalística associou a alegação. Dos 986 resultados, somente 11 foram catalogados como verdadeiros. Isso sustenta a hipótese de que as organizações verficam mais notícias como falsas do que verdadeiras.

Os rótulos associados às agências estão apresentados na Tabela 5.6. Os mais comuns são "falso/fake/errado" e "enganoso/enganador", destacados em negrito. A diversidade de rótulos além do binário simples (e.g., "distorcido", "sem contexto", "exagerado") também ilustra a granularidade perdida nos conjuntos de dados originais e parcialmente recuperada através desta API.

Esses resultados podem ser comparados com os da Tabela 2.1 de [30], com a ressalva de que o trabalho mencionado extraiu aproximadamente quatro vezes mais

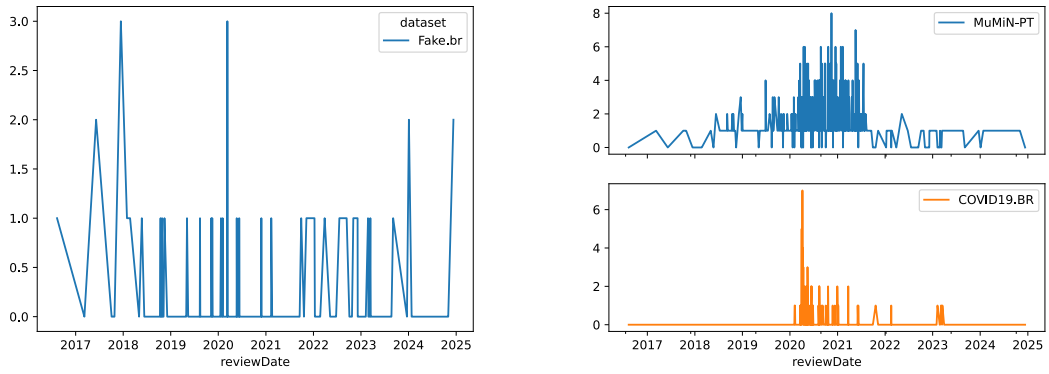


**Figura 5.15:** *Distribuição das datas de publicação das verificações encontradas via Google FactCheck API para Fake.br (esquerda) e COVID19.BR (direita).*

alegações, o que aumenta as chances de capturar categorias minoritárias. Além disso, na pesquisa de [30], a extração foi obtidas diretamente dos sites via *crawler*, enquanto neste trabalho foram extraídas da API de verificação de fatos do Google.

Embora os rótulos estejam distribuídos de forma semelhante, eles aparecem em ordens diferentes. Alguns termos não foram encontrados neste estudo, incluindo: "verdadeiro" e "impreciso" da agência Aos Fatos; "contraditório" e "subestimado" da Agência Lupa; "fato" e "não é bem assim" do G1: Fato ou Fake; e "Fato" do Estadão Verifica. Além disso, observou-se uma mudança na categorização atribuída pelo Boatos.org, que passou a ser "falso".

## 5.4 Análise Qualitativa dos Padrões nos Dados Enriquecidos

Para complementar a análise quantitativa, realizou-se uma análise qualitativa focada em identificar padrões recorrentes nos resultados do enriquecimento, especialmente em relação à natureza das evidências encontradas. Essa análise foi organizada considerando os dois cenários principais do fluxo da Figura 4.1: busca com correspondência direta e busca após extração de alegação.

### 5.4.1 Padrões em Casos de Correspondência Direta

Quando a busca inicial (CSE) encontrava alta similaridade com o texto original, dispensando a extração de alegação, observaram-se padrões distintos dependendo da veracidade do texto original (verdadeiro/falso) das alegações analisadas:

**Para alegações originalmente rotuladas como VERDADEIRAS:**

| Agências | Rótulos |
|---|---|
| Aos Fatos | **falso (246)**, distorcido (29), não é bem assim (2), insustentável (2), exagerado (2), contraditório (2) |
| UOL Notícias | **falso (130)**, **enganoso (10)**, insustentável (7), distorcido (2), sátira (1) |
| Observador | **errado (124)**, **enganador (23)** |
| Agência Lupa - UOL | **falso (100)**, verdadeiro (5), verdadeiro mas (5), exagerado (2), de olho (1), ainda é cedo para dizer (1) |
| Boatos.org | **falso (99)** |
| AFP Checamos | **falso (48)**, **enganoso (17)**, sem registro (2), verdadeiro (1), sem indícios (1), sem contexto (1), falta contexto (1) |
| Estadão Verifica | **enganoso (25)**, **falso (20)**, fora de contexto (2) |
| Projeto Comprova | **falso (24)**, **enganoso (21)** |
| G1: Fato ou Fake | **fake (11)** |
| BOL - UOL | **falso (2)** |
| Folha - UOL | **enganoso (1)**, **falso (1)** |
| Revista Piauí - UOL | **falso (2)** |

**Tabela 5.6:** *Distribuição dos rótulos de veracidade atribuídos pelas agências, conforme recuperados pela API do Google FactCheck (rótulos mais comuns destacados em negrito).*

**V1: Corroboração:** Os resultados de busca frequentemente contêm links para a fonte original da notícia ou outros veículos confiáveis que reportam a mesma informação factual. Este padrão fornece evidências que confirmam a veracidade da alegação por meio de corroboração independente, oferecendo suporte verificável para a informação apresentada.

**V2: Ausência de Confirmação Explícita:** Raramente se encontraram resultados de agências de checagem de fatos que confirmassem explicitamente uma alegação verdadeira por meio da CSE ou da API Fact Check. Esta observação alinha-se com a tendência observada na Tabela 5.6, na qual as agências focam primariamente em desmentir informações falsas ou enganosas em vez de afirmar declarações verdadeiras, sugerindo um viés sistemático no ecossistema de verificação de fatos.

**Para alegações originalmente rotuladas como FALSAS:**

**F1: Refutação Direta:** O resultado ideal para fins de verificação, incluem artigos ou checagens de fatos de fontes confiáveis (frequentemente agências de checagem) que refutam diretamente a alegação, fornecendo evidências contrárias claras e específicas que desmentem a informação falsa.

**F2: Reforço da *fake news*:** Um resultado problemático para a verificação automatizada, apresentam outras instâncias da mesma *fake news* sendo compartilhada em redes sociais, blogs ou sites não confiáveis. Este padrão inadvertidamente amplifica a

alegação falsa em vez de corrigi-la, criando um desafio significativo para sistemas que dependem apenas da correspondência textual sem avaliação da credibilidade da fonte.

**F3: Reconhecimento Acadêmico como *fake news*:** Um padrão distinto no qual os resultados de busca apontam para artigos acadêmicos, teses ou publicações científicas que discutem a *fake news* específica como um exemplo em seu contexto de pesquisa. Isso serve como meta-evidência, indicando que a alegação é reconhecida como falsa ou enganosa pela comunidade de pesquisa, proporcionando um tipo indireto mas valioso de verificação.

Para alegações verdadeiras, os resultados de busca tipicamente correspondem à própria notícia original (V1), como exemplificado pela mensagem de WhatsApp apresentada na Figura 5.16 sobre uma *startup* desenvolvendo teste de COVID-19. A ausência significativa do padrão V2 (verificação explícita de notícia verdadeira por agência) reforça a conclusão da análise quantitativa (Tabela 5.6) de que o foco primário das agências de verificação é identificar e desmentir a *fake news*, não confirmar informações verdadeiras.

**Exemplo de Alegação Verdadeira (COVID19.BR - Padrão V1)**

Projeto foi um dos seis primeiros selecionados em edital lançado pelo Pipe-Fapesp, para apoiar pesquisas sobre inovações. Pesquisadores da Biolinker – uma startup de biotecnologia (biotech) incubada no Centro de Inovação, Empreendedorismo e Tecnologia (Cietec) – estão desenvolvendo, com apoio do Programa Fapesp Pesquisa Inovativa em Pequenas Empresas (Pipe), um teste de diagnóstico da COVID-19 (doença causada pelo novo coronavírus) de baixo custo e alto desempenho, com insumos totalmente nacionais. https://dunapress.org/2020/06/11/covid-19-startup-busca-desenvolver-teste-de-diagnostico-totalmente-nacional/

**Figura 5.16:** *Exemplo de texto verdadeiro do COVID19.BR. Buscas por trechos relevantes retornam a notícia original em fontes confiáveis como a Agência FAPESP e outros veículos de comunicação que reportaram o mesmo fato. A presença de múltiplas fontes independentes reportando a mesma informação serve como forte evidência de veracidade (Padrão V1).*

Para alegações falsas (*fake news*), os três padrões identificados (F1, F2, F3) foram observados com frequências variáveis. O cenário ideal (F1), onde a busca retorna diretamente uma verificação de fonte confiável desmentindo a alegação, ocorreu em parte dos casos, mas não foi o padrão predominante. Frequentemente, o padrão F2 prevaleceu: a busca retornava a própria *fake news* sendo propagada em outras plataformas, como posts de Facebook (Figura 5.17) ou blogs não confiáveis, sem qualquer sinalização de que se tratava de conteúdo falso.


**Figura 5.17:** *Exemplo de retorno de busca (Padrão F2) onde o primeiro link aponta para a própria fake news sendo compartilhada no Facebook. Este tipo de resultado representa um desafio significativo para sistemas automatizados de verificação, pois a alta correspondência textual pode ser interpretada erroneamente como validação da alegação, quando na verdade representa apenas mais uma instância da mesma fake news sendo propagada.*

O padrão F2 é particularmente problemático para sistemas automáticos de verificação, pois uma correspondência forte com um resultado que replica a *fake news* pode ser interpretada erroneamente como validação se a credibilidade da fonte não for rigorosamente avaliada. Esta observação ressalta a importância crítica de incorporar avaliação de credibilidade de fontes em qualquer sistema automatizado de verificação de fatos, além da mera correspondência textual.

O padrão F3, onde a *fake news* é citada em publicações acadêmicas ou análises, também foi identificado com frequência significativa. Conforme detalhado na Tabela 5.7, foram encontrados 37 resultados de busca apontando para 23 publicações acadêmicas distintas (artigos, dissertações, TCCs) que utilizavam exemplos dos *corpora* analisados (majoritariamente do Fake.br e COVID19.BR) em suas análises sobre *fake news*. Este padrão representa uma forma interessante e não antecipada de verificação indireta, onde a menção da alegação em trabalhos acadêmicos que a classificam como *fake news* serve

como meta-evidência de sua falsidade.

| Referência | Área | Ano | Instituição | Tipo de Publicação | Tema Principal |
|---|---|---|---|---|---|
| [106] | C. Sociais | 2018 | USP | Periódico | Política |
| [33] | Comunicação | 2019 | PUC-RIO | Dissertação Mestrado | Política |
| [107] | Computação | 2019 | UEL | TCC | Geral |
| [56] | Comunicação | 2019 | PUC-RIO | Congresso | Política |
| [108] | Computação | 2020 | Unichristus | TCC | Geral |
| [34] | Comunicação | 2021 | PUC-RIO | Livro | Política |
| [10] | Comunicação | 2021 | UTP | Dissertação Mestrado | COVID-19 |
| [101] | Linguística | 2021 | UNL | Dissertação Mestrado | COVID-19 |
| [70] | Comunicação | 2021 | UFOP | Periódico | COVID-19 |
| [69] | Comunicação | 2021 | USP | TCC | Política |
| [9] | Política | 2021 | UFSC | Periódico | COVID-19 |
| [65] | Linguística | 2021 | UFC | Periódico | COVID-19, Política |
| [36] | Computação | 2021 | UFC | Dissertação Mestrado | COVID-19, Política |
| [87] | Biblioteconomia | 2021 | UFC | TCC | COVID-19 |
| [18] | Química | 2021 | UFPA | TCC | COVID-19 |
| [12] | Administração | 2021 | UFPE | Dissertação Mestrado | Geral |
| [32] | Política | 2022 | UFPE | Dissertação Mestrado | Política |
| [14] | Educação | 2022 | Unioeste | Periódico | COVID-19, Ciências |
| [113] | Computação | 2022 | UFC | TCC | Geral |
| [100] | Política | 2022 | UFSCAR | Dissertação Mestrado | Política |
| [97] | Linguística | 2023 | UEM | Periódico | Geral |
| [42] | Computação | 2023 | UFAM | TCC | Geral |

**Tabela 5.7:** *Publicações acadêmicas/analíticas identificadas nos resultados de busca (Padrão F3) que citam exemplos dos corpora. Esta tabela demonstra como as alegações falsas dos conjuntos de dados analisados se tornaram objetos de estudo acadêmico em diversas áreas do conhecimento, proporcionando uma forma indireta de validação da classificação dessas alegações como fake news.*

As publicações acadêmicas identificadas distribuem-se ao longo de um período de seis anos (2018-2023), com a seguinte distribuição temporal: 1 em 2018, 3 em 2019, 1 em 2020, 11 em 2021, 4 em 2022, e 2 em 2023. O aumento significativo de publicações em 2021 coincide com o período de maior intensidade da pandemia de COVID-19, refletindo o interesse acadêmico crescente no estudo da *fake news* relacionada à saúde pública durante esse período crítico.

