DISSERTAÇÃO DE MESTRADO PROFISSIONAL

FACTUAL: Uma Solução Baseada em Modelos de Linguagem para o Combate à Desinformação

Verônica Souza dos Santos

Programa de Pós-Graduação Profissional em Engenharia Elétrica
DEPARTAMENTO DE ENGENHARIA ELÉTRICA
FACULDADE DE TECNOLOGIA
UNIVERSIDADE DE BRASÍLIA

UNIVERSIDADE DE BRASÍLIA
Faculdade de Tecnologia

DISSERTAÇÃO DE MESTRADO PROFISSIONAL

FACTUAL: Uma Solução Baseada em Modelos de Linguagem para o Combate à Desinformação

Verônica Souza dos Santos

Dissertação de Mestrado Profissional submetida ao Departamento de Engenharia Elétrica como requisito parcial para obtenção do grau de Mestre em Engenharia Elétrica

Banca Examinadora

Prof Dra. Edna Dias Canedo, PPEE/UnB _____
Presidente Orientadora

Prof Dr. Daniel Alves da Silva, PPEE/UnB _____
Examinador Interno

Prof Dr. Gilmar dos Santos Marques, UPIS/DF _____
Examinador Externo

Prof Dr. Georges Daniel Amvame Nz, ENE/UnB _____
Suplente

FICHA CATALOGRÁFICA

SANTOS, VERÔNICA SOUZA DOS
FACTUAL: Uma Solução Baseada em Modelos de Linguagem para oCombate à Desinformação [Distrito Federal] 2025.
xvi, 62 p., 210 x 297 mm (ENE/FT/UnB, Mestre, Engenharia Elétrica, 2025).
Dissertação de Mestrado Profissional - Universidade de Brasília, Faculdade de Tecnologia.
Departamento de Engenharia Elétrica

1. Inteligência Artificial
2. ChatGPT
3. Métodos de Classificação
4. *Fake News*

I. ENE/FT/UnB
II. Título (série)

PUBLICAÇÃO: PPEE.MP.081

REFERÊNCIA BIBLIOGRÁFICA

SANTOS, V.S. (2025). *FACTUAL: Uma Solução Baseada em Modelos de Linguagem para oCombate à Desinformação*. Dissertação de Mestrado Profissional, Departamento de Engenharia Elétrica, Universidade de Brasília, Brasília, DF, 62 p.

Depto. de Engenharia Elétrica (ENE) - FT
Universidade de Brasília (UnB)
Campus Darcy Ribeiro
CEP 70919-970 - Brasília - DF - Brasil

DEDICATÓRIA

Dedico a minha querida mãe, Dalva de Jesus Souza, por sempre acreditar na minha capacidade.

Dedico este trabalho as minhas filhas, Letícia Souza de Mendonça e Paola Souza de Mendonça e ao meu marido Fábio Lúcio Lopes de Mendonça. Pelo impulsionamento para mais essa conquista em minha vida e também ao tempo que deixei de dedicar a nossa família para intensificar pesquisas e estudos para o sucesso da dissertação do tema.

AGRADECIMENTOS

Agradeço a minha orientadora, professora Dra. Edna dias Canedo, que me orientou de forma profissional e amiga na horas mais complicadas durante este trabalho e aturou tantas dúvidas e problemas relativos ao assunto e outros detalhes pertinentes à criação desta dissertação. Aos Professores do Programa de Pós Graduação em Engenharia Elétrica de Universidade de Brasília PPEE/UNB, Daniel Alves da Silva, Rafael Timóteo de Sousa Júnior, Georges D.Amvame Nze, Demétrio Antônio da Silva, Robson de Oliveira Albuquerque, William Ferreira Giozza e Geraldo Pereira Rocha Filho pelas grandes dicas, constante apoio, incentivo e amizade, essenciais para o desenvolvimento deste trabalho. Agradeço também aos membros da banca.

Agradeço o apoio técnico e computacional do Laboratório de Tecnologias para Tomada de Decisão - LATITUDE, da Universidade de Brasília, que conta com apoio do CNPq - Conselho Nacional de Pesquisa (Outorgas 312180/2019-5 PQ-2 e 465741/2014-2 INCT em Cibersegurança), da Advocacia Geral da União (Outorga AGU 697.935/2019), da Procuradoria Geral da Fazenda Nacional (Outorga PGFN 23106.148934/2019-67), da Polícia Federal (Outorga PF 03/2020), do Mestrado Profissional em Engenharia Elétrica, na área de concentração: Segurança Cibernética – 1ª Turma para Profissionais do Setor de Inteligência (Outorga ABIN 01/2019) ao Decanatos de Pesquisa e Inovação e de Pós-Graduação da Universidade de Brasília (Outorga 7129 FUB/EMENDA/DPI/COPEI/AMORIS) e do Projeto SISTER City –Sistemas Inteligentes Seguros e em Tempo Efetivo Real para Cidades Inteligentes (Outorga 625/2022) e a Fundação de Apoio a Pesquisa do Distrito Federal - FAP/DF

Aos meus irmãos pelo apoio e incentivo que foi dado durante todo o tempo em que estive envolvido neste trabalho.

Agradeço em especial ao Professor Geraldo Pereira Rocha Filho e aos meus amigos Bruno Praciano, Flávio Praciano e Paulo Henrique Batista Rodrigues que contribuíram de forma fundamental para a conclusão deste trabalho: meus sinceros agradecimentos.

Agradeço, acima de tudo, a Deus!

RESUMO

A disseminação de *fake news* representa um problema significativo no contexto digital, impactando setores tais como saúde pública, democracia e processos sociais. Apesar de avanços na área, ainda existem lacunas importantes, como a ausência de uma categorização abrangente das formas de desinformação e a limitação de modelos capazes de generalizar em diferentes domínios e conjuntos de dados heterogêneos. Diante disso, esta dissertação propõe o FACTUAL (*fake news* Analysis with Confidence and Trust Using AI and LLM), um sistema baseado em modelos de linguagem de grande escala (LLMs) para a detecção e mitigação de *fake news*. O FACTUAL utiliza uma arquitetura integrada que combina processamento assíncrono para escalabilidade, um mecanismo dinâmico de cálculo de confiança adaptado às análises realizadas e uma categorização de desinformação. Como contribuições, a dissertação apresenta um modelo que permite a análise de grandes volumes de dados, calcula dinamicamente o nível de confiança das predições e estrutura a desinformação em categorias específicas para aprimorar o processo de detecção. A avaliação do FACTUAL foi realizada utilizando os datasets FakeBR e *fake news* Dataset, complementados por dados coletados em portais de notícias. O sistema foi validado quanto à capacidade de identificar padrões de desinformação e quanto à classificação das notícias em categorias apropriadas. Os resultados evidenciam a funcionalidade do FACTUAL, bem como apontam melhorias necessárias, como o refinamento dos limiares de confiança e o balanceamento das categorias classificadas.

Palavras-chave: Inteligência Artificial, Aprendizado de Maquina, Bases de Dados Específicas, Notícias Falsas.

ABSTRACT

The dissemination of fake news represents a significant problem in the digital context, impacting sectors such as public health, democracy, and social processes. Despite advancements in the field, important gaps remain, such as the absence of a comprehensive categorization of misinformation types and the limitation of models capable of generalizing across different domains and heterogeneous datasets. In light of this, this dissertation proposes FACTUAL (Fake News Analysis with Confidence and Trust Using AI and LLM), a system based on large language models (LLMs) for the detection and mitigation of fake news. FACTUAL employs an integrated architecture that combines asynchronous processing for scalability, a dynamic confidence calculation mechanism adapted to the analyses performed, and a categorization of misinformation. As contributions, the dissertation presents a model that enables the analysis of large volumes of data, dynamically calculates the confidence level of predictions, and structures misinformation into specific categories to enhance the detection process. The evaluation of FACTUAL was conducted using the FakeBR and Fake News Dataset, complemented by data collected from news portals. The system was validated for its ability to identify patterns of misinformation and classify news into appropriate categories. The results demonstrate the functionality of FACTUAL, while also highlighting necessary improvements, such as refining confidence thresholds and balancing classified categories.

Keywords: Artificial Intelligence, Machine Learning, Specific Datasets, Fake News

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

1 INTRODUÇÃO

A disseminação de *fake news*, embora não seja um fenômeno recente, se intensificou consideravelmente com o avanço das tecnologias digitais e a crescente popularização das redes sociais. Atualmente, o termo *fake news* descreve a propagação desenfreada de desinformação, muitas vezes alcançando um público mais amplo do que as notícias verdadeiras (1, 2). Esses conteúdos podem ser inteiramente fabricados ou baseados em distorções de fatos reais e são divulgados com o intuito de manipular opiniões, influenciar emoções ou conquistar vantagens políticas, sociais e econômicas. Além disso, as *fake news* se apresentam em uma variedade de formatos, como textos, imagens e vídeos, e têm causado impactos profundos em diversos setores, incluindo mercados financeiros, processos decisórios, mecanismos democráticos e saúde pública (3).

Com o crescimento da internet, a disseminação de informações falsas se tornou mais frequente e acessível, pois qualquer pessoa pode compartilhar conteúdo sem restrições, utilizando plataformas como Facebook e X (4). Um exemplo marcante no Brasil ocorreu nas eleições presidenciais de 2018, quando falsas alegações sobre urnas eletrônicas foram amplamente divulgadas, gerando desconfiança pública e impactando o debate político nacional (5, 6). No cenário global, as eleições presidenciais dos EUA de 2016 evidenciaram como a desinformação pode minar processos democráticos, destacando a gravidade do problema (7, 4). Esses casos reforçam a urgência de implementar regulamentações eficazes e desenvolver soluções tecnológicas avançadas para combater a proliferação de *fake news*, especialmente considerando a velocidade com que essas informações se espalham.

No contexto brasileiro, o Marco Civil da Internet (8) define diretrizes essenciais para assegurar a liberdade de expressão e a proteção de dados pessoais na internet, promovendo uma convivência equilibrada entre os direitos dos usuários e a regulação de conteúdos e serviços digitais. No entanto, essa legislação apresenta limitações no enfrentamento de questões como as *fake news*, especialmente no que diz respeito à remoção rápida de conteúdos falsos e à responsabilização dos provedores de conteúdo. Em um contexto global, eventos como as eleições presidenciais dos EUA de 2016 evidenciam a gravidade do problema, mostrando como a desinformação pode comprometer a integridade de processos democráticos (7, 4). No cenário local, especialmente na saúde pública, a disseminação de informações falsas tem exacerbado crises sanitárias, como ilustrado pela resistência comunitária enfrentada por equipes que implementavam ações contra a dengue em Fortaleza, devido à propagação de notícias falsas nas redes sociais (9). Um fenômeno semelhante foi observado durante a pandemia de COVID-19, quando a disseminação de desinformação sobre tratamentos, vacinas e medidas de prevenção causou confusão e desconfiança nas orientações oficiais. A propagação de notícias falsas relacionadas à pandemia não apenas afetou a eficácia das campanhas de saúde pública, mas também contribuiu para o aumento de comportamentos de risco, prejudicando os esforços globais de contenção do vírus. Esse contexto evidencia a urgência de soluções eficazes para combater a desinformação, particularmente em tempos de crise sanitária.

Esses exemplos, tanto locais quanto globais, ressaltam a urgência de se desenvolver abordagens mais eficazes para mitigar a propagação de *fake news*, especialmente diante das graves consequências que podem resultar da disseminação rápida desse tipo de conteúdo.

Detectar *fake news* é uma tarefa complexa, dada a ambiguidade e subjetividade do conteúdo, a rapidez com que as informações se espalham, a sofisticação na criação de notícias falsas e a escassez de rótulos e dados confiáveis (4, 7). Nesse cenário, a Inteligência Artificial (IA) tem se mostrado uma solução promissora, especialmente através de Modelos de Linguagem de Grande Escala (LLMs). Esses modelos são capazes de processar grandes volumes de dados em tempo real, identificando padrões e indícios de desinformação. Modelos como LLaMA, Mistral e Phi se destacam pela sua habilidade em compreender as sutilezas textuais, permitindo a detecção de informações potencialmente falsas mesmo em contextos complexos e ambíguos (10, 11). Além disso, técnicas de Processamento de Linguagem Natural (PLN) permitem que esses modelos compreendam o contexto e o conteúdo dos textos, ampliando significativamente sua eficácia na verificação da veracidade das informações.