Quanto às áreas do conhecimento, os trabalhos dividem-se em: 6 da área de Comunicação, 5 de Computação, 3 de Linguística, 3 de Política, 1 de Biblioteconomia, 1 de Administração, 1 de Ciências Sociais, 1 de Educação e 1 de Química. Esta diversidade disciplinar demonstra como o fenômeno das *fake news* atravessa fronteiras acadêmicas

tradicionais, sendo objeto de estudo tanto em campos relacionados à tecnologia quanto em ciências humanas e sociais.

Entre as publicações da área de Computação, quatro são trabalhos de conclusão de curso (TCCs) e uma é dissertação de mestrado. Os TCCs [107, 108, 42, 113] abordam principalmente os aspectos técnicos da detecção automática de *fake news*, utilizando o *corpus* Fake.br como conjunto de dados para treinamento e avaliação de modelos. Estes trabalhos exploram desde modelos tradicionais de aprendizado de máquina até arquiteturas mais avançadas como BERT e LSTM, demonstrando a relevância dos conjuntos de dados analisados para o desenvolvimento tecnológico no campo da verificação automática.

A dissertação de mestrado em Computação [36] é de autoria do criador do *corpus* COVID19.BR. Concentra-se no processo de coleta de dados denominado "farol digital de monitoramento de WhatsApp Públicos" [37], destacando a importância de metodologias específicas para a coleta e análise de *fake news* em plataformas de mensagens privadas.

Das publicações na área de Linguística, [97] analisa diretamente o Fake.br, investigando as estratégias linguísticas de manipulação utilizadas na *fake news*. O estudo identifica a evidencialidade como uma tática comum, na qual alegações falsas ganham credibilidade aparente ao atribuírem informações a figuras públicas conhecidas. Por exemplo, em "Alckmin diz que por ele PSDB 'desembarca', mas não explica se utilizará o aparelho do filme MIB", o nome de uma figura política conhecida (Geraldo Alckmin) é instrumentalizado para conferir credibilidade à informação falsa.

Dos *corpora* que são diretamente mencionados, portanto, somente o Fake.br por [107, 42, 108, 113, 97] e o COVID19.BR por [36].

Três publicações [10, 14, 101] investigam especificamente como professores podem inadvertidamente propagar *fake news* sobre COVID-19 em contextos educacionais. Barbieri [10] e Cunha [14] focam em professores do ensino fundamental brasileiro, enquanto Quintanilha [101] analisa o conhecimento de professores de química sobre a COVID-19 em Portugal, sendo este o único dos 23 trabalhos analisados proveniente de uma instituição não brasileira.

A ampla adoção dos *corpora* Fake.br e COVID19.BR em pesquisas acadêmicas diversas não apenas valida a qualidade desses conjuntos de dados, mas também demonstra como o estudo da *fake news* se consolidou como campo de pesquisa interdisciplinar. Este fenômeno cria um ciclo interessante onde a identificação de exemplos de *fake news* alimenta estudos acadêmicos que, por sua vez, podem ser detectados por sistemas automatizados de verificação como evidência indireta da classificação dessas informações como falsas.

### 5.4.2 Padrões em Casos com Extração de Alegação

Nos casos em que a correspondência direta falhou e foi necessária a extração de alegação via LLM, a análise qualitativa focou na natureza da extração e na relevância dos resultados da segunda busca (realizada com a alegação extraída).

Observou-se que, em alguns casos, a extração de alegações consistiu em operações mais simples de limpeza textual, como a remoção de saudações, despedidas ou marcações específicas do texto original (e.g., colchetes, asteriscos). Essas operações podem ser importantes, pois tais elementos poderiam interferir na eficácia da busca, especialmente se contiverem caracteres especiais interpretados pelos mecanismos de busca. A Figura 5.18 apresenta dois exemplos: à esquerda, uma limpeza de "cabeçalho" e "rodapé" em uma mensagem de WhatsApp; à direita, a remoção de marcações (colchetes) em um texto do MuMiN-PT.

**Limpeza (COVID19.BR)**

~~BOM DIA A TODOS. ASSISTAM E DEPOIS COMPARTILHEM. É MUITO IMPORTANTE QUE TODOS SAIBAM DESTA INFORMAÇÃO.~~ A CURA DO CORONAVIRUS EXISTE E ESTÁ DENTRO DO NOSSO PRÓPRIO CORPO.O PROBLEMA É QUE PARA A INDÚSTRIA FARMACÊUTICA NÃO COMPENSA DIVULGAR POR QUE NÃO TEM VALOR FINANCEIRO PARA ELA. POR ISTO ESCONDEM ESTÁ INFORMAÇÃO DE TODOS NÓS. ~~POR FAVOR ASSISTAM E DIVULGUEM. É MUITO IMPORTANTE ESTÁ INFORMAÇÃO.~~

**Alegação Extraída:** *A cura do coronavírus existe e está dentro do nosso próprio corpo.*

**Remoção de Marcação (MuMiN-PT)**

[Há] 440 mil crianças [com idade entre 4 e 5 anos] sem vaga [na educação infantil] no Brasil

**Alegação Extraída:** *Há 440 mil crianças sem vaga na educação infantil no Brasil.*

**Figura 5.18:** *Exemplos de extração de alegação envolvendo limpeza textual e remoção de marcações. À esquerda, o LLM identificou e removeu elementos paratextuais característicos de mensagens de WhatsApp (saudações, chamadas à ação), isolando a alegação central. À direita, a extração removeu marcações e reconstitui a afirmação principal em forma direta e verificável.*

A capacidade do LLM de realizar essa sumarização abstrata foi geralmente eficaz em produzir consultas mais direcionadas para a segunda busca. No entanto, mesmo com uma alegação bem extraída, a busca subsequente nem sempre retornava resultados relevantes ou conclusivos. O exemplo Figura da B.1 ilustrou um caso onde a alegação foi bem extraída, mas a busca em novembro de 2024 falhou, enquanto a busca inicial na Figura 4.4 pelo texto original em dezembro de 2024 havia encontrado correspondência.

Este fenômeno sugere a influência da temporalidade e da dinâmica dos algoritmos de busca. A relevância e a indexação de um documento podem mudar ao longo do tempo nos índices dos mecanismos de busca. Estratégias como o uso de restrições temporais na busca, quando a data original da alegação é conhecida ou pode ser estimada (como no caso do Fake.br), poderiam mitigar parcialmente esse problema, aproximando a busca do contexto temporal original da alegação.

## 5.5 Resultados da Avaliação dos Dados

Esta seção detalha os resultados obtidos a partir da avaliação dos dados, conforme a estratégia delineada na Seção 4.5 e utilizando o ambiente computacional especificado na Seção 5.2. As Tabelas 5.8 e 5.9 apresentam, respectivamente, os desfechos do ajuste fino do modelo Bertimbau e da abordagem *few-shot* com o modelo Gemini 1.5 Flash, considerando as diferentes configurações de processamento de dados estabelecidas na Seção 4.5.

| Processamento | COVID19.BR | | Fake.br | |
|---|---|---|---|---|
| | Acurácia | F1 macro | Acurácia | F1 macro |
| 1. Original | 81,9 | 82,1 | **99,6** | **99,6** |
| 2. Validado | 81,1 | 81,4 | 98,9 | 98,9 |
| 3.a) Validado e Enriquecido completo | **82,1** | **82,4** | 99,2 | 99,2 |
| 3.b) Validado e Enriquecido filtrado | 77,9 | 78,3 | 98,7 | 98,8 |

**Tabela 5.8:** *Resultados de Acurácia e F1-Macro para o ajuste fino do Bertimbau base nas diferentes configurações de processamento de dados. Os modelos foram selecionados com base no maior F1-Macro obtido no conjunto de validação e, subsequentemente, avaliados no conjunto de teste. As maiores pontuações estão em negrito e as melhores avaliações após-validação estão sublinhadas.*

Os resultados indicam que os conjuntos de dados submetidos ao processo de validação (item 2 das tabelas) apresentaram desempenho inferior em termos de Acurácia e F1-Macro quando comparados aos dados originais (item 1), tanto para o Bertimbau (Tabela 5.8) quanto para o Gemini nos datasets COVID19.BR e Fake.br (Tabela 5.9). Este fenômeno pode ser atribuído às modificações realizadas durante o processo de validação, descritas na Seção 4.2.1, podem ter tornado a tarefa de classificação mais desafiadora.

A análise comparativa entre os resultados obtidos com os dados apenas validados (item 2) e os dados validados e enriquecidos (itens 3.a e 3.b) revela que a etapa de enriquecimento de dados geralmente proporcionou ganhos de desempenho. Esta tendência

| Processamento | COVID19.BR | | Fake.br | |
|---|---|---|---|---|
| | Acurácia | F1 macro | Acurácia | F1 macro |
| 1. Original | 75,2 | 75,2 | 81,5 | 81,0 |
| 2. Validado | 76,9 | 76,9 | **81,0** | **80,4** |
| 3.a) Val. e Enriq. completo | **79,3** | **79,3** | 77,6 | 76,7 |
| 3.b) Val. e Enriq. filtrado | 79,0 | 78,9 | 78,1 | 77,2 |

**Tabela 5.9:** *Resultados de Acurácia e F1-Macro para a abordagem few-shot com o Gemini 1.5 Flash nas diferentes configurações de processamento de dados. As maiores pontuações estão em negrito.*

foi observada para o Bertimbau no COVID19.BR (Tabela 5.8) e para o Gemini 1.5 Flash no mesmo conjunto (Tabela 5.9).

Para o dataset Fake.br, o enriquecimento completo (3.a) proporcionou uma ligeira melhora no desempenho do Bertimbau, mas resultou em queda de desempenho para o Gemini. Esta degradação no Gemini pode ser atribuída à qualidade dos exemplos utilizados na entrada do few-shot, onde aproximadamente um terço (5 dos 15) dos resultados de busca se mostraram irrelevantes – um problema possivelmente relacionado à temporalidade do *corpus*. Conforme indica a Figura 5.4, os resultados de busca podem não apresentar uma concentração temporal adequada com os dados mais antigos do Fake.br. Uma possível solução seria a seleção de exemplos temporalmente mais representativos para o *few-shot* e a aplicação de restrições temporais nas buscas, considerando as datas específicas do *corpus*.

A filtragem dos dados enriquecidos para excluir aqueles provenientes de redes sociais (item 3.b em comparação com 3.a) resultou em redução de desempenho para o Bertimbau em ambos os datasets, e também para o Gemini no COVID19.BR. Entretanto, para o Gemini no Fake.br, essa filtragem específica (3.b) superou o enriquecimento completo (3.a). A redução de desempenho observada na maioria dos cenários com a filtragem (excluindo redes sociais) sugere que: (i) informações relevantes para a classificação, presentes em fontes de redes sociais, foram eliminadas no processo de filtragem; ou (ii) as páginas web não classificadas como redes sociais continham características distintivas de notícias falsas/verdadeiras que foram perdidas com essa filtragem específica, ou ainda que o conteúdo de redes sociais pode ter sido mais ruidoso para esses casos.

Os resultados demonstram consistentemente que o ajuste fino obteve resultados superiores ao *few-shot learning* em todos os cenários avaliados. Esta observação corrobora a literatura científica que indica que, havendo quantidade suficiente de exemplos de treinamento, o ajuste fino tende a superar as abordagens de *few-shot* [99, 133, 88].

## 5.6 Resumo dos Resultados

Este trabalho desenvolveu e validou uma metodologia de enriquecimento de *corpora* de detecção de notícias falsas em português utilizando contexto externo. Os resultados apresentados demonstram tanto os benefícios quanto os desafios dessa abordagem, organizados em três etapas principais conforme a metodologia proposta.

### Seleção e Validação dos Conjuntos de Dados

O levantamento inicial identificou 18 conjuntos de dados públicos de *fake news* em português. Dentre os 11 *corpora* que continham alegações textuais, foram selecionados três com características distintas que atendem aos critérios metodológicos estabelecidos: **Fake.br** [83] (7.200 notícias web de domínio geral, 12/2018), **COVID19.BR** [76] (5.635 mensagens WhatsApp sobre saúde, 10/2021) e **MuMiN-PT** [90] (3.382 *tweets* de domínio geral, 02/2022).

A análise exploratória dos dados (Seção 5.1) revelou que o Fake.br era relativamente balanceado, enquanto o MuMiN-PT apresentava mais exemplos falsos (devido à sua natureza *bottom-up*) e o COVID19.BR, mais exemplos verdadeiros. Características textuais variaram com a plataforma: textos do MuMiN-PT eram mais curtos e homogêneos em tamanho entre rótulos; no Fake.br, notícias verdadeiras eram originalmente mais longas (normalizadas para o estudo); e no COVID19.BR, textos falsos tendiam a ser mais longos (“textões” de WhatsApp). A presença de URLs foi mais significativa no COVID19.BR, com alguns domínios fortemente correlacionados com o rótulo de veracidade (Tabela 5.2), justificando sua remoção no pré-processamento.