As aplicações de Inteligência Artificial (IA) vão além da detecção de *fake news*, abrangendo uma vasta gama de setores. Ferramentas baseadas em IA, como o ChatGPT, possibilitam a checagem de informações por meio de engenharia de prompt e integração de APIs, permitindo a validação de conteúdos em tempo real. No entanto, apesar de suas notáveis capacidades, essas ferramentas enfrentam desafios relacionados à precisão das respostas. Isso ocorre porque os modelos generativos dependem de dados de treinamento que podem incorporar vieses e inconsistências. Portanto, ao implementar sistemas de verificação baseados em IA, é fundamental levar em consideração essas limitações, a fim de garantir a confiabilidade e a qualidade das informações fornecidas.

Diversas abordagens têm sido exploradas para combater as *fake news*. Estudos como os de (12) e (13) investigam o uso de Modelos de Linguagem de Grande Escala (LLMs) e abordagens multiespecializadas para aprimorar a detecção em múltiplos domínios. No entanto, essas abordagens ainda enfrentam desafios relacionados à complexidade computacional e à presença de vieses. Por outro lado, pesquisas como as de (14) e (15) exploram arquiteturas baseadas em redes neurais convolucionais (CNNs) e abordagens multimodais, que combinam texto e imagem para melhorar a precisão na classificação de *fake news*. Apesar disso, o desequilíbrio de classes nos conjuntos de dados ainda limita a eficácia dessas soluções. Em uma abordagem diferente, (16) explora o uso de aprendizado por transferência com modelos transformadores para detectar *fake news* em idiomas de poucos recursos, enquanto (17) propõe um framework que integra preferências de usuários e redes sociais para otimizar a detecção. Contudo, apesar dos avanços, questões como a dependência de dados históricos e as limitações de escalabilidade continuam representando barreiras significativas para o desenvolvimento dessas tecnologias.

Esta dissertação concentra-se em duas lacunas principais, sendo elas:

- **Falta de uma categorização abrangente das formas de desinformação**: Isso inclui não apenas a desinformação deliberada, mas também informações falsas não intencionais e rumores, que têm impactos diferentes na sociedade e em como podemos combatê-las.
- **Escassez de modelos de LLMs eficientes para múltiplos domínios e dados diversos**: Muitos modelos atuais não conseguem generalizar bem em cenários variados, o que dificulta a criação de soluções mais robustas para detectar *fake news* em contextos distintos.

Essas lacunas representam desafios significativos, mas também oportunidades para impulsionar o estado da arte na área. A seguir, apresento o objetivo principal desta pesquisa, que visa superar esses obstá-

culos e contribuir com soluções inovadoras para o campo.

1.1 OBJETIVO

Desenvolver, implementar e avaliar um sistema baseado em modelos de linguagem de grande escala (LLMs) e frameworks assíncronos, denominado FACTUAL, para a detecção e mitigação de notícias falsas. O estudo busca aprimorar a precisão, a escalabilidade e a transparência no processo de identificação de desinformação, explorando metodologias inovadoras que superem limitações de abordagens tradicionais. Além disso, pretende-se compreender os desafios e impactos dessa abordagem na confiabilidade das informações e na percepção dos usuários, contribuindo para o avanço das soluções tecnológicas no combate à desinformação.

1.2 OBJETIVOS ESPECÍFICOS

Esses objetivos específicos garantem um aprofundamento metodológico e aplicado na construção do FACTUAL, reforçando seu papel como uma solução inovadora e robusta para o combate à desinformação no ambiente digital.

- **Conceituar e Estruturar a Base Teórica**
  - Definir os fundamentos teóricos e técnicos da detecção de notícias falsas, abordando modelos de linguagem de grande escala (LLMs), frameworks assíncronos e mecanismos de cálculo de confiança.
  - Caracterizar os datasets FakeBR e *fake news* Dataset, destacando suas particularidades e relevância para a validação do FACTUAL.
  - Descrever a arquitetura do FACTUAL, detalhando suas camadas funcionais, como banco de dados, modelos de IA, middleware e frameworks de execução.
- **Explorar e Compreender o Modelo**
  - Analisar o funcionamento do modelo LLaMA 3.2 e do framework Ollama na identificação de padrões textuais associados à desinformação.
  - Examinar os resultados do mecanismo dinâmico de cálculo de confiança, identificando padrões preditivos e possíveis vieses.
  - Contextualizar a solução FACTUAL dentro do panorama das principais abordagens acadêmicas e industriais voltadas à detecção de *fake news*.
- **Implementar e Aplicar a Solução**
  - Desenvolver o sistema FACTUAL, integrando os modelos LLaMA 3.2 e Ollama com frameworks assíncronos para garantir escalabilidade e eficiência.

  - Realizar testes empíricos utilizando os datasets selecionados, ajustando os parâmetros do modelo para maximizar sua precisão e confiabilidade.
  - Demonstrar a robustez e escalabilidade do sistema, avaliando seu desempenho sob diferentes cenários de carga e volume de dados.

- **Avaliar o Desempenho e Impacto**
  - Comparar os resultados do FACTUAL com outras abordagens de detecção de *fake news*, utilizando métricas como precisão, revocação e F1-score.
  - Categorizar os diferentes tipos de desinformação identificados pelo sistema, diferenciando conteúdos enganosos, rumores e informações imprecisas.
  - Investigar a influência do mecanismo de confiança na percepção dos usuários sobre a veracidade das informações processadas pelo FACTUAL.

- **Validar e Refinar a Abordagem**
  - Conduzir testes com usuários para avaliar a eficácia e usabilidade do FACTUAL, comparando suas previsões com percepções humanas.
  - Identificar pontos fortes e limitações da arquitetura proposta, sugerindo melhorias para otimização da precisão, escalabilidade e transparência do sistema.
  - Justificar a seleção dos datasets utilizados na pesquisa, analisando sua adequação ao propósito do estudo e sua representatividade na detecção de desinformação.

- **Expandir e Inovar**
  - Propor aprimoramentos para o FACTUAL, explorando novas técnicas de processamento de linguagem natural, aprendizado de máquina e análise multimodal.
  - Desenvolver um protótipo de interface interativa que permita aos usuários submeter conteúdos para análise e visualizar os resultados de forma intuitiva.
  - Elaborar diretrizes para a aplicação do FACTUAL em plataformas digitais e órgãos reguladores, contribuindo para estratégias eficazes de mitigação da desinformação.

1.3 CONTRIBUIÇÕES TÉCNICAS E CIENTÍFICAS

Esta dissertação alcançou tanto contribuições técnicas quanto científicas no contexto da detecção e mitigação de *fake news*. Tais contribuições se destacam tanto pelo avanço teórico no campo quanto pela aplicação prática de técnicas de IA. As principais contribuições deste trabalho incluem:

- Desenvolvimento de uma arquitetura integrada baseada em LLMs para detecção de *fake news*: O sistema FACTUAL foi modelado para lidar com desafios complexos de análise textual. Esta arquitetura não apenas maximiza a precisão da detecção, mas também fornece uma solução escalável e eficiente para análise em tempo real.

- Abordagem Escalável e Eficiente: Implementação de frameworks assíncronos para otimizar a integração com bancos de dados e permitir a execução paralela de tarefas. Essa abordagem aumenta a escalabilidade e eficiência do sistema, permitindo o processamento de grandes volumes de dados em tempo real.
- Categorização Abrangente de Desinformação: Proposta de um mecanismo de categorização para identificar padrões de desinformação, incluindo desinformação deliberada, informações falsas não intencionais e rumores. Essa categorização aprimora a compreensão do fenômeno das *fake news* e auxilia no refinamento das técnicas de detecção.
- Mecanismo Dinâmico de Cálculo de Confiança: Modelagem de um método para calcular o nível de confiança nas análises realizadas pelo sistema. Esse mecanismo considera palavras-chave e expressões indicativas de certeza ou incerteza presentes nas respostas do modelo, ajustando dinamicamente o nível de confiança e proporcionando ao usuário uma compreensão mais profunda da confiabilidade das classificações.

1.4 TRABALHOS PUBLICADOS

Durante o período de execução do mestrado, foram aceitos um total de quatro trabalhos, dos quais dois são artigos como autora principal e dois como coautora. Destaca-se que este trabalho foi laureado com o prêmio de Melhor Artigo da conferência. A seguir, serão elencados os trabalhos publicados.

- (18) VERÔNICA S. S. et al. Combate à Desinformação com FACTUAL: Uma Solução Baseada em Modelos de Linguagem de Grande Escala. 21ª Conferência Ibero Americana WWW/INTERNET (CIAWI), 2024.
  - Destaca-se que este trabalho foi laureado com o prêmio de Melhor Artigo da conferência.
- (19) VERÔNICA S. S. et al. Utilização de Inteligência Artificial para Verificar Notícias Falsas com a Utilização do ChatGPT. 20ª Conferência Ibero Americana WWW/INTERNET (CIAWI), 2023.
- (20) CANEDO, E. D.; VERÔNICA S. S. et al. Do you see what happens around you? Men's Perceptions of Gender Inequality in Software Engineering. SBES 2023: XXXVII Brazilian Symposium on Software Engineering, 2023, Campo Grande Brazil. Proceedings of the XXXVII Brazilian Symposium on Software Engineering. New York: ACM, 2023. v. 37. p. 464-474.
- (21) RABELO, B. S. ; VERÔNICA S. S. et al. Plataforma IoT para predição de falhas em congeladores através de modelo auto-regressivo integrado de média móvel (ARIMA). 19ª Conferência Ibero Americana WWW/INTERNET (CIAWI), 2022.

Os artigos (18) e (19) estão ligados diretamente ao tema dessa dissertação. Já os artigos (20) e (21) foram publicados no início das pesquisas e estão relacionados a pesquisas de Inteligência Artificial e Análise Preditiva, assuntos importantes para essa dissertação.

1.5 ESTRUTURA DA DISSERTAÇÃO

O restante desta dissertação está organizada da seguinte forma.

- Capítulo 2 - Fundamentação Teórica: Este capítulo aborda os princípios teóricos e os conceitos fundamentais que sustentam a formulação e o desenvolvimento da solução proposta, oferecendo a base conceitual necessária para a compreensão do trabalho.
- Capítulo 3 - Trabalhos Relacionados: Este capítulo apresenta diversas abordagens presentes na literatura que abordam o problema investigado nesta pesquisa, comparando as metodologias e resultados desses trabalhos com o modelo proposto nesta dissertação.
- Capítulo 4 - Solução Proposta: Este capítulo apresenta o FACTUAL, descrevendo sua arquitetura, os componentes principais, e as técnicas implementadas para a detecção e mitigação de *fake news*. Também são apresentados o fluxo de processamento e as estratégias adotadas para escalabilidade e eficiência.
- Capítulo 5 - Avaliação de Desempenho: Neste capítulo, são descritos os experimentos realizados para avaliar o desempenho do FACTUAL. Inclui a descrição do dataset, as métricas de avaliação aplicadas, e uma análise dos resultados obtidos.
- Capítulo 6 - Conclusão e Trabalhos Futuros: Neste capítulo, é apresenta uma síntese dos principais achados da pesquisa, além de apresentar direções para trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo estão presentes o conjunto de soluções e utilizadas para alcançar os objetivos da proposta. Alguns conceitos estudados também são apresentados para auxiliar o entendimento das tecnologias, que irão apoiar na execução deste trabalho, podendo assim garantir a possibilidade de replicação dos experimentos utilizados.

2.1 *FAKE NEWS*

O conceito de *fake news* refere-se a informações falsas ou enganosas apresentadas como se fossem notícias verdadeiras. Essas notícias falsas são disseminadas intencionalmente com o objetivo de enganar, desinformar ou manipular a opinião pública sobre determinado assunto. As *fake news* geralmente têm uma aparência de jornalismo legítimo, mas são produzidas com o intuito de causar confusão, induzir ao erro, ou influenciar comportamentos e crenças de maneira prejudicial (22).

Exemplos de *fake news*:

Figura 2.1: Exemplo de *fake news*.
Fonte: Redes Sociais

As *fake news* são um fenômeno perigoso que pode afetar significativamente a sociedade ao desinformar, manipular e causar impactos graves em diversas áreas, como política, saúde e relações sociais. No Caso da Figura 2.1, temos um exemplo de uma imagem alterada em que a notícia diz "Documentos comprovam que Lula foi alertado da Catástrofe das queimadas antes delas acontecerem". Nesse caso trata-se de uma

mensagem falsa, pois, segundo pesquisas em fontes de informações seguras, a mensagem correta foi "Governo Lula é alertado **sobre risco de seca e incêndios** desde o início do ano, mostram documentos"postada em 13 de setembro 2024 em vários veículos de comunicação. Entretanto, a informação anterior causa uma entonação dúbia, provocando uma desinformação, quando na realidade em 10 de setembro de 2024 foi lançando a nota "10 de set. de 2024 — Lula anuncia investimentos contra efeitos da seca e alerta para incêndios criminosos"através da fonte: "https://agenciagov.ebc.com.br/noticias"neste caso podemos verificar que após a noticia lançada, pessoas utilizaram da noticias para rapidamente criar uma desinformação ao invés de uma informação.