### Aplicação do Fluxo de Enriquecimento

O fluxo de enriquecimento adaptativo demonstrou comportamento diferenciado entre os *corpora*. A necessidade de extração de alegação variou conforme esperado: 95% (COVID19.BR), 80% (Fake.br) e 71% (MuMiN-PT), com a menor taxa no MuMiN-PT devido à sua origem em notícias já verificadas por agências. As alegações extraídas pelo Gemini 1.5 Flash apresentaram consistência de tamanho (11-12 palavras em média) e qualidade de limpeza textual.

A recuperação de evidências via **Google Custom Search Engine (CSE)** retornou predominantemente fontes confiáveis (domínios governamentais e grandes mídias), com 21,3% dos resultados para MuMiN-PT contendo verificações estruturadas (`ClaimReview`). A **Google FactCheck Claims Search API** apresentou taxas de sucesso diferenciadas: 58,8% (MuMiN-PT), 4,8% (COVID19.BR) e 0,7% (Fake.br), com ’Aos Fatos’ emergindo como principal fonte de verificação (Tabela 5.5).

A análise qualitativa identificou padrões consistentes no enriquecimento. Para alegações verdadeiras, os resultados geralmente corroboravam a informação (Padrão V1), com raras confirmações explícitas por agências (V2). Para alegações falsas, emergiram três padrões: refutação direta (F1), reforço por fontes não confiáveis (F2) e reconhecimento acadêmico como *fake news* (F3). O padrão F3 revelou 23 publicações acadêmicas citando exemplos dos *corpora* estudados, principalmente com pico em 2021.

## Avaliação Experimental do Impacto

A estratégia de avaliação experimental, aplicando *fine-tuning* do Bertimbau e *few-shot learning* com Gemini 1.5 Flash nas três configurações de dados, revelou resultados nuançados. Conforme previsto na metodologia, os dados apenas validados geralmente apresentaram desempenho inferior aos originais (Tabelas 5.8 e 5.9)), confirmando o aumento da complexidade pela remoção de sinais discriminativos.

O enriquecimento com contexto externo demonstrou benefícios seletivos: melhorias consistentes para Bertimbau no COVID19.BR e para Gemini em ambos os *corpora*, com exceção do Fake.br com Gemini, onde a idade do *corpus* pode ter impactado a relevância das evidências recuperadas. A filtragem dos resultados de enriquecimento apresentou resultados mistos, sugerindo que a qualidade supera a quantidade de evidências. Consistentemente, o *fine-tuning* do Bertimbau superou o *few-shot learning* com Gemini, alinhando-se com evidências da literatura sobre a eficácia de modelos menores especializados [99, 133, 88].

Em suma, o enriquecimento adiciona contexto valioso, mas sua eficácia depende da qualidade da busca, da extração de alegações, da cobertura das APIs e da temporalidade da informação. Os resultados sugerem que uma abordagem híbrida, combinando análise textual com avaliação crítica de evidências externas (incluindo credibilidade da fonte e data), é promissora para uma detecção de *fake news* mais robusta e adaptável.

A análise detalhada do desenvolvimento e dos resultados experimentais apresentada neste capítulo fornece a base para as conclusões e discussões sobre os próximos passos da pesquisa, que serão sumarizadas no Capítulo 6.

CAPÍTULO 6

# Conclusão

Esta dissertação partiu da constatação, validada por um levantamento inicial (Hipótese **H1**), da escassez de *corpora* em língua portuguesa adequados para a tarefa de Verificação Semi-Automática de Fatos baseada em evidências externas. A maioria dos recursos existentes concentra-se na classificação baseada em características intrínsecas do texto ou, quando incluem evidências, estas frequentemente se tornaram inacessíveis devido a restrições de APIs (como a do Twitter/X).

O objetivo central deste trabalho foi desenvolver um método para enriquecer conjuntos de dados de notícias em português já existentes, agregando evidências contextuais relevantes recuperadas de fontes externas, conforme proposto no Objetivo Geral e na Hipótese **H2**. A abordagem simulou o processo cognitivo de um usuário que verifica informações online, utilizando Modelos de Linguagem Grandes (LLMs) para extrair a alegação principal dos textos e, subsequentemente, empregando mecanismos de busca (API de busca do Google e API de busca de alegação do Google FactCheck) para recuperar evidências associadas.

Os resultados demonstraram a viabilidade da metodologia proposta. Três conjuntos de dados distintos (Fake.Br, COVID19.BR, MuMiN-PT), selecionados por suas diferentes características (fonte, método de coleta, temporalidade), foram estendidos com sucesso. A análise detalhada desses conjuntos estendidos constitui o resultado principal desta pesquisa.

Confirmou-se a Hipótese **H3**, de que a extração de alegações via LLM (utilizando o Gemini 1.5 Flash) é uma etapa útil no processo, especialmente para textos mais longos ou menos diretos, como os encontrados no Fake.Br e COVID19.BR, onde a necessidade de extração foi maior (80% e 95%, respectivamente). Contudo, observou-se que, em alguns casos, principalmente para textos mais curtos ou já próximos de uma alegação verificável (mais comum no MuMiN-PT, com 71% de necessidade de extração), a busca direta poderia retornar resultados relevantes sem a etapa de extração.

A análise comparativa dos *corpora* enriquecidos apoiou as Hipóteses **H4** (referente ao impacto do veículo de publicação) e **H5** (referente à natureza da coleta dos dados), revelando que o método de coleta (*top-down* vs. *bottom-up*) e a plataforma de

publicação (notícias web, WhatsApp, Twitter) exercem influência significativa nas propriedades dos dados e, por conseguinte, no processo de enriquecimento. O MuMiN-PT (*bottom-up*, Twitter) destacou-se pela maior correspondência com verificações de fatos preexistentes e menor demanda por extração de alegações. Adicionalmente, características textuais como o tamanho dos textos e a prevalência de quase duplicatas, notadamente alta no COVID19.BR (13,6%), demonstraram variar consideravelmente entre os *corpora*, impactando a análise de consistência e a interpretação dos dados enriquecidos.

Os aprendizados sobre extração de alegações foram a base para um método de normalização, validada na competição CLEF CheckThat! 2025 [5, 6]. O sistema proposto, utilizando o ajuste fino do modelo Mono-PTT5 [98], provou-se altamente competitivo ao alcançar a terceira posição para o português (METEOR 0.5290) [118].

A avaliação experimental, relacionada à Hipótese **H6**, indicou que, embora a etapa de validação dos dados, ao remover certos sinais (e.g., URLs explícitas), pudesse tornar a tarefa de classificação mais desafiadora para os modelos, o subsequente enriquecimento com conteúdo externo geralmente propiciou melhorias de desempenho sobre os dados apenas validados. Isso foi particularmente observado para o modelo Bertimbau e para o Gemini 1.5 Flash no *corpus* COVID19.BR. Consistentemente, o ajuste fino do Bertimbau demonstrou superioridade em relação ao *few-shot learning* com Gemini, sugerindo que o enriquecimento adiciona contexto valioso, mas sua eficácia final depende da qualidade da busca, da extração de alegações, da cobertura das APIs e da temporalidade da informação. Reconhece-se, contudo, que esta abordagem possui limitações intrínsecas, como a dependência de APIs comerciais e a ausência de uma avaliação granular de cada etapa do pipeline, discutidas em detalhe na Seção 6.1.

O processo de validação semi-automático, que incluiu a detecção de quase duplicatas e checagens de consistência de rótulos, mostrou-se fundamental para melhorar a qualidade dos dados base antes do enriquecimento. A análise dos resultados de busca revelou domínios de fontes recorrentes (governamentais, grandes portais de notícia) e diferenças importantes entre a busca web geral e a API do Google FactCheck. Notavelmente, a última retornou predominantemente verificações classificadas como falsas ou enganosas, apoiando a ideia de que as agências de checagem focam no desmentido. A análise qualitativa de casos com correspondência direta na busca web destacou a importância crucial da avaliação da credibilidade da fonte e identificou cenários recorrentes, como a citação de *fake news* em trabalhos acadêmicos (Padrão F3), com 23 publicações identificadas.

As contribuições desta dissertação, detalhadas na Seção 1.3, podem ser sumarizadas como: (1) um levantamento e análise comparativa aprofundada de *corpora* de *fake news* em português, incluindo características não exploradas anteriormente como métodos de coleta e quase duplicatas; (2) um processo de validação de dados que melhora a

confiabilidade dos *corpora*; (3) o desenvolvimento e aplicação de uma metodologia de enriquecimento utilizando LLMs e APIs de busca; (4) uma análise detalhada dos dados estendidos, fornecendo *insights* sobre o processo de extração, as fontes de evidência e as características dos diferentes tipos de *fake news*; e (5) uma avaliação experimental do impacto das etapas de validação e enriquecimento no desempenho de modelos de detecção de *fake news*.

O trabalho demonstra que é possível enriquecer sistematicamente *corpora* de verificação de fatos em português com evidências contextuais, proporcionando recursos mais robustos para a pesquisa e desenvolvimento de sistemas de combate à desinformação. A metodologia proposta, embora apresente limitações relacionadas à dependência de APIs comerciais e à ausência de avaliação granular, estabelece uma base sólida para futuras pesquisas na área.

## 6.1 Limitações e Trabalhos Futuros

As limitações desta pesquisa podem ser categorizadas em três dimensões principais:

**Limitações dos dados:** A temporalidade dos conjuntos de dados representa um desafio, tanto pela desatualização de certos tópicos (e.g., referências políticas específicas no Fake.br) quanto pela dinâmica da relevância dos resultados de busca, que pode decair com o tempo. A dependência de rótulos binários (verdadeiro/falso) nos *corpora* originais simplifica a complexidade da veracidade, embora a API do Google FactCheck tenha permitido recuperar alguma granularidade (e.g., “enganoso”). Uma restrição adicional é a falta de verificabilidade associada a determinadas afirmações, como aquelas relacionadas a experiências pessoais.

**Limitações metodológicas:** As limitações dos métodos incluem a dependência de tecnologias específicas (Gemini 1.5 Flash, Google APIs), o que afeta a replicabilidade e generalização dos resultados, agravada pela ausência de uma semente determinística (*seed*) na API do Gemini utilizada e pela natureza personalizada dos resultados de busca do Google. A abordagem atual não considera também *claim splitting*, isto é, a separação de várias alegações em um texto. Além disso, como uma limitação central do método, não foi conduzida uma avaliação objetiva e granular de cada etapa do processo de enriquecimento. Tal avaliação permitiria quantificar o impacto isolado da qualidade da extração de alegações e da relevância dos resultados de busca no desempenho final, oferecendo uma compreensão mais profunda sobre quais componentes da metodologia contribuem mais significativamente para o resultado.

**Limitações de escopo:** O estudo não aborda desafios emergentes como a desinformação gerada por LLMs nem explora a aplicação prática da metodologia em cenários reais de agências de checagem de fatos.

Com base nas limitações identificadas e nos resultados obtidos, propõem-se as seguintes direções para pesquisas futuras, organizadas em três grandes áreas:

**Aprimoramento metodológico:**

- Avaliar objetivamente cada etapa do processo nos *corpora* gerados, desenvolvendo métricas para aferir a qualidade dos *prompts* de extração de alegação e a relevância das evidências recuperadas com e sem a extração, de forma a quantificar a eficácia da tarefa de enriquecimento como um todo.
- Realizar buscas contextualmente mais precisas, restringindo o período temporal para simular a verificação no momento da publicação original, sobretudo para *corpora* como o Fake.br.
- Implementar a coleta completa do conteúdo de páginas Web, superando a superficialidade dos trechos retornados por APIs de busca, a fim de assegurar uma análise contextual mais robusta e abrangente.
- Desenvolver e testar técnicas de *claim splitting* para processar textos com múltiplas alegações, comparando-as com a abordagem focada na alegação principal.

**Expansão teórica e técnica:**

- Investigar a evidencialidade linguística e seu papel na verificação de fatos, examinando o tratamento de fontes e informações atribuídas nos dados e por modelos computacionais [97].
- Explorar modelos de verificação mais avançados, como agentes baseados em LLMs [86, 134], capazes de interação iterativa com mecanismos de busca e refinamento de consultas.
- Explorar a geração de texto sintético utilizando LLMs, como uma estratégia para aumentar o volume e a diversidade dos *corpora* em português. Conforme demonstrado por [27], essa técnica permite criar conjuntos de dados amplos e balanceados para tarefas de verificação de fatos, superando a dificuldade de encontrar exemplos naturais em quantidade suficiente.
- Investigar o combate à desinformação gerada por Modelos de Linguagem Grandes LLMs, um desafio crescente que não foi objeto deste estudo. Esta linha de pesquisa futura abordaria tanto a detecção de conteúdo sintético, que pode ser mais convincente e produzido em larga escala, quanto a adaptação da metodologia de enriquecimento de evidências para enfrentar essa nova ameaça, conforme explorado por [23, 71, 136].

**Aplicação prática e escalabilidade:**

- Mitigar desafios de reprodutibilidade e custos por meio da experimentação com mecanismos de busca alternativos (e.g., DuckDuckGo, Mojeek) e da busca por maior determinismo nos LLMs.