Outro caso, que podemos analisar é sobre uma mensagem do ex-presidente Jair Bolsonaro em 2020, afirmando que Bolsonaro havia sancionado uma lei que legalizava o abuso sexual de crianças e adolescentes, conforme Figura 2.2

Figura 2.2: Exemplo 2 de Fake News.
Fonte: Redes Sociais

Essa informação é completamente falsa e distorcia, o conteúdo de uma Medida Provisória assinada por Bolsonaro, que, na verdade, tratava de regularizações relacionadas a certidões de nascimento e não tinha nenhuma relação com abuso sexual.

Sintetizamos uma literatura crescente investigando por que as pessoas acreditam e compartilham notícias falsas ou altamente enganosas online. Ao contrário de uma narrativa comum pela qual a política impulsiona a suscetibilidade a notícias falsas, as pessoas são "melhores"em discernir a verdade da falsidade (apesar da maior crença geral) ao avaliar notícias politicamente concordantes. Em vez disso, o discernimento pobre da verdade está associado à falta de raciocínio cuidadoso e conhecimento relevante, e ao uso de heurísticas como familiaridade.

Além disso, há uma desconexão substancial entre o que as pessoas acreditam e o que compartilham nas mídias sociais. Essa dissociação é amplamente motivada pela desatenção, mais do que pelo compartilhamento proposital de desinformação. Assim, as intervenções podem incitar com sucesso os usuários de mídia social a se concentrarem mais na precisão. As classificações de veracidade crowd sourced também podem ser aproveitadas para melhorar os algoritmos de classificação de mídia social (23).

2.1.1 Características das *fake news*

As *fake news*, ou notícias falsas, tornaram-se um fenômeno alarmante na era digital, especialmente com o crescimento das redes sociais e suas implicações na política e na sociedade. As *fake news* geralmente apelam para emoções fortes, como medo e raiva, e são disseminadas rapidamente através das redes sociais (3). Essas informações são geradas com o intuito deliberado de enganar e manipular a opinião pública. Podemos explorar as características mais relevantes das *fake news* como:

- **Intencionalidade**: Diferentemente de erros comuns de jornalismo, as *fake news* são elaboradas de forma consciente para confundir e manipular. Muitas vezes, isso ocorre com objetivos políticos ou econômicos, visando desestabilizar adversários ou promover agendas específicas.
- **Apelo Emocional**: Essas notícias frequentemente utilizam uma linguagem sensacionalista e alarmista, apelando para as emoções dos leitores. Esse tipo de abordagem provoca reações intensas, o que facilita sua rápida disseminação, já que as pessoas tendem a compartilhar conteúdos que as tocam emocionalmente.
- **Viralidade**: A natureza das redes sociais permite que *fake news* se espalhem em um curto espaço de tempo. Graças ao seu apo emocional e à facilidade de compartilhamento, elas podem alcançar um grande público em questão de horas.
- **Falta de Fontes Confiáveis**: As *fake news* costumam carecer de referências legítimas. Muitas vezes, citam fontes inexistentes ou distorcem informações de fontes reais, o que dá uma falsa impressão de credibilidade.
- **Manipulação de Mídia Visual**: Imagens e vídeos frequentemente são editados de forma a criar narrativas enganosas. Essa manipulação é muitas vezes sutil, tornando difícil para o público discernir a verdade.
- **Objetivos de Manipulação**: Em contextos eleitorais, as *fake news* são ferramentas poderosas para manipular a opinião pública. Elas podem ser utilizadas para difamar concorrentes ou para influenciar a percepção do eleitorado.
- **Dificuldade de Desmentido**: Uma vez que uma *fake news* se espalha, desmenti-la se torna uma tarefa árdua. Mesmo após correções, a desinformação pode continuar a circular e influenciar a opinião das pessoas, perpetuando mitos e ideias errôneas.

Além das características citadas, as *fake news* têm implicações diretas na **segurança cibernética**, pois muitas vezes são utilizadas em golpes virtuais, que são formas de crimes digitais projetadas para enganar e explorar as vítimas. Esses golpes podem envolver ***phishing***, onde os atacantes se fazem passar por organizações legítimas para obter informações sensíveis, ou até mesmo ataques mais sofisticados como ***ransomware***, que sequestram dados e exigem pagamento para liberação. A manipulação de informações e a rápida propagação das *fake news* pode ser vista como uma ferramenta poderosa para **golpes virtuais**, prejudicando tanto indivíduos quanto organizações.

A **segurança da informação e a cibersegurança** são essenciais para prevenir e mitigar os impactos desses ataques, protegendo dados e sistemas contra acessos não autorizados. Com o aumento da complexidade das ameaças digitais, é necessário adotar uma abordagem integrada, que envolva desde a proteção de sistemas até o uso de **análise forense digital** para identificar e investigar ataques cibernéticos. A análise forense digital permite a coleta e análise de evidências de crimes cibernéticos, possibilitando a identificação dos responsáveis e a compreensão dos métodos utilizados nas fraudes digitais.

Outra prática importante no combate aos golpes virtuais é o uso de ***Open Source Intelligence - OSINT***, que se refere à coleta de informações disponíveis publicamente, como dados de redes sociais, sites de notícias e registros públicos. O monitoramento de redes sociais também é fundamental para identificar a disseminação de *fake news* e outros tipos de desinformação em tempo real. Essas ferramentas ajudam na identificação de campanhas fraudulentas e fornecem dados valiosos para a correção de informações falsas antes que causem danos maiores.

As *fake news* representam, portanto, um desafio significativo para a sociedade contemporânea, demandando um esforço coletivo em educação midiática, promoção do pensamento crítico, e adoção de práticas eficazes de segurança cibernética. Além disso, é fundamental que indivíduos e organizações se conscientizem da importância da verificação de informações e do monitoramento contínuo de atividades digitais, para criar um ambiente informativo mais saudável e responsável.

2.1.2 Tipos de *Fake News*

Este estudo categoriza diferentes tipos de *fake news*, incluindo sátira, paródia, conteúdo enganoso e conteúdo fabricado. A pesquisa destaca como cada tipo de fake news pode ser identificado e os desafios associados à sua detecção (24). Com a ascensão das redes sociais, é crucial identificar seus diversos tipos.

- **Notícias Fabricadas**: Esse tipo se refere a informações completamente inventadas, criadas sem qualquer base na realidade. Normalmente, essas histórias buscam gerar cliques e engajamento nas redes sociais, como relatos fictícios sobre celebridades ou eventos que nunca ocorreram.
- **Clickbait**: Trata-se de títulos sensacionalistas que atraem o leitor, mas muitas vezes não correspondem ao conteúdo real da matéria. Essa prática tem como objetivo aumentar o tráfego e, frequentemente, resulta em desinformação, já que o conteúdo pode ser irrelevante ou distorcido.
- **Desinformação**: Esse tipo abrange a disseminação de informações enganosas, que, embora possam ter uma origem verdadeira, são apresentadas fora de contexto ou manipuladas de forma a alterar sua interpretação.
- **Misinformação**: Ao contrário da desinformação, a misinformação se refere ao compartilhamento de informações incorretas sem a intenção de enganar. Isso pode ocorrer devido à falta de verificação de fatos ou ao compartilhamento apressado de conteúdos.
- **Rumores**: Esses são dados não confirmados que circulam entre as pessoas, frequentemente baseados em verdades parciais. Rumores tendem a se proliferar em contextos sensíveis, como crises sociais ou políticas, contribuindo para a confusão e desinformação.

- **Sátira ou Paródia**: Embora não sejam *fake news* no sentido estrito, conteúdos humorísticos que imitam a aparência de notícias podem ser mal interpretados. Embora sua intenção seja o entretenimento, eles podem levar leitores a acreditarem em informações falsas.
- Falsa Conexão: Isso ocorre quando há uma desconexão entre títulos, imagens e o conteúdo real da notícia. Tal distorção pode levar o público a conclusões erradas, apenas com base em uma manchete chamativa ou uma imagem enganosa.
- Falso Contexto: Esse tipo se refere ao uso de informações verdadeiras, mas apresentadas em um contexto enganoso. Um exemplo seria utilizar uma imagem real de um evento para contar uma história que nada tem a ver com o que realmente ocorreu.
- **Conteúdo Impostor**: Envolve a criação de sites ou perfis que imitam fontes legítimas, enganando os usuários sobre a autenticidade da informação divulgada.
- **Conteúdo Manipulado**: Esse tipo refere-se à edição de informações ou imagens com a intenção de enganar. Isso pode incluir a alteração de fotos ou vídeos para criar uma narrativa falsa.
- **Golpes Virtuais e Segurança Cibernética**:
  - A crescente interligação entre *fake news* e golpes virtuais.
  - Uso de *fake news* como ferramenta para enganar usuários e propagar golpes como *phishing, ransomware* e fraudes online.
  - Impacto negativo das *fake news* na segurança cibernética, explorando a vulnerabilidade emocional das vítimas para a coleta de dados ou execução de ataques.
- **Análise Forense Digital**:
  - Importância da análise forense digital na identificação e rastreamento da origem de golpes virtuais e *fake news*.
  - Auxílio na coleta de evidências digitais, permitindo a compreensão e resposta eficaz aos ataques cibernéticos.
  - Papel da análise forense na resolução de incidentes de segurança cibernética, incluindo a recuperação de dados e investigação de crimes virtuais.
- ***Open Source Intelligence - OSINT***:
  - Uso de fontes de dados públicas para monitorar e coletar informações relacionadas a *fake news* e golpes virtuais.
  - Aplicação de OSINT para detectar padrões de disseminação de *fake news* e rastrear possíveis origens de campanhas fraudulentas.
  - Ferramenta poderosa para analisar redes sociais e outras plataformas online para identificar informações falsas e fontes de desinformação.
- **Monitoramento de Redes Sociais**:

- Monitoramento das redes sociais como meio de rastrear e identificar campanhas de *fake news* em tempo real.
- Importância do monitoramento para identificar a disseminação de desinformação e golpes virtuais.
- Uso de ferramentas avançadas para identificar fontes e influenciadores de *fake news*, ajudando na prevenção e mitigação de ataques cibernéticos e desinformação.

Cada categoria traz suas próprias dificuldades, especialmente em relação à verificação de fatos que se torna cada vez mais complexa. Devido a esses desafios, a verificação de *fake news* exige uma abordagem multidisciplinar que envolve o uso de tecnologias avançadas, colaboração entre verificadores de fatos e a conscientização crítica dos usuários. Investir em educação midiática e fomentar práticas de checagem antes do compartilhamento pode ajudar a reduzir a propagação de informações falsas e seus impactos na sociedade.

2.1.3 Impactos das *fake news*

As *fake news* são conteúdos enganosos não apenas distorcem a realidade, mas também afetam profundamente diversos aspectos da vida social, política, econômica, incluindo danos à reputação, perda de confiança do público e impactos financeiros e como podem influenciar decisões empresariais e políticas públicas (25).

- Erosão da Confiança nas Instituições: Um dos principais impactos das *fake news* é a erosão da confiança nas instituições, como a mídia, o governo e outras entidades. Quando as pessoas são constantemente expostas a informações falsas, sua capacidade de confiar em fontes legítimas diminui. Isso cria um ambiente de ceticismo generalizado, onde até mesmo fatos verificáveis podem ser questionados, dificultando a comunicação efetiva entre instituições e cidadãos.
- Polarização Social As *fake news* alimentam a polarização social, exacerbando divisões entre diferentes grupos. Muitas vezes, essas notícias são usadas para reforçar preconceitos e estereótipos, resultando em um ambiente de hostilidade e desconfiança mútua. Esse fenômeno pode transformar o debate público em um campo de batalha, onde o diálogo construtivo se torna cada vez mais difícil.
- Ameaça à Democracia: O impacto das *fake news* na política é especialmente preocupante. Durante eleições e plebiscitos, informações enganosas podem manipular a opinião pública, influenciando resultados e comprometendo a integridade dos processos democráticos. A desinformação não apenas distorce a realidade, mas também mina a capacidade dos cidadãos de tomar decisões informadas.
- Impacto na Saúde Pública: Em crises sanitárias, como a pandemia de COVID-19, as *fake news* podem ter consequências graves. Informações falsas sobre vacinas e tratamentos podem levar a comportamentos prejudiciais à saúde, dificultando os esforços de controle e aumentando a carga sobre os sistemas de saúde. A disseminação de desinformação pode desviar recursos e atenção de iniciativas legítimas, prejudicando a resposta a emergências.