- Medir o impacto da metodologia em aplicações práticas, desenvolvendo sistemas de apoio à decisão para agências de checagem de fatos. Tal colaboração permitiria avaliar a eficácia do enriquecimento automático na otimização do fluxo de trabalho dos verificadores humanos e na aceleração do processo de desmentido.
- Mensurar a escalabilidade da metodologia, tanto em termos de custo computacional (especialmente o uso de APIs comerciais) quanto de vazão de processamento, para viabilizar sua aplicação em cenários de grande volume de dados, como o monitoramento contínuo de redes sociais.

Essas direções futuras visam não apenas superar as limitações identificadas, mas também expandir o impacto da pesquisa para cenários práticos e desafios emergentes no combate à desinformação.

# Referências


[1] ABBASIANTAEB, Z.; ALIANNEJADI, M. **Generate then retrieve: Conversational response retrieval using llms as answer and query generators**, 2024. Disponível em https://arxiv.org/abs/2403.19302.

[2] ABDULLAH, T.; AHMET, A. **Deep learning in sentiment analysis: Recent architectures**. *ACM Comput. Surv.*, 55(8), dec 2022. Disponível em https://doi.org/10.1145/3548772.

[3] AHMAD, T.; ALIAGA LAZARTE, E. A.; MIRJALILI, S. **A systematic literature review on fake news in the covid-19 pandemic: Can ai propose a solution?** *Applied Sciences*, 12(24), 2022. Disponível em https://www.mdpi.com/2076-3417/12/24/12727.

[4] AIMEUR, E.; AMRI, S.; BRASSARD, G. **Fake news, disinformation and misinformation in social media: a review**. *Social Network Analysis and Mining*, 13(1):30, 2023.

[5] ALAM, F.; STRUSS, J. M.; CHAKRABORTY, T.; DIETZE, S.; HAFID, S.; KORRE, K.; MUTI, A.; NAKOV, P.; RUGGERI, F.; SCHELLHAMMER, S.; SETTY, V.; SUNDRIYAL, M.; TODOROV, K.; V., V. **The clef-2025 checkthat! lab: Subjectivity, fact-checking, claim normalization, and retrieval**. In: Hauff, C.; Macdonald, C.; Jannach, D.; Kazai, G.; Nardini, F. M.; Pinelli, F.; Silvestri, F.; Tonellotto, N., editors, *Advances in Information Retrieval*, p. 467–478, Cham, 2025. Springer Nature Switzerland.

[6] ALAM, F.; STRUSS, J. M.; CHAKRABORTY, T.; DIETZE, S.; HAFID, S.; KORRE, K.; MUTI, A.; NAKOV, P.; RUGGERI, F.; SCHELLHAMMER, S.; SETTY, V.; SUNDRIYAL, M.; TODOROV, K.; VENKTESH, V. **Overview of the CLEF-2025 CheckThat! Lab: Subjectivity, fact-checking, claim normalization, and retrieval**. In: Carrillo-de Albornoz, J.; Gonzalo, J.; Plaza, L.; García Seco de Herrera, A.; Mothe, J.; Piroi, F.; Rosso, P.; Spina, D.; Faggioli, G.; Ferro, N., editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Sixteenth International Conference of the CLEF Association (CLEF 2025)*, 2025.

[7] ALI, I.; AYUB, M. N. B.; SHIVAKUMARA, P.; NOOR, N. F. B. M. **Fake news detection techniques on social media: A survey**. *Wireless Communications and Mobile Computing*, 2022, 2022.

[8] ALMEIDA, L.; FUZARO, V.; NIETO, F. V.; SANTANA, A. **Identificação de "fake news" no contexto político brasileiro: uma abordagem computacional**. In: *Anais do II Workshop sobre as Implicações da Computação na Sociedade*, p. 78–89, Porto Alegre, RS, Brasil, 2021. SBC. Disponível em https://sol.sbc.org.br/index.php/wics/article/view/15966.

[9] ARNDT, G. J.; TRINDADE, M. T.; DE OLIVEIRA ALVES, J.; DE BARROS PINTO MIGUEL, R. **Quem é de direita toma cloroquina, quem é de esquerda toma... vacina**. *Revista Psicología Política*, 21(51):608–626, 2021. Disponível em https://dialnet.unirioja.es/servlet/articulo?codigo=8093425.

[10] BARBIERI, A. **Tem dúvida? não compartilhe! o uso de fake news por professores de língua portuguesa do ensino fundamental ii com o propósito de desenvolver habilidades em educação midiática com seus alunos**. Dissertação (mestrado), Universidade Tuiuti do Paraná, Curitiba, 10 2021. Orientadora: Mônica Cristine Fort. Disponível em https://tede.utp.br/jspui/handle/tede/1856.

[11] BARBOSA, V. N.; MENDES NETO, F. M.; FILHO, S. A.; SILVA, L. **A comparative study of machine learning algorithms for the detection of fake news on the internet**. In: *Proceedings of the XVIII Brazilian Symposium on Information Systems*, SBSI '22, New York, NY, USA, 2022. Association for Computing Machinery. Disponível em https://doi.org/10.1145/3535511.3535550.

[12] BATISTA, S. M. **Onde os fatos não têm vez: uma análise foucaultiana das fake news relativas à cultura**. Dissertação (mestrado em administração), Universidade Federal de Pernambuco, Recife, 2 2020. Orientador: Sérgio Carvalho Benício de Mello. Disponível em https://repositorio.ufpe.br/handle/123456789/39074.

[13] BOJANOWSKI, P.; GRAVE, E.; JOULIN, A.; MIKOLOV, T. **Enriching word vectors with subword information**. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017. Disponível em https://aclanthology.org/Q17-1010.

[14] BORIN DA CUNHA, M.; TILSCHNEIDER GARCIA ROSA, B. **Fake science: proposta de análise**. *Góndola, Enseñanza y Aprendizaje de las Ciencias*, 17(3):520–538, 11 2022. Disponível em https://revistas.udistrital.edu.co/index.php/GDLA/article/view/18098.

[15] BRODER, A. Z. **Identifying and filtering near-duplicate documents**. In: Giancarlo, R.; Sankoff, D., editors, *Combinatorial Pattern Matching*, p. 1–10, Berlin, Heidelberg, 2000. Springer Berlin Heidelberg.

[16] BROWN, T. B.; MANN, B.; RYDER, N.; SUBBIAH, M.; KAPLAN, J.; DHARIWAL, P.; NEELAKANTAN, A.; SHYAM, P.; SASTRY, G.; ASKELL, A.; AGARWAL, S.; HERBERT-VOSS, A.; KRUEGER, G.; HENIGHAN, T.; CHILD, R.; RAMESH, A.; ZIEGLER, D. M.; WU, J.; WINTER, C.; HESSE, C.; CHEN, M.; SIGLER, E.; LITWIN, M.; GRAY, S.; CHESS, B.; CLARK, J.; BERNER, C.; MCCANDLISH, S.; RADFORD, A.; SUTSKEVER, I.; AMODEI, D. **Language models are few-shot learners**, 2020. Disponível em https://arxiv.org/abs/2005.14165.

[17] CABRAL, L.; MONTEIRO, J. M.; DA SILVA, J. W. F.; MATTOS, C. L. C.; MOURAO, P. J. C. **Fakewhastapp.br: Nlp and machine learning techniques for misinformation detection in brazilian portuguese whatsapp messages.** In: *ICEIS (1)*, p. 63–74, 2021.

[18] CAPISTRANO, F. F. D. **Fake news sobre a covid-19 nas aulas de química: uma abordagem didática na monitoria das práticas de ensino**. Trabalho de conclusão de curso (licenciatura em química), Universidade Federal do Pará, Ananindeua, 8 2022. Orientadora: Janes Kened Rodrigues dos Santos. Disponível em https://bdm.ufpa.br/jspui/handle/prefix/4440.

[19] CASELI, H. D. M.; NUNES, M. D. G. V.; PAGANO, A. **O que é pln?** In: Caseli, H. M.; Nunes, M. G. V., editors, *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*, book chapter 1. BPLN, 2 edition, 2024. Disponível em https://brasileiraspln.com/livro-pln/2a-edicao/parte-introducao/cap-introducao/cap-introducao.html.

[20] CHARLES, A. C.; RUBACK, L.; OLIVEIRA, J. **Faketruebr: Um corpus brasileiro de notícias falsas**. In: Pinheiro, V.; Gamallo, P.; Amaro, R.; Scarton, C.; Batista, F.; Silva, D.; Magro, C.; Pinto, H., editors, *Anais da XVIII Escola Regional de Banco de Dados*, p. 108–117, Porto Alegre, RS, Brasil, 2023. SBC. Disponível em https://sol.sbc.org.br/index.php/erbd/article/view/24352.

[21] CHASE, H. **LangChain**. Disponível em https://github.com/langchain-ai/langchain, 8 2022. Acesso em 6 de jul. de 2024.

[22] CHEN, C.; SHU, K. **Combating misinformation in the age of llms: Opportunities and challenges**. *AI Magazine*, 45(3):354–368, 2024. Disponível em https://onlinelibrary.wiley.com/doi/abs/10.1002/aaai.12188.

[23] CHEN, C.; SHU, K. **Combating misinformation in the age of llms: Opportunities and challenges**. *AI Magazine*, 45(3):354–368, 2024.

[24] CHEN, J.; ZHANG, R.; GUO, J.; FAN, Y.; CHENG, X. **Gere: Generative evidence retrieval for fact verification**. In: *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, p. 2184–2189, New York, NY, USA, 2022. Association for Computing Machinery. Disponível em https://doi.org/10.1145/3477495.3531827.

[25] CHO, S.; JEONG, S.; YANG, W.; PARK, J. **Query generation with external knowledge for dense retrieval**. In: Agirre, E.; Apidianaki, M.; Vulić, I., editors, *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, p. 22–32, Dublin, Ireland and Online, May 2022. Association for Computational Linguistics. Disponível em https://aclanthology.org/2022.deelio-1.3.

[26] CHOI, E. C.; FERRARA, E. **Automated claim matching with large language models: Empowering fact-checkers in the fight against misinformation**. In: *Companion Proceedings of the ACM on Web Conference 2024*, WWW '24, p. 1441–1449, New York, NY, USA, 2024. Association for Computing Machinery. Disponível em https://doi.org/10.1145/3589335.3651910.

[27] CHOI, E. C.; FERRARA, E. **Fact-gpt: Fact-checking augmentation via claim matching with llms**. In: *Companion Proceedings of the ACM Web Conference 2024*, WWW '24, p. 883–886, New York, NY, USA, 2024. Association for Computing Machinery.

[28] CHOI, E.; PALOMAKI, J.; LAMM, M.; KWIATKOWSKI, T.; DAS, D.; COLLINS, M. **Decontextualization: Making sentences stand-alone**. *Transactions of the Association for Computational Linguistics*, 9:447–461, 2021. Disponível em https://aclanthology.org/2021.tacl-1.27.

[29] CORDEIRO, P. R.; PINHEIRO, V. **Um corpus de noticias falsas do twitter e verificaçao automática de rumores em lingua portuguesa**. In: *Proceedings of the Symposium in Information and Human Language Technology*, p. 219–228, 2019.

[30] COUTO, J.; PIMENTA, B.; DE ARAÚJO, I. M.; ASSIS, S.; REIS, J. C. S.; DA SILVA, A. P.; ALMEIDA, J.; BENEVENUTO, F. **Central de fatos: Um repositório de checagens de fatos**. In: *Anais do III Dataset Showcase Workshop*, p. 128–137, Porto Alegre, RS, Brasil, 2021. SBC. Disponível em https://sol.sbc.org.br/index.php/dsw/article/view/17421.

[31] DA SILVA, F. R. M.; FREIRE, P. M. S.; DE SOUZA, M. P.; DE A. B. PLENAMENTE, G.; GOLDSCHMIDT, R. R. **Fakenewssetgen: A process to build datasets that support comparison among fake news detection methods**. In: *Proceedings of the Brazilian Symposium on Multimedia and the Web*, WebMedia '20, p. 241–248, New York, NY, USA, 2020. Association for Computing Machinery. Disponível em https://doi.org/10.1145/3428658.3430965.

[32] DE FREITAS MELO, U. M. B. **Feita sob medida: a estrutura de uma notícia falsa e seu papel no convencimento do eleitor**. Master's thesis, Universidade Federal de Pernambuco, Recife, 2 2022. Dissertação (Mestrado em Ciência Política). Orientador: Sérgio Carvalho Benício de Mello. Disponível em https://repositorio.ufpe.br/handle/123456789/44709.

[33] DE MELO, M. C. **A pauta da desinformação: "fake news" e categorizações de pertencimento nas eleições presidenciais brasileiras de 2018**. Dissertação (mestrado), Pontifícia Universidade Católica do Rio de Janeiro, 4 2019. Orientadora: Adriana Andrade Braga. Disponível em https://www.dbd.puc-rio.br/pergamum/tesesabertas/1712862_2019_completo.pdf.

[34] DE MELO, M. C. **A pauta da desinformação: as ideias por trás das "fake news" nas eleições de 2018**. Fafich/Selo PPGCOM/UFMG, Belo Horizonte, 1 edition, 2021. Disponível em https://seloppgcomufmg.com.br/publicacao/a-pauta-da-desinformacao/.

[35] DE MORAIS, J. I.; ABONIZIO, H. Q.; TAVARES, G. M.; DA FONSECA, A. A.; BARBON JR, S. **A multi-label classification system to distinguish among fake, satirical, objective and legitimate news in brazilian portuguese**. *iSys - Brazilian Journal of Information Systems*, 13(4):126–149, Jul. 2020. Disponível em https://journals-sol.sbc.org.br/index.php/isys/article/view/833.

[36] DE SÁ, I. C. **Digital lighthouse: a platform for monitoring misinformation in whatsapp public groups**. Dissertação (mestrado), Universidade Federal do Ceará, 2021. Orientação: Prof. Dr. José Maria da Silva Monteiro Filho. Disponível em https://repositorio.ufc.br/handle/riufc/59268.

[37] DE SÁ, I. C.; MONTEIRO, J.; DA SILVA, J. F.; MEDEIROS, L.; MOURÃO, P.; DA CUNHA, L. C. **Digital lighthouse: A platform for monitoring public groups in whatsapp**. In: *Proceedings of the 23rd International Conference on Enterprise Information Systems - Volume 1: ICEIS,*, p. 297–304. INSTICC, SciTePress, 2021. Disponível em https://www.scitepress.org/Link.aspx?doi=10.5220/0010480102970304.

[38] DEKA, P.; JUREK-LOUGHREY, A.; P, D. **Evidence extraction to validate medical claims in fake news detection**. In: Traina, A.; Wang, H.; Zhang, Y.; Siuly, S.; Zhou, R.; Chen, L., editors, *Health Information Science*, p. 3–15, Cham, 2022. Springer Nature Switzerland.

[39] DEVLIN, J.; CHANG, M.-W.; LEE, K.; TOUTANOVA, K. **BERT: Pre-training of deep bidirectional transformers for language understanding**. In: Burstein, J.; Doran, C.; Solorio, T., editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, p. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. Disponível em https://aclanthology.org/N19-1423.

[40] DHALL, S.; DWIVEDI, A. D.; PAL, S. K.; SRIVASTAVA, G. **Blockchain-based framework for reducing fake or vicious news spread on social media/messaging platforms**. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 21(1), 11 2021. Disponível em https://doi.org/10.1145/3467019.

[41] DING, J.; HU, Y.; CHANG, H. **Bert-based mental model, a better fake news detector**. In: *Proceedings of the 2020 6th International Conference on Computing and Artificial Intelligence*, ICCAI '20, p. 396–400, New York, NY, USA, 2020. Association for Computing Machinery. Disponível em https://doi.org/10.1145/3404555.3404607.

[42] DOS SANTOS GUSMÃO, F. **Estudo comparativo de modelos de classificação textual aplicados na classificação de fake news**. Trabalho de conclusão de curso (bacharelado em engenharia da computação), Universidade Federal do Amazonas, Manaus, 6 2023. Orientador: José Luiz de Souza Pio. Disponível em https://riu.ufam.edu.br/bitstream/prefix/6934/2/TCC_FelipeGusm%C3%A3o.pdf.

[43] EL ANIGRI, S.; HIMMI, M. M.; MAHMOUDI, A. **How bert's dropout fine-tuning affects text classification?** In: Fakir, M.; Baslam, M.; El Ayachi, R., editors, *Business Intelligence*, p. 130–139, Cham, 2021. Springer International Publishing.

[44] FAUSTINI, P.; COVÕES, T. **Fake news detection using one-class classification**. In: *2019 8th Brazilian Conference on Intelligent Systems (BRACIS)*, p. 592–597, 2019.

[45] GANGI REDDY, R.; CHINTHAKINDI, S. C.; FUNG, Y. R.; SMALL, K.; JI, H. **A zero-shot claim detection framework using question answering**. In: Calzolari, N.; Huang, C.-R.; Kim, H.; Pustejovsky, J.; Wanner, L.; Choi, K.-S.; Ryu, P.-M.; Chen, H.-H.; Donatelli, L.; Ji, H.; Kurohashi, S.; Paggio, P.; Xue, N.; Kim, S.; Hahm, Y.;

He, Z.; Lee, T. K.; Santus, E.; Bond, F.; Na, S.-H., editors, *Proceedings of the 29th International Conference on Computational Linguistics*, p. 6927–6933, Gyeongju, Republic of Korea, Oct. 2022. International Committee on Computational Linguistics. Disponível em https://aclanthology.org/2022.coling-1.603.

[46] GANGI REDDY, R.; CHINTHAKINDI, S. C.; FUNG, Y. R.; SMALL, K.; JI, H. **A zero-shot claim detection framework using question answering**. In: Calzolari, N.; Huang, C.-R.; Kim, H.; Pustejovsky, J.; Wanner, L.; Choi, K.-S.; Ryu, P.-M.; Chen, H.-H.; Donatelli, L.; Ji, H.; Kurohashi, S.; Paggio, P.; Xue, N.; Kim, S.; Hahm, Y.; He, Z.; Lee, T. K.; Santus, E.; Bond, F.; Na, S.-H., editors, *Proceedings of the 29th International Conference on Computational Linguistics*, p. 6927–6933, Gyeongju, Republic of Korea, Oct. 2022. International Committee on Computational Linguistics. Disponível em https://aclanthology.org/2022.coling-1.603.

[47] GARCIA, G. L.; AFONSO, L. C. S.; PAPA, J. P. **Fakerecogna: A new brazilian corpus for fake news detection**. In: Pinheiro, V.; Gamallo, P.; Amaro, R.; Scarton, C.; Batista, F.; Silva, D.; Magro, C.; Pinto, H., editors, *Computational Processing of the Portuguese Language*, p. 57–67, Cham, 2022. Springer International Publishing.

[48] GARG, D.; KHAN, S.; ALAM, M. **Integrative use of iot and deep learning for agricultural applications**. In: Singh, P. K.; Panigrahi, B. K.; Suryadevara, N. K.; Sharma, S. K.; Singh, A. P., editors, *Proceedings of ICETIT 2019*, p. 521–531, Cham, 2020. Springer International Publishing.

[49] GOMES, J.; NETO, V.; BARBOSA, J.; DE LIMA, E. **A rapid tertiary review at the fake news domain**. In: *Anais da XI Escola Regional de Informática de Goiás*, Porto Alegre, RS, Brasil, 2023. SBC.

[50] GRIESSHABER, D.; MAUCHER, J.; VU, N. T. **Fine-tuning BERT for low-resource natural language understanding via active learning**. In: Scott, D.; Bel, N.; Zong, C., editors, *Proceedings of the 28th International Conference on Computational Linguistics*, p. 1158–1171, Barcelona, Spain (Online), dec 2020. International Committee on Computational Linguistics. Disponível em https://aclanthology.org/2020.coling-main.100/.

[51] GUO, Z.; SCHLICHTKRULL, M.; VLACHOS, A. **A survey on automated fact-checking**. *Transactions of the Association for Computational Linguistics*, 10:178–206, 2022. Disponível em https://aclanthology.org/2022.tacl-1.11.

[52] HANGLOO, S.; ARORA, B. **Combating multimodal fake news on social media: methods, datasets, and future perspective**. *Multimedia Systems*, 28(6):2391–

2422, Dec 2022. Disponível em https://doi.org/10.1007/s00530-022-00966-y.

[53] HAR-PELED, S.; INDYK, P.; MOTWANI, R. **Approximate Nearest Neighbors: Towards Removing the Curse of Dimensionality**. *Theory of Computing*, 8(1):321–350, 2012. Disponível em https://theoryofcomputing.org/articles/v008a014.

[54] HARTMANN, N.; FONSECA, E.; SHULBY, C.; TREVISO, M.; SILVA, J.; ALUÍSIO, S. **Portuguese word embeddings: Evaluating on word analogies and natural language tasks**. In: Paetzold, G. H.; Pinheiro, V., editors, *Proceedings of the 11th Brazilian Symposium in Information and Human Language Technology*, p. 122–131, Uberlândia, Brazil, Oct. 2017. Sociedade Brasileira de Computação. Disponível em https://aclanthology.org/W17-6615.

[55] HU, B.; SHENG, Q.; CAO, J.; SHI, Y.; LI, Y.; WANG, D.; QI, P. **Bad actor, good advisor: exploring the role of large language models in fake news detection**. 2024. Disponível em https://doi.org/10.1609/aaai.v38i20.30214.

[56] ITUASSU, A.; LIFSCHITZ, S.; CAPONE, L.; MANNHEIMER, V. **De donald trump a jair bolsonaro: democracia e comunicação política digital nas eleições de 2016, nos estados unidos, e 2018, no brasil**. In: *Anais do 8º Congresso da Associação Brasileira de Pesquisadores em Comunicação e Política*, Brasília, 5 2019. Universidade de Brasília, Associação Brasileira de Pesquisadores em Comunicação e Política - Compolítica. Disponível em https://compolitica.org/novo/anais/2019_gt4_Ituassu.pdf.

[57] JAKOBSON, R. **Linguistics and communications theory**. American mathematical society, 1961.

[58] KAMOI, R.; GOYAL, T.; DIEGO RODRIGUEZ, J.; DURRETT, G. **WiCE: Real-world entailment for claims in Wikipedia**. In: Bouamor, H.; Pino, J.; Bali, K., editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, p. 7561–7583, Singapore, Dec. 2023. Association for Computational Linguistics. Disponível em https://aclanthology.org/2023.emnlp-main.470.

[59] KARPUKHIN, V.; OGUZ, B.; MIN, S.; LEWIS, P.; WU, L.; EDUNOV, S.; CHEN, D.; YIH, W.-T. **Dense passage retrieval for open-domain question answering**. In: Webber, B.; Cohn, T.; He, Y.; Liu, Y., editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 6769–6781,

Online, Nov. 2020. Association for Computational Linguistics. Disponível em https://aclanthology.org/2020.emnlp-main.550.

[60] KAZEMI, A.; LI, Z.; PÉREZ-ROSAS, V.; HALE, S. A.; MIHALCEA, R. **Matching tweets with applicable fact-checks across languages**. In: *Proceedings of De-Factify: Workshop on Multimodal Fact Checking and Hate Speech Detection, CEUR*, 2022.

[61] KIM, K.; LEE, S.; HUANG, K.-H.; CHAN, H. P.; LI, M.; JI, H. **Can llms produce faithful explanations for fact-checking? towards faithful explainable fact-checking via multi-agent debate**, 2024. Disponível em https://arxiv.org/abs/2402.07401.

[62] KONDAMUDI, M. R.; SAHOO, S. R.; CHOUHAN, L.; YADAV, N. **A comprehensive survey of fake news in social networks: Attributes, features, and detection approaches**. *Journal of King Saud University - Computer and Information Sciences*, 35(6):101571, 2023. Disponível em https://www.sciencedirect.com/science/article/pii/S1319157823001258.

[63] KOTITSAS, S.; KOUNOUDIS, P.; KOUTLI, E.; PAPAGEORGIOU, H. **Leveraging fine-tuned large language models with LoRA for effective claim, claimer, and claim object detection**. In: Graham, Y.; Purver, M., editors, *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, p. 2540–2554, St. Julian's, Malta, Mar. 2024. Association for Computational Linguistics. Disponível em https://aclanthology.org/2024.eacl-long.156.

[64] KRISHNA, A.; RIEDEL, S.; VLACHOS, A. **ProoFVer: Natural logic theorem proving for fact verification**. *Transactions of the Association for Computational Linguistics*, 10:1013–1030, 2022. Disponível em https://aclanthology.org/2022.tacl-1.59.

[65] LEURQUIN, E. V. L. F.; LEURQUIN, C. **Fake news, desinformação e necessidade de formar leitores críticos**. *Scripta*, 25(54):265–295, 11 2021. Disponível em https://periodicos.pucminas.br/index.php/scripta/article/view/26681.

[66] LEWIS, P.; PEREZ, E.; PIKTUS, A.; PETRONI, F.; KARPUKHIN, V.; GOYAL, N.; KÜTTLER, H.; LEWIS, M.; YIH, W.-T.; ROCKTÄSCHEL, T.; RIEDEL, S.; KIELA, D. **Retrieval-augmented generation for knowledge-intensive nlp tasks**. In: *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc.

[67] LI, M.; PENG, B.; GALLEY, M.; GAO, J.; ZHANG, Z. **Self-checker: Plug-and-play modules for fact-checking with large language models**. In: Duh, K.; Gomez, H.; Bethard, S., editors, *Findings of the Association for Computational Linguistics: NAACL 2024*, p. 163–181, Mexico City, Mexico, June 2024. Association for Computational Linguistics. Disponível em https://aclanthology.org/2024.findings-naacl.12.

[68] LI, Y.; JIANG, B.; SHU, K.; LIU, H. **Mm-covid: A multilingual and multimodal data repository for combating covid-19 disinformation**, 2020.

[69] LIMA, A. G. **A propagação de fake news e seus impactos: um estudo sobre a onda conservadora na política ocidental contemporânea**. Trabalho de conclusão de curso (bacharelado em comunicação social com habilitação em relações públicas), Universidade de São Paulo, São Paulo, 2019. Orientador: Luiz Alberto de Farias. Disponível em https://bdta.abcd.usp.br/item/003051841.