- Danos Econômicos: As *fake news* também causam danos econômicos significativos. Empresas podem enfrentar perdas financeiras devido a boatos que afetam sua reputação. Além disso, a desinformação pode impactar os mercados financeiros, resultando em flutuações injustificadas nos preços de ações e outros ativos, gerando incerteza econômica.

- Desinformação e Tomada de Decisões: A propagação de *fake news* contribui para a desinformação geral, dificultando a capacidade da população de tomar decisões informadas. Isso afeta desde escolhas cotidianas, como consumo, até decisões políticas, prejudicando o bem-estar individual e coletivo.

- Propagação de Ódio e Violência: Em casos extremos, as *fake news* podem incitar ódio e violência. Notícias falsas que demonizam grupos específicos podem levar a ataques físicos e verbais, além de fomentar um ambiente de intolerância. Esse cenário não apenas compromete a segurança, mas também agrava tensões sociais.

Os impactos das *fake news* são profundos e multifacetados, afetando não apenas a esfera política e econômica, mas também a saúde, o convívio social e o bem-estar individual. Combater a desinformação exige esforços conjuntos, que incluem a educação midiática, a promoção de uma cultura de verificação dos fatos e o desenvolvimento de ferramentas tecnológicas capazes de identificar e limitar a disseminação de conteúdos falsos. Dessa forma, é possível minimizar os danos e fortalecer a confiança nas instituições e na informação compartilhada.

2.1.4 Combate às *fake news*

Neste tópico serão abordados algumas das principais estratégias para o enfrentamento às Fake New, mostrando como essas tecnologias são essenciais para desenvolver ferramentas que combatam as *fake news* (26).

- Educação Midiática:Uma das ferramentas mais eficazes no combate às *fake news* é a educação midiática. Essa abordagem visa capacitar cidadãos a se tornarem consumidores críticos de informações. A alfabetização digital é fundamental, pois ensina como identificar fontes confiáveis e verificar informações. Além disso, o desenvolvimento do pensamento crítico ajuda a questionar a veracidade das notícias antes de compartilhá-las. Programas em países como a Finlândia, que introduziram a educação midiática nas escolas, são exemplos positivos a serem seguidos.

- Verificação de Fatos O fortalecimento de agências de verificação de fatos é essencial para desmantelar a desinformação. Essas organizações analisam e desmentem notícias falsas, fornecendo informações corretas ao público. Iniciativas como o "Fato ou Fake" no Brasil exemplificam esforços bem-sucedidos nesse sentido. A colaboração entre plataformas digitais e agências de checagem é crucial para rotular informações enganosas e ajudar os usuários a discernir entre verdade e mentira.

- Uso de Tecnologia A tecnologia também desempenha um papel importante no combate às *fake news*. Ferramentas baseadas em inteligência artificial e algoritmos podem detectar conteúdos enganosos antes que se espalhem. Sistemas que analisam padrões de compartilhamento são essenciais para

identificar informações potencialmente falsas. Além disso, criar mecanismos onde os usuários possam relatar conteúdos suspeitos ajuda a promover um ambiente mais seguro e informativo.

- Responsabilidade das Plataformas Digitais As empresas de tecnologia têm a responsabilidade de implementar políticas eficazes no combate às *fake news*. Regulamentações claras sobre a divulgação de informações falsas e suas consequências para usuários e páginas são fundamentais. Campanhas de conscientização promovidas por essas plataformas também são vitais, incentivando a verificação de informações antes do compartilhamento e fomentando uma cultura de responsabilidade digital.

- Papel das Instituições e Governos Governos e instituições têm um papel crucial no enfrentamento das *fake news*. A implementação de legislações que responsabilizem os criadores de *fake news*, respeitando a liberdade de expressão, é essencial. Além disso, é importante promover iniciativas culturais que incentivem a ética na comunicação e a valorização da verdade.

- Colaboração Internacional A desinformação é um problema global que requer uma resposta coordenada entre países. A troca de melhores práticas e estratégias pode fortalecer os esforços no combate às *fake news* em nível internacional.

- Responsabilidade Individual Cada cidadão também desempenha um papel vital nesse combate. Verificar a veracidade das informações antes de compartilhá-las, denunciar conteúdos falsos e promover uma cultura de pensamento crítico nas redes sociais são atitudes que podem fazer a diferença.

O combate às *fake news* não exige apenas uma abordagem integrada que envolva educação, tecnologia, regulamentação e conscientização pública. mas também por meio de esforços coordenados será possível diminuir os impactos negativos da desinformação na sociedade.

2.2 INTELIGÊNCIA ARTIFICIAL - IA

Conforme descrito por (27) em *"Artificial Intelligence: A Modern Approach"*, um agente inteligente é um sistema que percebe seu ambiente através de sensores e atua sobre ele através de atuadores, com o objetivo de maximizar uma função de desempenho. Essa definição operacional permite a modelagem de sistemas que vão desde programas de computador para jogos até robôs autônomos.

A Inteligência Artificial -*IA* é um ramo da ciência da computação que se dedica ao desenvolvimento de sistemas e algoritmos capazes de realizar tarefas que, normalmente, requereriam inteligência humana. Isso inclui habilidades como:

- Raciocínio e tomada de decisões: Processar informações e tomar decisões com base em dados.

- Aprendizado: Adaptar-se e melhorar o desempenho através de experiências, muitas vezes utilizando técnicas de aprendizado de máquina.

- Reconhecimento de padrões: Identificar padrões em dados, o que é fundamental para tarefas como reconhecimento de imagens e processamento de linguagem natural.

- Processamento de linguagem natural: Entender e gerar linguagem humana, permitindo a comunicação entre pessoas e máquinas.

Segundo (28) a *IA* permite que os sistemas técnicos percebam o ambiente que os rodeia, lidem com o que percebem e resolvam problemas, agindo no sentido de alcançar um objetivo específico. O computador recebe dados (já preparados ou recolhidos através dos seus próprios sensores, por exemplo, com o uso de uma câmara), processa-os e responde. Os sistemas de *IA* são capazes de adaptar o seu comportamento, até certo ponto, através de uma análise dos efeitos das ações anteriores e de um trabalho autônomo. Contam com os *Large Language Models*, que em português pode ser traduzido como Modelos de Linguagem de Grande Escala. Esses modelos são sistemas de inteligência artificial, geralmente baseados em arquiteturas de redes neurais profundas, como os *transformers*, e são treinados com grandes quantidades de dados textuais. A seguir, alguns pontos importantes sobre os LLMs:

- Treinamento em Grandes Corpora:Os LLMs são alimentados com vastos conjuntos de dados provenientes de diversas fontes, permitindo-lhes aprender uma ampla gama de padrões e estruturas da linguagem. Esse treinamento massivo permite a generalização em múltiplas tarefas de processamento de linguagem natural (PLN).

- Arquitetura *Transformer*: A maioria dos LLMs modernos utiliza a arquitetura *transformer*, introduzida por Vaswani et al. (2017). Essa arquitetura se destaca por sua capacidade de capturar dependências de longo alcance em textos, o que é essencial para a compreensão e geração de linguagem natural de forma coerente.

- Número de Parâmetros: Uma característica marcante dos LLMs é a grande quantidade de parâmetros que eles possuem, frequentemente na ordem de milhões ou até bilhões. Esse alto número de parâmetros permite uma modelagem mais detalhada e complexa das nuances linguísticas, embora também imponha desafios em termos de computação e interpretabilidade.

- Aplicações: Devido à sua capacidade de entender e gerar texto de forma sofisticada, os LLMs são empregados em uma variedade de tarefas, como:Tradução automática, Resumo de textos Geração de conteúdo, Resposta a perguntas e Assistentes virtuais

- Desafios e Considerações Éticas: Embora os LLMs representem avanços significativos no campo do PLN, eles também trazem desafios, como a necessidade de grande poder computacional, questões de viés nos dados de treinamento, e preocupações relacionadas à privacidade e à ética no uso das tecnologias de *IA*.

A inteligência em máquinas remonta a trabalhos pioneiros, como o de (29), que propôs o Teste de *Turing* em 1950 como um critério para a avaliação da inteligência de uma máquina. Ao longo das décadas, a *IA* passou por diversas fases, incluindo períodos de grande otimismo e os chamados "invernos da *IA*", quando o financiamento e o interesse diminuíram devido às limitações das tecnologias disponíveis na época. Atualmente, o ressurgimento e os avanços em áreas como o aprendizado de máquina e o processamento de grandes volumes de dados *(big data)* têm impulsionado novas aplicações e debates éticos sobre o papel da *IA* na sociedade.

Em síntese, a Inteligência Artificial é uma área robusta e multifacetada, que combina fundamentos teóricos com aplicações práticas, e continua a evoluir à medida que novas técnicas e paradigmas são desenvolvidos. Seu caráter interdisciplinar e os desafios associados à replicação da inteligência humana fazem dela um campo dinâmico e repleto de questões tanto técnicas quanto filosóficas (27).

Neste sentido, a *IA* tem influência diretamente neste trabalho, quando se pensa em buscar noticias falsas ou verdadeiras em grandes massa de informações, utilizando aprendizado de máquina e o processamento de grandes volumes de dados *(big data)* que têm impulsionado sobre mensagens distribuídas da internet para que possamos tomar decisões diante das informações coletadas.

2.2.1 Aprendizado de Máquina (ML)

Aprendizado de Máquina, do inglês *Machine learning - ML*, é o estudo científico de algoritmos e modelos estatísticos que sistemas computacionais utilizam para executar uma tarefa específica sem serem explicitamente programados. Algoritmos de aprendizado são a base de muitas aplicações que usamos diariamente, como, por exemplo, toda vez que um motor de busca como o *Google* é usado para pesquisar na internet, um dos motivos pelo qual ele funciona tão bem é devido a um algoritmo de aprendizado que aprendeu como classificar páginas da web. Esses algoritmos são usados para várias finalidades, como mineração de dados, processamento de imagens, análise preditiva, entre outras. (30).

É um dos campos pertencentes à computação moderna, em que diversos estudos foram realizados para tornar as máquinas inteligentes. *ML* pode ser classificado de duas formas: 1) supervisionado; 2) não supervisionado. Um aprendizado é considerado supervisionado somente se os dados utilizados estão antecipadamente rotulados, ou se a resposta objetiva já é conhecida. De outro modo, o aprendizado de máquina é tido como não supervisionado quando os dados não possuem um atributo alvo e não foram previamente rotulados. No presente estudo, o conceito e a aplicação do aprendizado de máquina são referente ao aprendizado supervisionado (31).

Além disso, Podemos dizer que *ML* é o subconjunto da inteligência artificial que se concentra na construção de sistemas que aprendem, ou melhoram o desempenho, com base nos dados que consomem. A inteligência artificial é um termo amplo que se refere a sistemas ou máquinas que imitam a inteligência humana (31).

O *machine learning* e a *IA* são frequentemente abordados juntos, e os termos às vezes são usados de forma intercambiável, mas não significam a mesma coisa. Uma distinção importante é que, embora todo machine learning seja *IA*, nem toda *IA* é *machine learning*.

A principal vantagem de usar *machine learning* é que, uma vez que um algoritmo aprende o que fazer com os dados, ele pode realizar seu trabalho automaticamente.

2.2.2 Processamento de Linguagem Natural (PLN)

O Processamento de Linguagem Natural (PLN) do inglês *Natural Language Processing (NLP)* é um campo interdisciplinar que integra conceitos de linguística, ciência da computação e inteligência artificial para permitir que máquinas compreendam, interpretem e gerem linguagem humana de forma significativa.

Essa área busca desenvolver métodos e algoritmos capazes de lidar com a complexidade e a ambiguidade inerentes à comunicação natural, abordando desafios como polissemia, ambiguidade sintática e semântica, além da variação linguística e contextual (32).

Historicamente, o PLN evoluiu a partir de abordagens baseadas em regras e métodos estatísticos para técnicas avançadas de aprendizado de máquina e deep learning. O surgimento de modelos baseados em redes neurais, como os transformers, revolucionou o campo ao possibilitar a captura de dependências de longo alcance e a modelagem de contextos complexos, aprimorando tarefas como tradução automática, análise de sentimentos, resumo automático de textos e reconhecimento de fala.

Além das aplicações práticas, o PLN tem implicações teóricas importantes, contribuindo para a compreensão de como a linguagem é estruturada e processada tanto por humanos quanto por sistemas computacionais. As pesquisas atuais buscam não só melhorar a acurácia e eficiência dos sistemas de processamento, mas também abordar questões éticas e de viés que emergem no tratamento de dados linguísticos, promovendo o desenvolvimento de tecnologias mais inclusivas e responsáveis (33).

O Processamento de Linguagem Natural (PLN) possui características, vantagens e desvantagens que o tornam uma área de intensa pesquisa e aplicação prática. A seguir, apresenta-se uma síntese desses aspectos:

2.2.2.1 Características

- Interdisciplinaridade: O PLN integra conhecimentos de linguística, ciência da computação, estatística e inteligência artificial para interpretar e gerar linguagem humana.
- Complexidade Linguística: Envolve o tratamento de desafios inerentes à linguagem, como ambiguidade, polissemia, variações regionais e contextuais, o que exige abordagens sofisticadas para lidar com diferentes níveis de linguagem (morfologia, sintaxe, semântica e pragmática).
- Abordagens Baseadas em Dados: Historicamente evoluiu de métodos baseados em regras para técnicas estatísticas e, atualmente, para modelos de deep learning (por exemplo, redes neurais e transformers), que dependem de grandes quantidades de dados para aprendizado.
- Aplicabilidade Multidisciplinar: As técnicas de PLN são aplicadas em diversas tarefas, incluindo tradução automática, análise de sentimentos, reconhecimento de fala, sumarização automática e extração de informações.

2.2.2.2 Vantagens

- Automatização de Tarefas: Permite a automação de processos que envolvem o processamento de grandes volumes de texto e voz, otimizando fluxos de trabalho e reduzindo custos operacionais.
- Melhoria na Comunicação: Facilita a interação entre humanos e sistemas computacionais por meio de interfaces mais naturais, como assistentes virtuais e chatbots, melhorando a experiência do usuário.

- Aprimoramento de Decisões: Através da análise de grandes volumes de dados textuais, o PLN auxilia na extração de insights que podem embasar decisões estratégicas em diversas áreas, como marketing, saúde e finanças.
- Evolução Tecnológica: Os avanços recentes em modelos pré-treinados (por exemplo, BERT, GPT) proporcionaram melhorias significativas na acurácia e na capacidade de compreensão contextual, ampliando as possibilidades de aplicação.

2.2.2.3 Desvantagens

- Ambiguidade e Contexto: Apesar dos avanços, os modelos de PLN ainda enfrentam dificuldades em lidar com ambiguidade e nuances contextuais, o que pode levar a interpretações equivocadas em certos casos.
- Dependência de Dados: O desempenho dos sistemas de PLN é altamente dependente da qualidade e quantidade dos dados de treinamento. A escassez ou viés nos dados pode comprometer a eficácia e a imparcialidade dos resultados.
- Complexidade Computacional: Modelos avançados, especialmente os baseados em deep learning, requerem elevado poder computacional para treinamento e inferência, o que pode limitar sua implementação em ambientes com recursos restritos.
- Viés e Ética: Sistemas de PLN podem reproduzir e amplificar vieses presentes nos dados de treinamento, levantando questões éticas e de justiça na tomada de decisão automatizada.

Nesse sentido, o Processamento de Linguagem Natural representa uma área de intensa atividade científica e tecnológica, cujo desenvolvimento contínuo tem impacto significativo em diversas esferas da sociedade, facilitando a interação entre humanos e máquinas e ampliando as possibilidades de análise e gerenciamento de grandes volumes de informação textual.

2.3 IMPACTO DA IA NA SOCIEDADE

2.3.1 Governança e IA segundo o OpenIA

Segundo (34) é concebível que, nos próximos dez anos, os sistemas de IA excedam o nível de habilidade especializada na maioria dos domínios e realizem tanta atividade produtiva quanto uma das maiores corporações de hoje.Em termos de possíveis vantagens e desvantagens, a superinteligência será mais poderosa do que outras tecnologias com as quais a humanidade teve que lidar no passado. Podemos ter um futuro dramaticamente mais próspero; mas temos que administrar o risco para chegar lá. Dada a possibilidade de risco existencial, não podemos ser apenas reativos. A energia nuclear é um exemplo histórico comumente usado de uma tecnologia com essa propriedade; a biologia sintética é outro exemplo.

Também devemos mitigar os riscos da tecnologia de IA de hoje, mas a superinteligência exigirá tratamento e coordenação especiais. Existem muitas ideias importantes para termos uma boa chance de navegar

com sucesso nesse desenvolvimento; aqui apresentamos nosso pensamento inicial sobre três deles.

Primeiro, precisamos de algum grau de coordenação entre os principais esforços de desenvolvimento para garantir que o desenvolvimento da superinteligência ocorra de uma maneira que nos permita manter a segurança e ajudar na integração suave desses sistemas com a sociedade. Existem muitas maneiras de implementar isso; os principais governos ao redor do mundo poderiam estabelecer um projeto do qual muitos esforços atuais se tornem parte, ou poderíamos concordar coletivamente (com o apoio de uma nova organização como a sugerida abaixo) que a taxa de crescimento na capacidade de IA na fronteira é limitada a uma determinada taxa por ano.(35)

Em segundo lugar, é provável que eventualmente precisemos de algo como uma IAEApara esforços de superinteligência; qualquer esforço acima de um determinado limite de capacidade (ou recursos como computação) precisará estar sujeito a uma autoridade internacional que pode inspecionar sistemas, exigir auditorias, testar a conformidade com os padrões de segurança, impor restrições aos graus de implantação e níveis de segurança, etc. Rastrear o uso de computação e energia pode ajudar bastante e nos dar alguma esperança de que essa ideia possa ser implementada. Como primeiro passo, as empresas poderiam concordar voluntariamente em começar a implementar elementos do que tal agência poderia um dia exigir e, como segundo, países individuais poderiam implementá-lo. Seria importante que tal agência se concentrasse na redução do risco existencial e não em questões que deveriam ser deixadas para países individuais, como definir o que uma IA deveria ter permissão para dizer.

Em terceiro lugar, precisamos de capacidade técnica para tornar uma superinteligência segura. Esta é uma questão de pesquisa em aberto na qual nós e outros estamos nos empenhando muito.

Mas a governança dos sistemas mais poderosos, bem como as decisões relativas à sua implantação, devem ter forte supervisão pública. Acreditamos que as pessoas ao redor do mundo devem decidir democraticamente sobre os limites e padrões dos sistemas de IA. Ainda não sabemos como projetar tal mecanismo, mas planejamos experimentar seu desenvolvimento. Continuamos a pensar que, dentro desses limites amplos, os usuários individuais devem ter muito controle sobre como a IA que usam se comporta.(35)

Na OpenAI, temos dois motivos fundamentais. Primeiro, acreditamos que levará a um mundo muito melhor do que podemos imaginar hoje (já estamos vendo exemplos iniciais disso em áreas como educação, trabalho criativo e produtividade pessoal). O mundo enfrenta muitos problemas que precisaremos de muito mais ajuda para resolver; essa tecnologia pode melhorar nossas sociedades, e a capacidade criativa de todos para usar essas novas ferramentas certamente nos surpreenderá. O crescimento econômico e o aumento da qualidade de vida serão surpreendentes (36).

Em segundo lugar, acreditamos que seria arriscado e difícil impedir a criação de superinteligência. Como as vantagens são enormes, o custo para construí-lo diminui a cada ano, o número de atores que o constroem está aumentando rapidamente e é inerentemente parte do caminho tecnológico em que estamos, interrompê-lo exigiria algo como um regime de vigilância global e mesmo isso não é garantido para funcionar. Então temos que acertar.

2.3.2 Discriminação algorítmica

Segundo (37) existem muitos algoritmos de tomada de decisão a serem considerados aqui, como eleger representantes, votar por maioria, empregar democracia líquida e tomar decisões por uma amostra aleatória da população , também conhecida como júri ou sortition.

Algoritmos de tomada de decisão em IA são métodos que permitem a um agente escolher ações com base em seu ambiente e objetivos, visando otimizar uma determinada função de desempenho ou utilidade. Essas abordagens podem variar bastante, desde métodos determinísticos baseados em regras até técnicas probabilísticas e de aprendizado. A seguir, descrevo algumas das principais categorias e exemplos:

Sistemas Baseados em Regras e Lógica:

- Sistemas Especialistas: Utilizam conjuntos de regras condicionais ("se... então...") para inferir conclusões ou tomar decisões com base em um banco de conhecimento.
- Lógica Formal e Raciocínio Baseado em Conhecimento: Empregam técnicas de inferência lógica para deduzir ações apropriadas a partir de um conjunto de fatos e regras.

Algoritmos de Busca e Planejamento:

- Busca em Espaço de Estados: Algoritmos como A*, busca em largura e profundidade são usados para encontrar sequências de ações que conduzam a um estado objetivo.
- Planejamento Automático: Métodos que geram planos de ação (por exemplo, planejamento hierárquico) para atingir metas em ambientes dinâmicos e complexos.

Aprendizado por Reforço (Reinforcement Learning):

- Q-Learning e SARSA: Permitem que um agente aprenda uma política de ação através da interação com o ambiente, atribuindo valores (recompensas) a pares estado-ação e ajustando suas escolhas para maximizar a recompensa acumulada.
- Métodos Baseados em Políticas (Policy Gradients): Aprendem diretamente a política que mapeia estados em ações, sendo úteis em ambientes com espaços de ação contínuos ou de alta dimensão.

Algoritmos de Otimização e Busca Heurística:

- Algoritmos Genéticos e Métodos Evolutivos: Inspirados na evolução biológica, esses algoritmos geram e selecionam soluções com base em um processo iterativo de mutação, cruzamento e seleção, explorando um grande espaço de soluções.
- Simulated Annealing: Uma técnica probabilística que busca escapar de mínimos locais, permitindo decisões que, inicialmente, podem parecer subótimas para alcançar uma solução global melhor.

Métodos Probabilísticos e de Decisão sob Incerteza:

- Processos de Decisão Markovianos (MDP) e POMDP: Modelam ambientes onde o resultado das ações envolve incerteza, utilizando probabilidades para representar a transição entre estados e associando recompensas a essas transições.
- Redes Bayesianas: São utilizadas para representar e raciocinar sobre incertezas, permitindo a tomada de decisão mesmo com informações parciais ou incompletas.

Nesse sentido, os algoritmos de tomada de decisão em IA são fundamentais para o desenvolvimento de sistemas autônomos e agentes inteligentes. Eles possibilitam que esses sistemas ajam de forma adaptativa e eficiente em ambientes complexos, mesmo na presença de incertezas e grandes espaços de soluções. Cada abordagem tem suas vantagens e limitações, e a escolha do algoritmo adequado depende do problema específico, das restrições computacionais e da natureza do ambiente em que o agente atua.

2.3.3 Transparência e responsabilidade da IA

IA responsável Nós apenas tivemos uma amostra do verdadeiro impacto que a IA generativa está tendo no varejo. Mas a tecnologia está avançando incrivelmente rápido. É por isso que é extremamente importante garantir que a IA seja usada com responsabilidade. Isso significa definir proteções para adquirir, refinar e implantar dados. Também significa pensar nas operações de segurança cibernética. O gerenciamento de riscos regulatórios e de privacidade é um bom começo, e os varejistas podem dar um passo além, certificando-se de que a tecnologia que estão usando seja responsável por design.

Não demorará muito para que ter fortes recursos de IA generativa seja um requisito básico para qualquer varejista que queira acompanhar o ritmo de seus pares. Varejistas de sucesso serão aqueles que se apoiarem fortemente nessa nova e empolgante tecnologia e começarem a explorar como ela pode reinventar cada parte de seus negócios (38)

2.4 REGULAÇÃO DA IA

2.4.1 Marco regulatório da IA no Brasil

A estratégia adotada pelo Brasil para a Transformação Digital, E-Digital, foi aprovada em março de 2018, pelo Decreto n. 9.319/2018 e regulamentada pela Portaria MCTIC n. 1.556/2018, indicou-se a importância de priorizar o tema da IA em razão de suas oportunidades e impactos. Mais adiante, por meio da Portaria MCTIC no 1.122/2020, a inteligência artificial foi definida como prioridade pelo MCTI, na categoria de tecnologias habilitadoras (arts. 2º e 4º da referida portaria), em projetos de pesquisa, desenvolvimento de tecnologias e inovações, para o período de 2020 a 2023 (39).

Consoante a esse objetivo, em abril de 2021, o Ministério da Ciência, Tecnologia e Inovações (MCTI), apresentou a Estratégia Brasileira de Inteligência Artificial (EBIA) (40), com a finalidade de promover o avanço científico e estabelecer um plano de desenvolvimento para a IA no país. Em linhas gerais, o documento teve o objetivo de guiar a atuação da administração pública brasileira na definição das ações de estímulo à pesquisa e inovação no âmbito da IA, além de promover o uso ético.

Segundo o próprio documento, para a formulação da EBIA, foi feito um trabalho de benchmarking, especialmente observadas as estratégias lançadas por outros países. Nota-se que, em geral, essas estratégias apresentam tópicos em comum. Nesse sentido, a EBIA estabelece eixos verticais e transversais para definir seu mecanismo de ação. A EBIA foi alvo de severas críticas por especialistas da área, por, na visão deles, consistir em mera compilação de princípios e dados, sem indicar metas, orçamento ou planejamento de implementação (40).