[70] LIMA, G.; CALAZANS, M.; MASSI, L. **Mensagens falsas sobre o novo coronavírus: legitimidade e manipulação na luta de classes**. *Chasqui. Revista Latinoamericana de Comunicación*, 1(147):259–280, 08 2021. Disponível em https://revistachasqui.org/index.php/chasqui/article/view/4408.

[71] LIU, A.; SHENG, Q.; HU, X. **Preventing and detecting misinformation generated by large language models**. In: *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, p. 3001–3004, New York, NY, USA, 2024. Association for Computing Machinery.

[72] LIU, Q.; TAO, X.; WU, J.; WU, S.; WANG, L. **Can large language models detect rumors on social media?**, 2024. Disponível em https://arxiv.org/abs/2402.03916.

[73] LIU, Y.; OTT, M.; GOYAL, N.; DU, J.; JOSHI, M.; CHEN, D.; LEVY, O.; LEWIS, M.; ZETTLEMOYER, L.; STOYANOV, V. **Roberta: A robustly optimized bert pretraining approach**, 2019. Disponível em https://arxiv.org/abs/1907.11692.

[74] MAIA, D.; DA SILVA, N. F. F. **Enhancing stance detection in low-resource Brazilian Portuguese using corpus expansion generated by GPT-3.5**. In: Gamallo, P.; Claro, D.; Teixeira, A.; Real, L.; Garcia, M.; Oliveira, H. G.; Amaro, R., editors, *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1*, p. 503–508, Santiago de Compostela, Galicia/Spain, Mar. 2024. Association for Computational Lingustics. Disponível em https://aclanthology.org/2024.propor-1.51.

[75] MANNING, C. D.; RAGHAVAN, P.; SCHÜTZE, H. **Introduction to Information Retrieval**. Cambridge University Press, USA, 2008.

[76] MARTINS, A. D. F.; CABRAL, L.; MOURAO, P. J. C.; DE SÁ, I. C.; MONTEIRO, J. M.; MACHADO, J. **Covid19.br: A dataset of misinformation about covid-19 in brazilian portuguese whatsapp messages**. In: *Anais do III Dataset Showcase Workshop*, p. 138–147. SBC, 2021.

[77] MARTINS, A.; CABRAL, L.; MOURÃO, P. J.; MONTEIRO, J.; MACHADO, J. **Detection of misinformation about covid-19 in brazilian portuguese whatsapp messages using deep learning**. In: *Anais do XXXVI Simpósio Brasileiro de Bancos de Dados*, p. 85–96, Porto Alegre, RS, Brasil, 2021. SBC. Disponível em https://sol.sbc.org.br/index.php/sbbd/article/view/17868.

[78] MARTÍN, A.; HUERTAS-TATO, J.; ÁLVARO HUERTAS-GARCÍA.; VILLAR-RODRÍGUEZ, G.; CAMACHO, D. **Facter-check: Semi-automated fact-checking through semantic similarity and natural language inference**. *Knowledge-Based Systems*, 251:109265, 2022. Disponível em https://www.sciencedirect.com/science/article/pii/S0950705122006323.

[79] MEEL, P.; VISHWAKARMA, D. K. **Fake news, rumor, information pollution in social media and web: A contemporary survey of state-of-the-arts, challenges and opportunities**. *Expert Systems with Applications*, 153:112986, 2020. Disponível em https://www.sciencedirect.com/science/article/pii/S0957417419307043.

[80] META. **How-to guides: Prompting**. Disponível em https://llama.meta.com/docs/how-to-guides/prompting/#role-based-prompts. Acesso em 6 de jul. de 2024.

[81] MIKOLOV, T.; CHEN, K.; CORRADO, G.; DEAN, J. **Efficient estimation of word representations in vector space**, 2013.

[82] MIRSARRAF, M.; SHAIRI, H.; AHMADPANAH, A. **Social semiotic aspects of instagram social network**. In: *2017 IEEE International Conference on INnovations in Intelligent SysTems and Applications (INISTA)*, p. 460–465, 2017.

[83] MONTEIRO, R. A.; SANTOS, R. L. S.; PARDO, T. A. S.; DE ALMEIDA, T. A.; RUIZ, E. E. S.; VALE, O. A. **Contributions to the study of fake news in portuguese: New corpus and automatic detection results**. In: *Computational Processing of the Portuguese Language: 13th International Conference, PROPOR 2018, Canela, Brazil, September 24–26, 2018, Proceedings*, p. 324–334, Berlin, Heidelberg, 2018.

Springer-Verlag. Disponível em https://doi.org/10.1007/978-3-319-99722-3_33.

[84] MOREIRA, V. P. **Recuperação de informação**. In: Caseli, H. M.; Nunes, M. G. V., editors, *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*, book chapter 19. BPLN, 2 edition, 2024. Disponível em https://brasileiraspln.com/livro-pln/2a-edicao/parte-aplicacoes/cap-ir/cap-ir.html.

[85] MORENO, J. A.; BRESSAN, G. **Factck.br: A new dataset to study fake news**. In: *Proceedings of the 25th Brazillian Symposium on Multimedia and the Web*, WebMedia '19, p. 525–527, New York, NY, USA, 2019. Association for Computing Machinery. Disponível em https://doi.org/10.1145/3323503.3361698.

[86] NAKANO, R.; HILTON, J.; BALAJI, S.; WU, J.; OUYANG, L.; KIM, C.; HESSE, C.; JAIN, S.; KOSARAJU, V.; SAUNDERS, W.; JIANG, X.; COBBE, K.; ELOUNDOU, T.; KRUEGER, G.; BUTTON, K.; KNIGHT, M.; CHESS, B.; SCHULMAN, J. **Webgpt: Browser-assisted question-answering with human feedback**, 2022. Disponível em https://arxiv.org/abs/2112.09332.

[87] NASCIMENTO, J. G. D. **Disseminação de desinformação sobre a covid-19 em um núcleo familiar: um estudo de caso**. Trabalho de conclusão de curso (bacharelado em biblioteconomia), Universidade Federal do Ceará, Fortaleza, 2021. Orientador: Antonio Wagner Chacon Silva. Disponível em https://repositorio.ufc.br/handle/riufc/71926.

[88] NGUYEN, C. V.; SHEN, X.; APONTE, R.; XIA, Y.; BASU, S.; HU, Z.; CHEN, J.; PARMAR, M.; KUNAPULI, S.; BARROW, J.; WU, J.; SINGH, A.; WANG, Y.; GU, J.; DERNONCOURT, F.; AHMED, N. K.; LIPKA, N.; ZHANG, R.; CHEN, X.; YU, T.; KIM, S.; DEILAMSALEHY, H.; PARK, N.; RIMER, M.; ZHANG, Z.; YANG, H.; ROSSI, R. A.; NGUYEN, T. H. **A survey of small language models**, 2024. Disponível em: https://arxiv.org/abs/2410.20011.

[89] NI, J.; SHI, M.; STAMMBACH, D.; SACHAN, M.; ASH, E.; LEIPPOLD, M. **AFaCTA: Assisting the annotation of factual claim detection with reliable LLM annotators**. In: Ku, L.-W.; Martins, A.; Srikumar, V., editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, p. 1890–1912, Bangkok, Thailand, Aug. 2024. Association for Computational Linguistics. Disponível em https://aclanthology.org/2024.acl-long.104/.

[90] NIELSEN, D. S.; MCCONVILLE, R. **Mumin: A large-scale multilingual multimodal fact-checked misinformation social network dataset**. In: *Proceedings of the 45th*

*International ACM SIGIR Conference on Research and Development in Information Retrieval*, p. 3141–3153, 2022.

[91] NOGUEIRA, G. **Br fake news detection**. Disponível em https://github.com/Talendar/br_fake_news_detection, 2021. Acesso em 2 de fev. de 2025.

[92] NOGUEIRA, R. **Desvendando a arquitetura transformer: Fundamentos, aplicações e perspectivas futuras**. Disponível em https://www.youtube.com/watch?v=CP2B-OWvtF8, 12 2023. Acesso em 3 de jul. de 2024.

[93] OUSIDHOUM, N.; YUAN, Z.; VLACHOS, A. **Varifocal question generation for fact-checking**. In: Goldberg, Y.; Kozareva, Z.; Zhang, Y., editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, p. 2532–2544, Abu Dhabi, United Arab Emirates, Dec. 2022. Association for Computational Linguistics. Disponível em https://aclanthology.org/2022.emnlp-main.163.

[94] PAES, A.; VIANNA, D.; RODRIGUES, J. **Modelos de linguagem**. In: Caseli, H. M.; Nunes, M. G. V., editors, *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*, book chapter 15. BPLN, 2 edition, 2024. Disponível em https://brasileiraspln.com/livro-pln/2a-edicao/parte-modelos/cap-modelos-linguagem/cap-modelos-linguagem.html.

[95] PAGE, L.; BRIN, S.; MOTWANI, R.; WINOGRAD, T. **The pagerank citation ranking: Bringing order to the web.** Technical report, Stanford infolab, 1999.

[96] PENNINGTON, J.; SOCHER, R.; MANNING, C. **GloVe: Global vectors for word representation**. In: Moschitti, A.; Pang, B.; Daelemans, W., editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 1532–1543, Doha, Qatar, Oct. 2014. Association for Computational Linguistics. Disponível em https://aclanthology.org/D14-1162/.

[97] PEREIRA, L. P.; ANTONIO, J. D. **É verdade ou fake news? estratégicas linguísticas de manipulação em textos que promovem a desinformação**. *Revista USP*, (138):27–38, 11 2023. Disponível em https://www.revistas.usp.br/revusp/article/view/218040.

[98] PIAU, M.; LOTUFO, R.; NOGUEIRA, R. **ptt5-v2: A closer look at continued pretraining of t5 models for the portuguese language**. In: Paes, A.; Verri, F. A. N., editors, *Intelligent Systems*, p. 324–338, Cham, 2025. Springer Nature Switzerland.

[99] QIU, Y.; JIN, Y. **Chatgpt and finetuned bert: A comparative study for developing intelligent design support systems**. *Intelligent Systems with Applications*,

21:200308, 2024. Disponível em https://www.sciencedirect.com/science/article/pii/S2667305323001333.

[100] QUESSADA, M. **Desinformação e esquerda brasileira: o discurso por trás das fake news.** Dissertação (mestrado), Universidade Federal de São Carlos, São Carlos, 2 2022. Orientador: Thales Haddad Novaes de Andrade. Disponível em https://anpocs.org.br/wp-content/uploads/2023/07/42CPM.pdf.

[101] QUINTANILHA, V. D. C. **Combatendo as fake news sobre o sars-cov-2: o revisor como fact-checker**. Dissertação (mestrado), Universidade Nova, 9 2021. Orientadora: Matilde Gonçalves. Disponível em https://run.unl.pt/handle/10362/128081.

[102] RAJAPAKSE, T. C.; YATES, A.; DE RIJKE, M. **Simple transformers: Open-source for all**. In: *Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, SIGIR-AP 2024, p. 209–215, 2024. Disponível em https://doi.org/10.1145/3673791.3698412.

[103] REID, M.; SAVINOV, N.; TEPLYASHIN, D.; LEPIKHIN, D.; LILLICRAP, T.; ALAYRAC, J.-B.; SORICUT, R.; LAZARIDOU, A.; FIRAT, O.; SCHRITTWIESER, J.; OTHERS. **Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context**. *arXiv preprint arXiv:2403.05530*, 2024.

[104] REIMERS, N.; GUREVYCH, I. **Sentence-BERT: Sentence embeddings using Siamese BERT-networks**. In: Inui, K.; Jiang, J.; Ng, V.; Wan, X., editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, p. 3982–3992, Hong Kong, China, Nov. 2019. Association for Computational Linguistics. Disponível em https://aclanthology.org/D19-1410/.

[105] REIS, J. C. S.; CORREIA, A.; MURAI, F.; VELOSO, A.; BENEVENUTO, F. **Supervised learning for fake news detection**. *IEEE Intelligent Systems*, 34(2):76–81, 2019.

[106] RIBEIRO, M. M.; ORTELLADO, P. **O que são e como lidar com as notícias falsas**. *Sur - Revista Internacional de Direitos Humanos*, 15(27):71–83, 2018. Disponível em https://repositorio.usp.br/item/002993181.

[107] SAKURAI, G. Y. **Processamento de linguagem natural - detecção de fake news**. Trabalho de conclusão de curso (bacharelado em ciência da computação), Universidade Estadual de Londrina, Londrina, 2019. Orientador: Sérgio Montazzolli

Silva. Disponível em https://www.uel.br/cce/dc/wp-content/uploads/TCC_GUILHERME_SAKURAI.pdf.

[108] SANTOS, M. G. **DetecÇÃo de fake news: Um comparativo entre os modelos de aprendizado supervisionado passive agressive e multinomial naive bayes**. Trabalho de conclusão de curso (bacharelado em sistemas de informação, Centro Universitário Christus - Unichristus, Fortaleza, 8 2020. Orientador: Daniel Nascimento Teixeira. Disponível em https://repositorio.unichristus.edu.br/jspui/handle/123456789/1745.