Paralelamente, tramitavam no Senado Federal os Projetos de Lei n° 5.051/2019 (41), de autoria do Senador Styvenson Valetim (PODEMOS/RN), que estabelece princípios para o uso da IA no Brasil; o Projeto de Lei n° 21/2020 (41), de autoria do Deputado Eduardo Bismarck (PDT-CE), que estabelece princípios, direitos e deveres para o uso de inteligência artificial no Brasil; e o Projeto de Lei n° 872/2021 (29), de autoria do Senador Veneziano Vital do Rêgo (MDB-PB), que dispõe sobre o uso da IA.

Em 06 de julho de 2021, a Câmara dos Deputados aprovou o regime de urgência para o Projeto de Lei n° 21/2020 (29). Em 29 de setembro de 2021, o Plenário da Câmara aprovou o projeto, na forma do Substitutivo apresentado pela relatora da Comissão de Ciência e Tecnologia, Comunicação e Informática, a deputada Luísa Canziani (PTB-PR). O projeto inicial é constituído de 16 artigos, trazendo, logo no art. 2°, algumas definições importantes para a regulação da IA, dentre elas, a definição de sistema de inteligência artificial: **"o sistema baseado em processo computacional que pode, para um determinado conjunto de objetivos definidos pelo homem, fazer previsões e recomendações ou tomar decisões que influenciam ambientes reais ou virtuais"**. O PL cria ainda a figura do agente de IA, que pode ser aquele que desenvolve e implanta um sistema de IA (agente de desenvolvimento) ou o que opera (agente de operação) o sistema. O projeto cria também o relatório de impacto de IA, a ser elaborado pelos agentes de IA, com informações como a descrição da tecnologia, medidas de gerenciamento e de mitigação de riscos.

O substitutivo, aprovado pela Câmara, é mais enxuto e traz 10 artigos, com significativas alterações em relação ao seu projeto inicial. Cumpre destacar que o PL 21/20, assim como outros projetos mundo afora, encontra inspiração na recomendação sobre IA da OCDE (42). Logo no art. 2°, do texto substitutivo, observa-se uma mudança importante na definição de sistema de inteligência artificial, cujo texto passa a ser redigido da seguinte forma, in verbis:

Art. 2° Para os fins desta Lei, considera-se sistema de inteligência artificial o sistema baseado em processo computacional que, a partir de um conjunto de objetivos definidos por humanos, pode, por meio do processamento de dados e informações, aprender a perceber, interpretar e interagir com o ambiente externo, fazendo predições, recomendações, classificações ou decisões, e que utiliza técnicas como os seguintes exemplos, sem a eles se limitar: I – sistemas de aprendizagem de máquina (machine learning), incluindo aprendizagem supervisionada, não supervisionada e por reforço; II – sistemas baseados em conhecimento ou em lógica; III – abordagens estatísticas, inferência bayesiana, métodos de pesquisa e otimização.

O artigo 3° dispõe sobre os objetivos da IA no Brasil, dentre eles, a promoção do desenvolvimento econômico sustentável, inclusivo e do bem-estar; o da competitividade, da produtividade e inserção competitiva nas cadeias globais de valor; a melhoria na prestação dos serviços públicos e na implementação de políticas públicas; a promoção da pesquisa e desenvolvimento para estímulo da inovação.

O artigo 4°, por sua vez, traz os fundamentos para o desenvolvimento e aplicação da IA no Brasil.

São eles o desenvolvimento científico, tecnológico e a inovação; a livre manifestação de pensamento e livre expressão; a não discriminação, a pluralidade e o respeito às diversidades regionais; o estímulo à autorregulação por meio de códigos de conduta e guias de boas práticas; segurança, a privacidade e a proteção dos dados pessoais e da informação; a defesa e a soberania nacional; a harmonização com a LGPD (Lei n° 13.709/2018), o Marco Civil da Internet (Lei n° 12.965/2014), o Sistema Brasileiro de Defesa da Concorrência (Lei n° 12.529/2011, o CDC (Lei n° 8.078/1990) e a LAI (Lei n° 12.527/2011).

Já o artigo 5° trata dos princípios, sendo eles: finalidade benéfica para a humanidade; centralidade no ser humano (dignidade humana, privacidade, proteção de dados pessoais e direitos fundamentais); neutralidade e não discriminação; transparência e direito de informação sobre a utilização da IA (43); segurança e prevenção; inovação responsável; e disponibilidade de dados.

O texto prevê ainda, no artigo 6°, diretrizes a serem observadas pelo poder público, quando disciplinar a aplicação da IA, com destaque para:

- (i) intervenção subsidiária, pela qual deverão ser desenvolvidas regras específicas apenas quando necessário;
- (ii) atuação setorial, devendo considerar o contexto e as normas regulatórias específicas de cada setor.
- (iii) gestão baseada em risco, pela qual deve-se considerar os riscos concretos e a probabilidade de ocorrência desses riscos em comparação com potenciais benefícios sociais e econômicos e os riscos apresentados por sistemas similares sem inteligência artificial;
- (iv) participação social e interdisciplinar (baseada em evidências e precedida por consulta pública);
- (v) análise de impacto regulatório, nos termos do Decreto n° 10.411, de 2020 e Lei n° 13.874, de 2019;
- (vi) responsabilidade subjetiva; salvo disposição legal em contrário, as normas sobre responsabilidade dos agentes devem levar em consideração a efetiva participação desses agendes e os danos específicos que se deseja evitar ou remediar

No que tange à responsabilidade civil, quando o sistema de IA envolver relações de consumo (Art.5°, §3°), o texto previa que, observado o CDC, o agente responderá independentemente de culpa pela reparação dos danos causados aos consumidores e no limite de sua participação efetiva (responsabilidade objetiva). Ademais, pessoas jurídicas de direito público ou privado prestadoras de serviços públicos responderão pelos danos que seus agentes causarem a terceiros, assegurado o direito de regresso contra o responsável nos casos de dolo ou culpa (Art. 5°, §4°). Com relação à gestão baseada em risco, a proposta determina que a administração pública poderá, em casos concretos de alto risco, solicitar informações sobre as medidas de segurança e prevenção e respectivas salvaguardas (Art.5°, §2°).

De modo geral, as preocupações levantadas nas sessões semipresenciais de consulta pública e através da submissão de documentos pela da página do Senado Federal na internet convergem. Primeiramente, foram apontadas colaborações acerca da definição de IA, considerada por alguns insuficiente ou excessivamente restritiva, devendo, portanto, ser mais genérica para poder abranger também a IA autônoma, enquanto outros consideraram que uma definição restrita possui a vantagem de limitar o escopo de aplicação da lei

e evitar o excesso de regulação. Foi criticado também o modelo principiológico adotado pelo PL, por não apresentar mecanismos efetivos de implementação. Por outro lado, as associações representativas das empresas do setor tendem a apoiar a abordagem principiológica, tendo em vista sua natureza menos restritiva.

Outro ponto de destaque nos debates trazidos pelas consultas públicas é a edição de uma lei geral para a IA em oposição a regulamentações setoriais. As associações do setor privado posicionam-se pela adoção de regras setoriais para não criar gargalos ao desenvolvimento e à adoção de sistemas de IA que ainda serão criados. Fala-se ainda de uma "autorregulação regulada", pela qual haveria um estímulo à autorregulação pelo poder público, combinado com regulações setoriais para mitigar eventuais riscos.

Segundo o atual cronograma da Comissão, estão previstas mais três reuniões presenciais, nos dias 24 de novembro, 01 de dezembro, data na qual se pretende realizar a deliberação final do projeto, e 07 de dezembro, prazo fatal para a conclusão dos trabalhos. Nesta última data está prevista a apresentação da proposta final de regulação. Resta aguardar para avaliar os resultados do que se tornou este longo e cuidadoso processo de revisão do PL 21/20, ao que parece, mais aderente às necessidades e expectativas das partes interessadas na regulação dos sistemas de IA no Brasil.

2.5 CHATGPT

Segundo, (44) o ChatGPT é um agente conversacional construído sobre os modelos GPT que por sua vez são modelos de linguagem de uso geral que podem realizar uma ampla variedade de tarefas, desde criar conteúdo original até escrever código, resumir texto e extrair dados de documentos, lançados previamente pela OpenAI.

Essa série de modelos começou com o GPT-1, seguido pelo GPT-2, GPT-3 e, por fim, o GPT-3.5, que serve como base para o ChatGPT. Esses modelos utilizam uma arquitetura de rede neural chamada transformer, desenvolvida para resolver o problema de tradução automática neural.

Essa arquitetura permite prever a próxima palavra de uma frase de entrada com base em probabilidades calculadas a partir de dados de treinamento, resultando em uma frase de saída. O ChatGPT foi ajustado com a rotulagem de treinadores humanos usando Aprendizagem por reforço com feedback Humano (RLHF) para otimizar o modelo anterior (3.5) para conversas de bate-papo.

No caso específico de *fake news*, (45) conduziu um estudo para avaliar a capacidade do ChatGPT de verificar a veracidade de fatos. Os autores usaram um dataset coletado do popular site de verificação de fatos PolitiFact. O conjunto de dados contém 21.152 declarações de 2007 até 2022 incluídas que são verificadas por especialistas em uma das seis categorias: verdadeiro, quase verdadeiro, meio verdadeiro, principalmente falso, falso e "pants on fire", o último dos quais é categoria do PolitiFact para declarações que são flagrantemente falsas.

O ChatGPT categorizou as declarações corretamente em 72 dos casos, com precisão significativamente maior na identificação de afirmações verdadeiras (80) do que afirmações falsas (67). Sendo assim, embora não sendo uma ferramenta sem falhas, o ChatGPT tem certa credibilidade para classificar se informações

são falsas ou não. Além da verificação das notícias em si, é possível utilizar o ChatGPT para verificar a credibilidade de plataformas como um todo(46). desenvolveram uma metodologia para classificar a credibilidade de plataformas utilizando engenharia de prompt no ChatGPT.

Os autores desenvolveram uma estrutura para conseguir receber uma resposta bem estruturada do chat em relação à credibilidade do veículo de informação. A adaptação correspondente da estrutura para o português é mostrada na Figura 2.3 em um exemplo real com o ChatGPT.

Nesse sentido, o primeiro estudo deste trabalho conforme descrito, foi a utilização do ChatGpt para verificação de *fake news*, utilizando o ChatGPT para verificar a credibilidade de uma notícia a partir da URL do site, Tal mecanismo visava a criação de um modelo composto pela coleta e envio de dados através das bases de informações já existentes para consolidação do tipo de mensagem, com o intuito de fazer a detecção preventiva e contínua de *fake news*. Como forma de prova de conceito, esta pesquisa utilizou a ferramenta ChatGPT devido a sua capacidade de processamento em grande escala, acessos via API e a capacidade de correlacionar dados em função dos modelos de linguagem utilizadas em treinamento de aprendizado de máquina.

Além disso, teve como objetivo explorar o potencial do ChatGPT como ferramenta de auxílio na verificação de notícias, capaz de identificar rapidamente suspeitas acerca da veracidade de uma informação e de acionar novos níveis de checagem. A proposta deste estudo é realizar a verificação da credibilidade de veículos de informação a partir de uma dada notícia, enviando comandos de prompt. Na primeira fase, a principal proposta seria validar as respostas fornecidas pelo modelo generativo, as quais foram analisadas manualmente em contraste com outras fontes de referência. Do ponto de vista de segurança da informação, a proposta buscava avaliar a integridade dos dados publicados e sua relação com a possibilidade de verificação da origem da notícia e seu possível impacto considerando aspectos de confiança, já que a aceitação do conteúdo passa por uma avaliação humana de checagem.

Assim, o que pode verificar na primeira fase do trabalho foi possível identificar seus principais prós e contras, complexidade de implementação e particularidades em relação à criação e validação dos algoritmos, a primeira parte explorou o potencial do ChatGPT como uma ferramenta promissora para auxiliar na verificação de notícias. Em seguida os resultados do teste de verificação revelaram que o ChatGPT foi capaz de atribuir ratings coerentes com a verificação manual das notícias, sugerindo sua capacidade de discernir informações precisas e confiáveis. Essa descoberta é significativa, pois destacava a utilidade do ChatGPT como uma ferramenta eficaz na luta contra a propagação de desinformação e na promoção de um ambiente de informação mais confiável. Entretanto havia um problema na classificação das mensagens, pois embora o ChatGPT não deva substituir a verificação manual de notícias, sua capacidade de fornecer insights consistentes pode ser uma adição valiosa ao arsenal de ferramentas utilizadas por jornalistas, pesquisadores e leitores preocupados com a veracidade das informações. À medida que avançamos em direção a uma era digital cada vez mais complexa, o potencial do ChatGPT como uma ferramenta de auxílio confiável na verificação de notícias é uma perspectiva emocionante para explorar e desenvolver.

dessa situação, foi proposto então uma ferramente "Factual"que utilizasse de sus próprios datasets para verificação e validação de notícias falsas. Pois foi necessário uma busca de casos para realização desse teste, sendo possível perceber que a divulgação de *fake news* por sites é bem escassa quando comparada com a divulgação por posts e mensagens em redes sociais. Sendo assim, a segunda fase do trabalho focou

em investigar a validação de tais informações extraídas das redes sociais utilizando o método próprio, conforma previsto nos capítulos 4 e 5 deste trabalho.

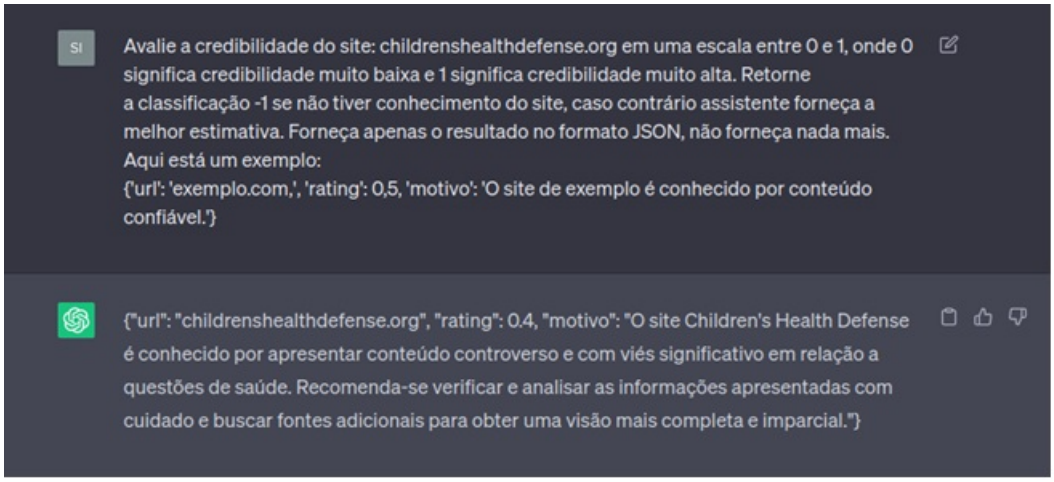


Figura 2.3: Método para classificar a credibilidade de sites de informação com o ChatGPT.
Fonte: fYang e Menczer (2023)

2.6 FERRAMENTAS DE DETECÇÃO DE *FAKE NEWS*

Existem várias ferramentas e iniciativas voltadas para a detecção de *fake news*, essas ferramentas são recursos úteis, mas a alfabetização digital e o senso crítico também são essenciais para detectar *fake news*, pois muitas vezes as informações falsas apelam para emoções ou narrativas sensacionalistas. Dentre elas podemos citar:

- Lupa - É uma agência de fact-checking brasileira que verifica a veracidade de notícias e informações divulgadas nas redes sociais e pela imprensa;
- E-Farsas - Uma plataforma que desmascara hoaxes, lendas urbanas e boatos que circulam na internet, especialmente no Brasil;
- Boatos.org - Focado em desmentir boatos e notícias falsas que circulam principalmente nas redes sociais;
- Snopes - Uma das mais antigas plataformas de verificação de fatos, muito usada em nível internacional para esclarecer rumores e *fake news*;
- Fato ou Fake - Projeto da Globo que verifica a veracidade de informações em circulação no Brasil;
- Google Fact Check Tools - Ferramenta do Google que permite procurar verificações de fatos realizadas por agências de fact-checking em todo o mundo.
- FactCheck.org - Site americano que revisa a veracidade de informações, principalmente relacionadas à política.

A verificação de notícias não é sempre preto no branco. Muitas vezes, as *fake news* misturam fatos reais com informações falsas, o que dificulta a análise automática. Além disso, a interpretação dos fatos pode variar, dependendo de contextos políticos, sociais e culturais.

2.7 PYTON

A linguagem Python (47) foi projetada com a filosofia de enfatizar a importância do esforço do programador sobre o esforço computacional. Prioriza a legibilidade do código sobre a velocidade da execução, tendo uma sintaxe concisa e clara. Além disso, conta com os recursos poderosos para implementação de algoritmos de Inteligência Artificial – IA, pois, possui frameworks desenvolvidos por terceiros, bastante adequada para efeito da validação dos resultados deste trabalho.

segundo (48) Python é uma linguagem de programação de alto nível, interpretada e de uso geral, criada por *Guido van Rossum* e lançada em 1991. Ela foi projetada com foco na legibilidade e simplicidade, o que facilita a escrita e manutenção do código. Possui uma sintaxe projetada com ênfase na clareza do código, com sua sintaxe simples permitindo que conceitos complexos sejam expressos de forma concisa, facilitando a escrita, leitura e manutenção do código. Essa característica é especialmente valiosa em ambientes colaborativos e em projetos de grande escala. Um dos pontos fortes de Python é seu vasto ecossistema de bibliotecas e frameworks. A biblioteca padrão já oferece módulos para diversas tarefas, como manipulação de arquivos, expressões regulares e operações com redes. Além disso, frameworks como Django e Flask facilitam o desenvolvimento web, enquanto bibliotecas como NumPy, Pandas, Matplotlib e TensorFlow são amplamente utilizadas em ciência de dados, análise numérica e machine learning.

o Python conta com uma das comunidades mais vibrantes e colaborativas no mundo da programação. Essa comunidade não só contribui com a criação e manutenção de pacotes e frameworks, mas também oferece suporte por meio de fóruns, tutoriais, conferências e grupos de usuários, facilitando a resolução de problemas e a evolução contínua da linguagem. É uma linguagem altamente portável, isto quer dizer que o desenvolvimento de eventuais classes ou soluções podem ser migrados para outras plataformas sem grande esforço. Outro ponto importante é da flexibilidade da linguagem utilizados em algoritmo de inteligência artificial, possuindo uma extensa biblioteca para tais fins (49).

O uso de Python como linguagem de programação neste trabalho, deve-se a vários motivos, cabendo principalmente no que se relaciona a parte de coleta de informações “busca de fontes” para confrontar com as informações geradas pelo ChatGPT, conforme demonstrados nos algoritmos utilizados e nos resultados dos capítulos 4 e 5. Além de, alguns pontos fortes dessa linguagem que influenciaram na implementação do modelo de busca de validação deste trabalho.

2.8 CARACTERÍSTICAS DOS CUDA CORES

Os **CUDA cores** (Cores de Unidade de Processamento de Dados Paralelos) são componentes essenciais das GPUs da NVIDIA. Eles permitem a execução paralela de operações, proporcionando um aumento

significativo de desempenho em tarefas que envolvem grandes volumes de dados. A seguir, são descritas as principais características dos CUDA cores.

2.9 CARACTERÍSTICAS DOS CUDA CORES

- **Processamento Paralelo:** Cada CUDA core permite a execução simultânea de múltiplas operações em diferentes dados. Esse modelo de computação paralela acelera significativamente o desempenho em tarefas de grande escala, como aprendizado de máquina e gráficos computacionais.
- **Arquitetura SIMD (Single Instruction, Multiple Data):** A arquitetura dos CUDA cores segue o modelo SIMD, onde a mesma instrução é executada em múltiplos dados ao mesmo tempo. Isso é eficiente para tarefas como multiplicação de matrizes, amplamente utilizada em algoritmos de aprendizado profundo e processamento gráfico.
- **Número de CUDA Cores:** O número de CUDA cores varia entre diferentes modelos de GPUs. Por exemplo, a NVIDIA GeForce RTX 3060 Ti possui 4.864 cores, enquanto modelos mais avançados, como a Tesla V100, possuem até 5.120 cores. Quanto maior o número de cores, maior a capacidade de processamento paralelo.
- **Alta Eficiência em Computação Paralelizada:** CUDA cores são otimizados para operações matemáticas simples e eficientes, como adições e multiplicações, essenciais para algoritmos que requerem processamento intensivo, como redes neurais artificiais.
- **Aceleração de Tarefas Específicas:** Os CUDA cores se destacam em tarefas específicas, como o treinamento de modelos de aprendizado de máquina, processamento de dados em larga escala e renderização de gráficos 3D. Isso é possível pela sua capacidade de processar operações matemáticas em grande volume de dados simultaneamente.
- **Memória e Latência:** Cada CUDA core pode acessar diferentes tipos de memória, como memória global e memória compartilhada. Essa estrutura permite um processamento eficiente de dados em larga escala, reduzindo a latência e aumentando o desempenho.
- **Desempenho Escalável:** O desempenho dos CUDA cores é escalável. Com o aumento no número de cores nas GPUs, o desempenho em tarefas paralelizadas também aumenta, permitindo a execução de operações complexas em um menor período de tempo.
- **Compatibilidade com o CUDA Toolkit:** Os CUDA cores são totalmente compatíveis com o `CUDA Toolkit`, que oferece ferramentas de desenvolvimento para escrever código otimizado em diversas linguagens como C, C++, Fortran e Python, facilitando a criação de software que aproveita o poder de computação das GPUs.
- **Eficiência Energética:** Mesmo com o aumento significativo no poder de processamento, os CUDA cores são projetados para operar de maneira eficiente em termos de consumo de energia, proporcionando um desempenho alto sem um grande aumento no consumo de energia.

- **Suporte a Frameworks de Inteligência Artificial:** CUDA cores são amplamente utilizados em frameworks de aprendizado profundo, como TensorFlow, PyTorch e Caffe, que são otimizados para tirar proveito do poder de processamento das GPUs e proporcionar aceleração no treinamento de modelos de IA.

Os **CUDA cores** desempenham um papel fundamental no aumento de desempenho em tarefas que envolvem processamento intensivo de dados. Eles são projetados para realizar operações em paralelo, oferecendo uma solução eficiente e escalável para o processamento de grandes volumes de dados, especialmente em áreas como inteligência artificial, aprendizado profundo e gráficos computacionais.

3 TRABALHOS RELACIONADOS

A detecção de *fake news* tem sido amplamente explorada devido aos seus impactos em diversas áreas, tais como saúde pública, processos eleitorais e plataformas digitais. Este capítulo apresenta diferentes abordagens propostas na literatura para enfrentar esse problema, destacando as metodologias, contribuições e limitações de cada trabalho. A seguir, serão apresentados os principais estudos relacionados ao tema, com o objetivo de contextualizar a pesquisa desenvolvida nesta dissertação. Ao final do capítulo, serão apresentadas as discussões e as considerações finais, confrontando os trabalhos analisados ao modelo proposto neste estudo.

3.1 TRABALHOS RELACIONADOS NA DETECÇÃO DE *FAKE NEWS*

O trabalho de (50) propõe o framework Event-Radar, modelado para detectar *fake news* em mídias multimodais. A abordagem combina aprendizado *multi-view* e inconsistências a nível de eventos entre texto e imagem. Para isso, o Event-Radar modela grafos de eventos, capturando relações sujeito-predicado em notícias, e utiliza distribuições Beta para ajustar a credibilidade das modalidades. Essa abordagem apresenta limitações, como a falta de exploração de relações causais entre eventos e a dependência de ferramentas externas para reconhecimento de entidades nomeadas.

O trabalho de (12) investiga o uso de LLMs, como o GPT-3.5, para detectar *fake news*. Os autores realizam um estudo empírico sobre a capacidade desses modelos em detectar desinformação, propondo uma abordagem em que os LLMs atuam como conselheiros fornecendo justificativas que complementam o desempenho de modelos menores e específicos (SLMs), como o BERT. A metodologia inclui o desenvolvimento de uma Rede de Orientação Adaptativa de Justificativas (ARG), que integra as percepções dos LLMs ao julgamento dos SLMs. Essa abordagem difere do modelo proposto nesta dissertação, uma vez que aqui o sistema verifica a veracidade das respostas e oferece justificativas contextuais como parte da classificação.

Em (51), os autores apresentam um modelo que integra múltiplas fontes e granularidades de informação, tais como conteúdo de notícias, redes sociais, gráficos de conhecimento e dados externos, para a detecção de *fake news*. A metodologia envolve a construção de dois grafos heterogêneos: (i) um para modelar características sociais; e (ii) outro para capturar o conhecimento contido nas notícias. Um módulo de raciocínio explicável é implementado para gerar explicações sobre os resultados. O trabalho enfrenta desafios, como a construção de grafos em larga escala e o risco de inclusão de ruído irrelevante em informações provenientes de múltiplas fontes.