[109] SCHLICHT, I. B.; FERNANDEZ, E.; CHULVI, B.; ROSSO, P. **Automatic detection of health misinformation: a systematic review**. *Journal of Ambient Intelligence and Humanized Computing*, p. 1–13, 2023.

[110] SCIRÈ, A.; GHONIM, K.; NAVIGLI, R. **FENICE: Factuality evaluation of summarization based on natural language inference and claim extraction**. In: Ku, L.-W.; Martins, A.; Srikumar, V., editors, *Findings of the Association for Computational Linguistics: ACL 2024*, Bangkok, Thailand, Aug. 2024. Association for Computational Linguistics. Disponível em https://aclanthology.org/2024.findings-acl.841/.

[111] SHAHI, G. K.; NANDINI, D. **FakeCovid- A Multilingual Cross-domain Fact Check News Dataset for COVID-19**. ICWSM, Jun 2020.

[112] SHARMA, K.; QIAN, F.; JIANG, H.; RUCHANSKY, N.; ZHANG, M.; LIU, Y. **Combating fake news: A survey on identification and mitigation techniques**. *ACM Trans. Intell. Syst. Technol.*, 10(3), 4 2019. Disponível em https://doi.org/10.1145/3305260.

[113] SOUSA, F. J. D. S. **Transferência de conhecimento para detecção automática de fake news com aprendizagem profunda**. Trabalho de conclusão de curso (bacharelado em ciência da computação), Universidade Federal do Ceará, Cratéus, 2022. Orientador: Livio Antonio de Melo Freire. Disponível em https://repositorio.ufc.br/handle/riufc/70014.

[114] SOUZA, F.; NOGUEIRA, R.; LOTUFO, R. **Bertimbau: Pretrained bert models for brazilian portuguese**. In: Cerri, R.; Prati, R. C., editors, *Intelligent Systems*, p. 403–417, Cham, 2020. Springer International Publishing.

[115] STAHL, P. M. **Lingua-py**. Disponível em https://github.com/pemistahl/lingua-py, 2023. Acesso em 3 de abril de 2024.

[116] STATISTA. **Most popular social networks worldwide as of january 2024, ranked by number of monthly active users**, 2024. Disponível em https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/.

[117] SUNDRIYAL, M.; CHAKRABORTY, T.; NAKOV, P. **From chaos to clarity: Claim normalization to empower fact-checking**. In: *Findings of the Association for Computational Linguistics: EMNLP 2023*, p. 6594–6609, Singapore, 2023. Association for Computational Linguistics.

[118] SUNDRIYAL, M.; CHAKRABORTY, T.; NAKOV, P. **Overview of the CLEF-2025 CheckThat! lab task 2 on claim normalization**. In: Faggioli, G.; Ferro, N.; Rosso, P.; Spina, D., editors, *Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum*, CLEF 2025, Madrid, Spain, 2025.

[119] TAN, X.; ZOU, B.; AW, A. T. **Evidence-based interpretable open-domain fact-checking with large language models**, 2023. Disponível em https://arxiv.org/abs/2312.05834.

[120] TAN, X.; ZOU, B.; AW, A. T. **Improving explainable fact-checking with claim-evidence correlations**. In: Rambow, O.; Wanner, L.; Apidianaki, M.; Al-Khalifa, H.; Eugenio, B. D.; Schockaert, S., editors, *Proceedings of the 31st International Conference on Computational Linguistics*, p. 1600–1612, Abu Dhabi, UAE, Jan. 2025. Association for Computational Linguistics.

[121] TANG, L.; LABAN, P.; DURRETT, G. **MiniCheck: Efficient fact-checking of LLMs on grounding documents**. In: Al-Onaizan, Y.; Bansal, M.; Chen, Y.-N., editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, p. 8818–8847, Miami, Florida, USA, Nov. 2024. Association for Computational Linguistics. Disponível em https://aclanthology.org/2024.emnlp-main.499/.

[122] TAUSCZIK, Y. R.; PENNEBAKER, J. W. **The psychological meaning of words: Liwc and computerized text analysis methods**. *Journal of Language and Social Psychology*, 29(1):24–54, 2010. Disponível em https://doi.org/10.1177/0261927X09351676.

[123] TEAM, G.; ANIL, R.; BORGEAUD, S.; ALAYRAC, J.-B.; YU, J.; SORICUT, R.; SCHALKWYK, J.; DAI, A. M.; HAUTH, A.; MILLICAN, K.; SILVER, D.; JOHNSON, M.; ANTONOGLOU, I.; SCHRITTWIESER, J.; GLAESE, A.; CHEN, J.; PITLER, E.; LILLICRAP, T.; LAZARIDOU, A.; FIRAT, O.; MOLLOY, J.; ISARD, M.; BARHAM, P. R.; HENNIGAN, T.; LEE, B.; VIOLA, F.; REYNOLDS, M.; XU, Y.; DOHERTY, R.; ETAL,

E. C. **Gemini: A family of highly capable multimodal models**, 2024. Disponível em https://arxiv.org/abs/2312.11805.

[124] VARGAS, F.; JAIDKA, K.; PARDO, T.; BENEVENUTO, F. **Predicting sentence-level factuality of news and bias of media outlets**. In: Mitkov, R.; Angelova, G., editors, *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, p. 1197–1206, Varna, Bulgaria, Sept. 2023. INCOMA Ltd., Shoumen, Bulgaria. Disponível em https://aclanthology.org/2023.ranlp-1.127.

[125] VARMA, R.; VERMA, Y.; VIJAYVARGIYA, P.; CHURI, P. P. **A systematic survey on deep learning and machine learning approaches of fake news detection in the pre-and post-covid-19 pandemic**. *International Journal of Intelligent Computing and Cybernetics*, 14(4):617–646, 2021.

[126] VASWANI, A.; SHAZEER, N.; PARMAR, N.; USZKOREIT, J.; JONES, L.; GOMEZ, A. N.; KAISER, L. U.; POLOSUKHIN, I. **Attention is all you need**. In: Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. Disponível em https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[127] VILLAR-RODRÍGUEZ, G.; ÁLVARO HUERTAS-GARCÍA.; MARTÍN, A.; HUERTAS-TATO, J.; CAMACHO, D. **Distrack: a new tool for semi-automatic misinformation tracking in online social networks**, 2024. Disponível em https://arxiv.org/abs/2408.00633.

[128] WAGNER FILHO, J. A.; WILKENS, R.; IDIART, M.; VILLAVICENCIO, A. **The brWaC corpus: A new open resource for Brazilian Portuguese**. In: Calzolari, N.; Choukri, K.; Cieri, C.; Declerck, T.; Goggi, S.; Hasida, K.; Isahara, H.; Maegaard, B.; Mariani, J.; Mazo, H.; Moreno, A.; Odijk, J.; Piperidis, S.; Tokunaga, T., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA). Disponível em https://aclanthology.org/L18-1686/.

[129] WANG, A.; SINGH, A.; MICHAEL, J.; HILL, F.; LEVY, O.; BOWMAN, S. **GLUE: A multi-task benchmark and analysis platform for natural language understanding**. In: Linzen, T.; Chrupała, G.; Alishahi, A., editors, *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for*

*NLP*, p. 353–355, Brussels, Belgium, Nov. 2018. Association for Computational Linguistics. Disponível em https://aclanthology.org/W18-5446.

[130] WANG, H.; SHU, K. **Explainable claim verification via knowledge-grounded reasoning with large language models**. In: Bouamor, H.; Pino, J.; Bali, K., editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, p. 6288–6304, Singapore, Dec. 2023. Association for Computational Linguistics. Disponível em https://aclanthology.org/2023.findings-emnlp.416.

[131] WANG, Y.; GANGI REDDY, R.; MUJAHID, Z. M.; ARORA, A.; RUBASHEVSKII, A.; GENG, J.; MOHAMMED AFZAL, O.; PAN, L.; BORENSTEIN, N.; PILLAI, A.; AUGENSTEIN, I.; GUREVYCH, I.; NAKOV, P. **Factcheck-bench: Fine-grained evaluation benchmark for automatic fact-checkers**. In: Al-Onaizan, Y.; Bansal, M.; Chen, Y.-N., editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, p. 14199–14230, Miami, Florida, USA, Nov. 2024. Association for Computational Linguistics. Disponível em https://aclanthology.org/2024.findings-emnlp.830/.

[132] WU, Y.; NGAI, E. W.; WU, P.; WU, C. **Fake news on the internet: a literature review, synthesis and directions for future research**. *Internet Research*, 32:1662–1699, 1 2022. Disponível em https://doi.org/10.1108/INTR-05-2021-0294.

[133] YANG, J.; JIN, H.; TANG, R.; HAN, X.; FENG, Q.; JIANG, H.; ZHONG, S.; YIN, B.; HU, X. **Harnessing the power of llms in practice: A survey on chatgpt and beyond**. *ACM Trans. Knowl. Discov. Data*, 18(6), apr 2024. Disponível em https://doi.org/10.1145/3649506.

[134] YAO, S.; ZHAO, J.; YU, D.; DU, N.; SHAFRAN, I.; NARASIMHAN, K.; CAO, Y. **React: Synergizing reasoning and acting in language models**. *arXiv preprint arXiv:2210.03629*, 2022.

[135] ZENG, F.; GAO, W. **JustiLM: Few-shot Justification Generation for Explainable Fact-Checking of Real-world Claims**. *Transactions of the Association for Computational Linguistics*, 12:334–354, 04 2024. Disponível em https://doi.org/10.1162/tacl_a_00649.

[136] ZHANG, Y.; SHARMA, K.; DU, L.; LIU, Y. **Toward mitigating misinformation and social media manipulation in llm era**. WWW '24, p. 1302–1305, New York, NY, USA, 2024. Association for Computing Machinery.

[137] ZHAO, W. X.; ZHOU, K.; LI, J.; TANG, T.; WANG, X.; HOU, Y.; MIN, Y.; ZHANG, B.; ZHANG, J.; DONG, Z.; DU, Y.; YANG, C.; CHEN, Y.; CHEN, Z.; JIANG, J.;

REN, R.; LI, Y.; TANG, X.; LIU, Z.; LIU, P.; NIE, J.-Y.; WEN, J.-R. **A survey of large language models**. *arXiv preprint arXiv:2303.18223*, 2023. Disponível em http://arxiv.org/abs/2303.18223.

APÊNDICE A

# Exemplos Ilustrativos do Fluxo de Validação Semiautomático

Este apêndice fornece exemplos concretos para ilustrar as diferentes etapas do fluxo de validação semiautomático descrito na Seção 4.2.1. O objetivo é clarificar os critérios aplicados durante a filtragem e correção dos dados. É importante notar que a remoção de URLs dos textos, mencionada como uma etapa geral para evitar viés de domínio, foi realizada após as etapas de validação aqui exemplificadas que pudessem depender da presença dessas URLs (como a resolução de contradições baseada em URL ou a filtragem específica do Fake.br por URL de origem). Os exemplos abaixo mostram os textos antes dessa remoção final de URLs, mas com outros processamentos de validação específica sendo ilustrados.

## A.1 Filtragem Inicial Automatizada

Esta etapa removeu exemplos que não atendiam a critérios básicos de conteúdo ou formato. A Figura A.1 ilustra três tipos de remoção nesta fase: textos compostos unicamente por uma URL, textos considerados excessivamente curtos após tokenização e remoção de *stopwords*, e duplicatas exatas dentro do mesmo *corpus*.

## A.2 Filtragem de Idioma

Exemplos identificados como não estando primariamente em português foram removidos. A ferramenta Lingua foi usada para detecção automática, seguida de verificação manual em casos ambíguos. A Figura A.2 mostra exemplos removidos por estarem em inglês e espanhol.

**Texto composto apenas por URL (COVID19.BR) - Removido**

https://fanoticias.com.br/covid-19-jovem-sobre-os-efeitos-parecia-furar-meu-peito/'

**Texto com poucos tokens relevantes (COVID19.BR) - Removido**

Kkkkk se parece com o corona.kkkk

**Duplicata Exata (Fake.br) - Exemplo `true_0069` removido**

Suplicy participará de programa de Doria na web nesta quinta
O vereador Eduardo Suplicy (PT-SP) participará nesta quinta-feira (10), às 20h30, do programa "Olho no Olho", transmitido pelas redes sociais do prefeito João Doria (PSDB-SP).
Doria já recebeu no quadro aliados e personalidades como o cantor Lobão, o apresentador José Luiz Datena, o ex-jogador de basquete Oscar, o cantor Roger, do Ultraje a Rigor, e a jornalista Joice Hasselman.
No começo de sua gestão, o tucano poupou de críticas seu antecessor Fernando Haddad (PT-SP), mas, nos últimos meses, passou a acusar o ex-prefeito de deixar um rombo de R$ 7 bilhões na prefeitura. Haddad nega e diz que deixou as contas da cidade em ordem.
Suplicy, por sua vez, é uma das vozes críticas à Doria na Câmara dos Vereadores.

**Figura A.1:** *Exemplos de remoções realizadas na etapa de filtragem inicial automatizada.*

## A.3 Resolução de contradições

Esta etapa abordou inconsistências de rótulos entre exemplos semanticamente muito similares ou que referenciavam a mesma fonte. A Figura A.3 ilustra um caso de textos quase idênticos no *corpus* COVID19.BR que possuíam rótulos de veracidade conflitantes (verdadeiro vs. falso). Após análise manual, baseada no contexto e fontes externas quando disponíveis, um dos rótulos foi corrigido ou, em casos irresolúveis, ambos os exemplos foram removidos para garantir a consistência.

## A.4 Verificação Externa de Rótulos

Utilizou-se a API do Google FactCheck para identificar potenciais erros de rotulação nos dados originais. Quando a verificação externa sugeria um rótulo diferente do original, o exemplo era revisado manualmente. A Figura A.4 demonstra um caso no *corpus* MuMiN-PT onde o rótulo original ('true') foi corrigido para 'fake' após a verificação externa e análise manual confirmarem a incorreção inicial.

**Exemplo em Inglês (COVID19.BR) - Removido**

German state minister kills himself as coronavirus hits economy
- https://www.aljazeera.com/news/2020/03/german-state-minister-kills-coronavirus-hits-economy-200329165242615.html

**Exemplo em Espanhol (MuMiN) - Removido**

Vídeo de fraude en las urnas de Flint (Michigan) durante las elecciones de Estados Unidos

**Figura A.2:** *Exemplos filtrados por não estarem em português.*

**Texto Quase-Duplicado A (COVID19.BR) - Rótulo Original: Falso**

Em meio à grave crise do CORONAVÍRUS, o Congresso Nacional se recusou a abrir mão do dinheiro destinado ao fundo eleitoral para auxílio ao combate da pandemia. Como sempre esses sangue-sugas demonstram se importar apenas com os próprios interesses. Vamos assinar a petição pelo fechamento do Congresso Nacional!
👉 https://peticaopublica.org/fechamento-congresso-nacional/

**Trecho dado como verdadeiro (COVID19.BR)**

Em meio à grave crise do CORONAVÍRUS, o Congresso Nacional se recusou a abrir mão do dinheiro destinado ao fundo eleitoral para auxílio ao combate da pandemia. Como sempre esses sangue-sugas demonstram se importar apenas com os próprios interesses. Vamos assinar a petição pelo fechamento do Congresso Nacional!
[redacted]https://peticaopublica.org/fechamento-congresso-nacional/

**Figura A.3:** *Exemplo de par de textos quase duplicados no corpus COVID19.BR que apresentavam rótulos de veracidade contraditórios. A resolução manual foi necessária para determinar o rótulo correto ou remover o par.*

## A.5 Tratamento Específico ao Fake.br

O *corpus* Fake.br possui características particulares que demandaram etapas de tratamento adicionais. A Figura A.5 ilustra a remoção de textos quase duplicados que compartilhavam a mesma URL de origem, uma situação específica deste *corpus*. Além disso, outras duas regras específicas foram aplicadas (não ilustradas com exemplos visuais detalhados por brevidade, mas descritas abaixo):

(a) **Remoção de textos incompletos:** Exemplos cujos textos normalizados (conforme 5.1.1) foram identificados como truncados ou incompletos em relação à notícia original (abordado na Issue #7 do repositório do Fake.br) foram removidos.

(b) **Manutenção de pares:** Dado que o Fake.br é estruturado em pares de notícias (verdadeira e falsa sobre o mesmo evento), se um dos elementos do par fosse removido por qualquer um dos critérios de validação anteriores, o elemento correspondente também era removido para preservar a integridade da estrutura pareada do conjunto

**Exemplo de Correção de Rótulo (MuMiN-PT)**

**Corpus:** MuMiN-PT

**Texto original:** A Associação Médica Americana retirou restrições e passou a recomendar a hidroxicloroquina contra a covid-19

**Rótulo:** ~~`true`~~ `fake`

**Consulta (Alegação extraída):** A Associação Médica Americana recomenda a hidroxicloroquina contra a covid-19.

**Resultado do CSE:**

> **Título:** `<b>Hidroxicloroquina</b> não é recomendada como tratamento precoce ...`
>
> **Trecho:** `1 day ago <b>...</b> De tempos em tempos, o medicamento <b>hidroxicloroquina</b> volta a ser apontado como tratamento precoce eficaz <b>contra a Covid</b>-<b>19</b>.`

**Resultado do FactCheck:**

> **Agência:** Aos Fatos | **Rótulo:** Falso | **Alegação revisada:** Associação Médica Americana não recomenda hidroxicloroquina ...'

**Figura A.4:** *Ilustração do processo de verificação externa de rótulos. Um exemplo do MuMiN-PT teve seu rótulo original (`true`) corrigido para `fake` com base em evidências da API do Google FactCheck e subsequente confirmação manual.*

de dados.

**Quase-Duplicatas com Mesma URL de Origem (Fake.br) - Exemplo `true_3023` removido**

O paraíso dos infratores Para conseguir um terço dos votos dos deputados, Temer decreta perdão de multas Os projetos de privatização para melhorar o desempenho das contas públicas estão sendo golpeados no processo de conquista de um terço dos votos da Câmara para livrar o presidente Temer de responder a investigação por crimes de organização criminosa e obstrução da Justiça. Mais agora, depois que foi atendido o condenado Waldemar da Costa Neto, vulgo Boy, que exigiu a retirada do aeroporto de Congonhas do pacote anunciado, que traria 6 bilhões de renda, em troca de manter o segundo maior aeroporto do País sob controle da Infraero. Está na cara que o esquema de corrupção dos governos anteriores é mantido. . . .

**Figura A.5:** *Exemplo de remoção específica do Fake.br: textos quase duplicados (`true_0251` e `true_3023`) originados da mesma URL fonte. Um deles (`true_3023`) foi removido para reduzir redundância originada da coleta.*

APÊNDICE B

# Exemplos pós-expansão dos dados

Neste apêndice, apresentam-se exemplos ilustrativos dos conjuntos de dados após o processo de enriquecimento descrito na Seção 5.3. Cada exemplo inclui metadados do conjunto de dados original (*Corpus*, rótulo, texto original com a parte da consulta sublinhada), a consulta utilizada para a busca (seja o texto sublinhado ou uma alegação extraída), o trecho principal do resultado do CSE e do Google FactCheck obtidos.

## B.1 Correspondência: sem extração de alegação

O primeiro cenário ilustra casos em que a busca inicial pelo texto original encontrou um resultado com alta correspondência (maior que 0,8), dispensando a extração de alegação. Os casos são organizados conforme os padrões identificados na análise qualitativa (Seção 5.4).

## B.2 Com extração de alegação

Este cenário abrange os casos em que a busca inicial não encontrou resultados com correspondência suficiente (‘match < 0.8‘ para todos os resultados). Nesses casos, uma alegação (‘claim‘) foi extraída via LLM, e uma segunda busca foi realizada com essa alegação. Apresentam-se exemplos onde a segunda busca retornou resultados considerados relevantes (Subseção B.2.1) e onde os resultados não foram considerados diretamente relevantes para a verificação (Subseção B.2.2).

### B.2.1 Resultados relevantes após extração

O exemplo a seguir ilustra um caso onde a busca inicial falhou, a alegação foi extraída com sucesso, e a busca com a alegação extraída retornou resultados potencialmente úteis para verificação. Este padrão demonstra o valor da extração de alegações como etapa intermediária no processo de enriquecimento.

Exemplo com correspondência direta (Padrão F1)

***Corpus*:** MuMiN-PT

**Texto original:** O novo coronavírus pode ser transmitido através de encomendas enviadas da China

**Rótulo** `fake`

**Consulta (Texto original)** O novo coronavírus pode ser transmitido através de encomendas enviadas da China

**Resultado do CSE:**

> **Título:** `<b>O novo coronavírus</b> não sobrevive em <b>encomendas enviadas</b> pelo ...`
>
> **Trecho:** `4 de mar. de 2020 <b>...</b> ... <b>coronavírus pode ser transmitido através de encomendas enviadas</b> pelo correio da <b>China</b> para outros países. A alegação, que serviu de base  ...`

**Resultado do FactCheck:** `None`

**Figura B.1:** *Exemplo de busca direta satisfatória (Cenário 1). A consulta encontra um resultado que refuta diretamente a alegação, sem necessidade de extração de alegação. Ilustra o Padrão F1 (Refutação Direta) onde o próprio título do resultado já indica a refutação ("não sobrevive"). O resultado origina-se do MuMiN-PT e demonstra a eficácia do sistema em localizar verificações relevantes quando a consulta contém termos precisos.*

Exemplo com extração e resultados relevantes

***Corpus*** COVID19.BR

**Texto original** Gente ja tivemos pacientes que fizeram exame de COVID-19 , inclusive com exame feito no sabin e tendo dado negativo e repetiu aqui no Hospital Oswaldo Cruz no teste rápido e deu positivo, lembre-se o teste de swab nasal tem chance de dar negativo. Fiquem atentos, pra não estar com um falso negativo e contaminando outros. Olho se tiver sintomas não descuide pois alguns exames são falsos negativos.

**Rótulo** `true`

**Consulta (Alegação extraída)** Exames de COVID-19 podem apresentar falsos negativos, mesmo com testes em laboratórios renomados.

**Resultado do CSE:**

> **Match:** 0.875
>
> **Título:** `<b>Covid</b>-<b>19</b>: Pesquisador da Fiocruz tira dúvidas sobre <b>testes</b> de ...`
>
> **Trecho:** `Jan 15, 2021 <b>...</b> ... <b>falsos negativos podem</b> se dever à baixa especificidade e baixa sensibilidade analítica do <b>teste</b>.`

**Resultado do FactCheck** `None`

**Tabela B.1:** *Exemplo com extração de alegação e resultados relevantes na segunda busca (Subcenário 2.1). A busca inicial falhou em encontrar correspondência direta com o texto original (que continha detalhes específicos de relato pessoal), mas o LLM conseguiu extrair a alegação central de forma concisa e precisa. A busca com a alegação extraída encontrou informações cor-*

Reforço da *fake news* no padrão F2

**Corpus:** Fake.br

**Texto original:** Governo Temer irá acabar com a farmácia popular. Quem nunca ouviu ou até mesmo precisou da farmácia popular, não é mesmo? O sistema oferece a distribuição de remédios gratuitos para diversos tipos de doenças, esses recursos estão disponíveis até o mês de agosto de 2017 (truncado)

**Rótulo:** `fake`

**Consulta (Texto original):** Governo Temer irá acabar com a farmácia popular.

**Resultado do CSE:**

**Título:** `<b>Governo Temer</b> fecha <b>Farmácia Popular</b> e pretende extinguir o SUS ...`

**Trecho:** `Feb 22, 2018 <b>...</b> Você <b>vai</b> deixar? Notícias &middot; ascom ... O Ministério da Saúde <b>acaba</b> de fechar as 517 <b>farmácias</b> populares mantidas pelo <b>governo</b> federal no país.`

**Resultado do FactCheck:**

**Agência:** Lupa - UOL | **Rótulo:** Falso | **Alegação revisada:** Temer não 'oficializa fim do projeto Farmácia Popular'

**Figura B.2:** *Exemplo ilustrando o Padrão F2 (Reforço da fake news). A busca do CSE retornou um resultado que reforça a alegação falsa ("Governo Temer fecha Farmácia Popular"), amplificando a desinformação em vez de corrigi-la. No entanto, o Google FactCheck conseguiu localizar uma verificação relevante que refuta a alegação. Este caso demonstra a importância da integração de múltiplas fontes de verificação, pois o resultado do CSE isoladamente poderia levar a conclusões equivocadas.*

### B.2.2 Resultados não relevantes após extração

O exemplo a seguir demonstra um caso onde, mesmo após a extração bem-sucedida de uma alegação, a busca não retornou resultados diretamente relevantes para a verificação da alegação específica. Este cenário representa um desafio persistente para sistemas automatizados de verificação de fatos.

Exemplo com extração e sem resultados relevantes.

***Corpus*** MuMiN-PT

**Texto Original** Pessoal, todo mundo precisa se cadastrar no conectesus para vacinar. Sugiro fazer já. Provavelmente o site nao aguentará os acessos quando for o momento. https://conectesus-paciente.saude.gov.br/ É um cadastro no SUS. Quem tomou a vacina da Febre Amarela em 2018 já tem. Ou quem usou o SUS nos últimos anos. O aplicativo funciona mais ou menos como esses apps de carteira de motorista ou título de eleitor

**Rótulo** `fake`

**Consulta (Alegação extraída)** Pessoal, todo mundo precisa se cadastrar no conectesus para vacinar.

**Resultado do CSE:**

> **Título:** `Obter o Certificado Nacional de <b>Vacinação</b> COVID-19`
> **Trecho:** `Os dados já foram enviados, mas possui algum erro de informação. Quais os possíveis erros: CNS duplicado no <b>cadastro</b> do SUS.O cidadão deverá se dirigir a ...`

**Resultado do FactCheck** `None`

**Tabela B.2:** *Exemplo de busca que retorna resultados não diretamente relevantes para a alegação específica (Subcenário 2.2). Neste caso, o LLM identificou corretamente a alegação central (obrigatoriedade do cadastro para vacinação), mas a busca retornou apenas informações sobre o certificado de vacinação, sem abordar a questão da obrigatoriedade do cadastro prévio.*