Em um contexto voltado para línguas de poucos recursos, (16) explora o aprendizado por transferência com modelos transformadores pré-treinados, como o mBERT e o XLM-RoBERTa, ajustados para detectar *fake news*. A abordagem utiliza dados em inglês e línguas dravidianas para aprimorar a classificação em cenários com recursos limitados. O trabalho enfrenta limitações como a dependência de recursos da língua inglesa e a necessidade de mais dados anotados em línguas de baixo recurso.

O trabalho de (13) propõe um modelo para a detecção de *fake news* em múltiplos domínios, utilizando *soft-labels* para capturar características de notícias provenientes de áreas como política, entretenimento e finanças. A metodologia emprega o mecanismo Leap-GRU, que ignora palavras irrelevantes no processamento de texto, e grupos de especialistas para extrair características de cada domínio. As limitações incluem o processamento de textos longos e a dependência de especialistas, que podem introduzir vieses ou inconsistências no julgamento.

O artigo (14) apresenta uma abordagem baseada no BERT (*Bidirectional Encoder Representations from Transformers*) para a detecção de *fake news* em mídias sociais. O modelo combina o BERT com uma rede neural convolucional (CNN) de camada única, utilizando tamanhos de *kernel* variados para capturar dependências semânticas de longa distância. A arquitetura apresenta complexidade computacional, dificultando sua aplicação em cenários de larga escala ou com restrições de recursos.

Em (17), os autores propõem o framework UPFD (*User Preference-aware Fake News Detection*), que utiliza preferências endógenas dos usuários e o contexto exógeno de propagação de notícias nas redes sociais para detectar *fake news*. A metodologia combina representações de engajamento dos usuários, extraídas de seus históricos de postagens, com um grafo de propagação de notícias baseado em cascatas de compartilhamento, utilizando Redes Neurais Gráficas (GNNs). O UPFD depende de dados históricos e enfrenta desafios relacionados à escalabilidade devido à complexidade da integração de múltiplas fontes de informação.

O trabalho de (15) adota uma abordagem multimodal para a detecção de *fake news*, utilizando textos e imagens. Os autores empregam o dataset Fakeddit, com seis categorias de notícias, e uma arquitetura baseada em Redes Neurais Convolucionais (CNNs) para combinar texto e imagem na classificação de *fake news*. O desequilíbrio do dataset utilizado afeta o desempenho do modelo em categorias menos representadas, como conteúdos impostores.

O trabalho de (52) propõe um modelo de segurança cibernética baseado em redes neurais profundas para detectar e prevenir ataques relacionados a desinformação em plataformas digitais. Os autores exploram a interseção entre segurança cibernética e desinformação, com foco em técnicas de defesa baseadas em IA para melhorar a resistência a ataques de *fake news*.

Em (53), os pesquisadores investigam como as ameaças cibernéticas podem ser mitigadas por meio de técnicas de aprendizado de máquina que identificam padrões de manipulação em notícias. O estudo também aborda como ataques cibernéticos podem explorar vulnerabilidades em sistemas de verificação de notícias.

O trabalho de (54) discute as implicações da segurança cibernética na proteção contra a manipulação de notícias em ambientes digitais, analisando as ferramentas e as práticas de segurança utilizadas para detectar bots e outras fontes de desinformação.

Em (55), é abordado como as ameaças cibernéticas podem comprometer a integridade dos sistemas de detecção de *fake news*, propondo um sistema de defesa adaptativo que utiliza criptografia e blockchain para assegurar a autenticidade das fontes de notícias verificadas.

O artigo (56) examina como os ataques cibernéticos podem ser empregados para disseminar *fake news*, e propõe uma solução baseada em Inteligência Artificial (IA) para detectar e prevenir tais ataques, focando

em sistemas que protejam dados sensíveis em plataformas de mídia social.

Em (57), os autores discutem estratégias de defesa em redes sociais contra ataques de *fake news* e como a segurança cibernética pode ser aplicada para minimizar o impacto dessas notícias manipuladas nas plataformas digitais, com ênfase na privacidade e proteção de dados dos usuários.

O estudo de (58) analisa as ameaças digitais associadas à disseminação de desinformação e como as soluções de segurança cibernética podem ser integradas com sistemas de monitoramento para garantir que notícias falsas não comprometam a confiança nas plataformas digitais.

Em (59), é explorado como técnicas de segurança cibernética podem validar a veracidade das informações postadas em plataformas digitais, utilizando algoritmos de hash e criptografia para garantir a autenticidade das fontes.

Por fim o artigo (60) aborda a privacidade dos dados em sistemas de detecção de *fake news*, propondo um modelo de segurança cibernética que utiliza técnicas de anonimização para proteger os usuários enquanto verifica a veracidade das notícias disseminadas.

3.2 DISCUSSÃO SOBRE OS TRABALHOS RELACIONADOS

Os trabalhos relacionados analisados neste capítulo apresentam diferentes abordagens e metodologias para a detecção de *fake news*. Para contextualizar as contribuições da solução proposta nesta dissertação, a Tabela 3.1 apresenta uma analise comparativa entre os trabalhos citados e a solução proposta. A Tabela 3.1 resume as características positivas da solução proposta e a compara com os trabalhos citados, permitindo uma análise dos avanços proporcionados pela solução proposta. Três aspectos centrais foram selecionados para essa comparação: (i) a presença de uma categorização abrangente das diferentes formas de desinformação; (ii) a existência de um mecanismo dinâmico de cálculo de confiança para as previsões realizadas; e (iii) a escalabilidade das soluções.

Tabela 3.1: Comparativo entre os trabalhos relacionados e a solução proposta.

Trabalhos Relacionados	**Categorização Abrangente de Desinformação**	**Mecanismo Dinâmico de Cálculo de Confiança**	**Escalável**
Ma et al. (2024) (50)	Não	Não	Sim
Hu et al. (2024) (12)	Não	Não	Sim
Han et al. (2024) (51)	Não	Não	Sim
Raja et al. (2023) (16)	Não	Não	Sim
Wang et al. (2023) (13)	Sim	Sim	Não
Kaliyar et al. (2021) (14)	Não	Não	Não
Dou et al. (2021)(17)	Não	Não	Sim
Segura et al. (2022) (15)	Não	Não	Sim
Solução Proposta	Sim	Sim	Sim

Ao observar a Tabela 3.1, evidencia-se que a solução proposta se destaca ao abordar três pontos principais que não foram explorados pelos demais trabalhos. Enquanto o trabalho (50) introduz o framework

Event-Radar para detecção multimodal de *fake news*, ele não contempla uma categorização abrangente das diferentes formas de desinformação nem possui mecanismos para calcular dinamicamente a confiança nas previsões realizadas. Abordagens como as de (12) e (51) focam na integração de múltiplas fontes e granularidades, mas não oferecem soluções para a confiança das previsões ou para a categorização detalhada de desinformação.

O trabalho de (16) apresenta avanços no uso de aprendizado por transferência em línguas de poucos recursos e destaca a escalabilidade como um de seus pontos fortes. No entanto, não aborda a questão da categorização de desinformação nem implementa mecanismos de cálculo dinâmico de confiança. Já o modelo de (13) introduz *soft-labels* e Leap-GRU para detecção em múltiplos domínios, mas carece de escalabilidade e de soluções que aumentem a interpretabilidade e a confiabilidade dos resultados.

Por outro lado, abordagens como as de (14), (15), e (17) exploram diferentes combinações de métodos baseados em aprendizado profundo e análises multimodais. No entanto, não propõem categorizações detalhadas de desinformação nem fornecem suporte para mecanismos de confiança dinâmicos, o que limita a transparência e a aplicabilidade de seus modelos em contextos diversos.

4 FACTUAL - ANÁLISE DE FAKE NEWS COM CONFIANÇA E CREDIBILIDADE UTILIZANDO IA E LLMS

Neste capítulo, será apresentado o FACTUAL (***F**ake News **A**nalysis with **C**onfidence and **T**rust **U**sing **AI** and **LLM***), um sistema baseado em LLM desenvolvido para detectar e mitigar a disseminação de *fake news*. O FACTUAL foi projetado para fornecer uma análise confiável de conteúdos potencialmente falsos, permitindo que os usuários tomem decisões informadas de forma eficaz e em tempo real. O objetivo principal do FACTUAL é oferecer uma estrutura eficiente e escalável para o treinamento e uso de LLM, abordando os desafios relacionados à detecção de *fake news*, ao mesmo tempo em que possibilita a identificação e mitigação da desinformação de maneira ágil e precisa. Neste capítulo, será apresentado o funcionamento do FACTUAL, suas funcionalidades e as estratégias adotadas para alcançar eficiência, destacando como a solução se posiciona diante dos desafios relacionados à disseminação de *fake news*.

4.1 VISÃO GERAL DO FUNCIONAMENTO DO FACTUAL

A Figura 4.1 apresenta o modelo conceitual da arquitetura do FACTUAL. O FACTUAL é composto por quatro componentes principais: (i) a camada de banco de dados; (ii) os modelos de IA; (iii) o middleware; e (iv) os frameworks de execução.

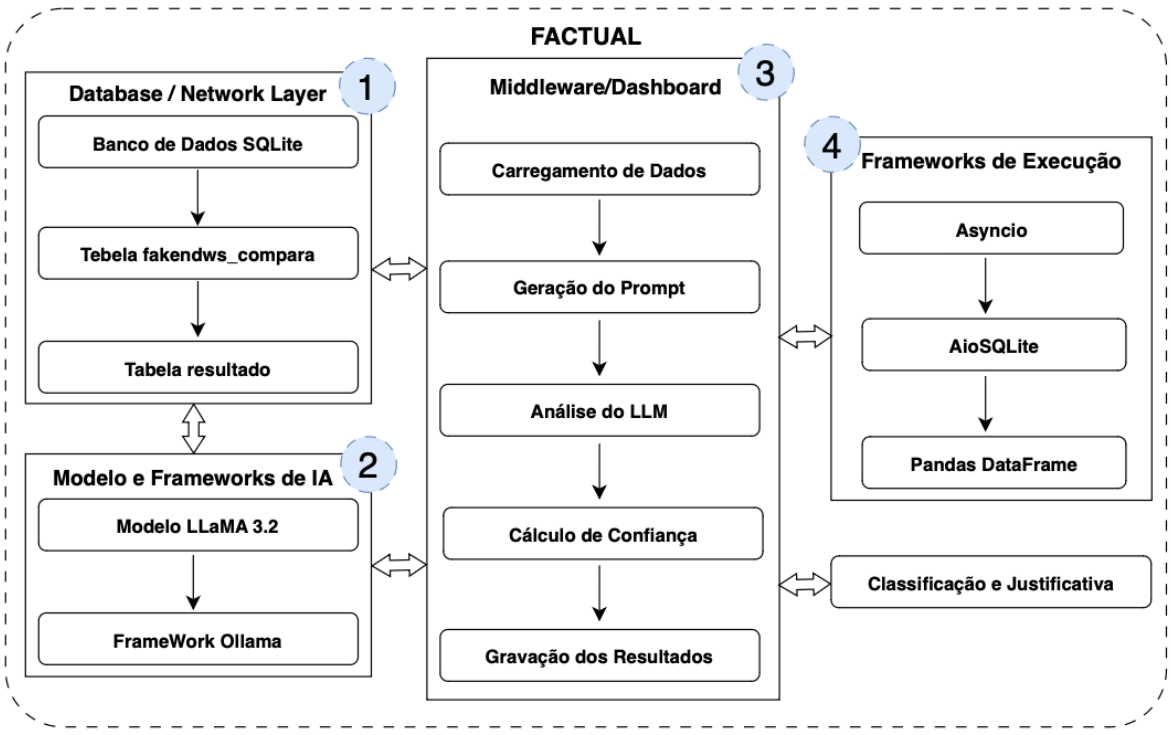


Figura 4.1: Modelo conceitual/Arquitetura
Fonte: autoria própria

A Camada de Banco de Dados (Rótulo 1, Figura 4.1) é responsável pelo armazenamento e gerenciamento dos dados utilizados no processo de detecção. Essa camada no FACTUAL utiliza um banco de dados SQLite para armazenar os registros relacionados a *fake news* que mantém informações comparativas, em que são gravadas as saídas da análise realizada pelo FACTUAL. Já a Camada de Modelos e Frameworks de IA (Rótulo 2, Figura 4.1) concentra o núcleo da análise do FACTUAL, utilizando o modelo LLaMA 3.2 para processar grandes volumes de dados textuais. O framework Ollama é utilizado para integrar o modelo ao FACTUAL, permitindo a análise e a identificação de padrões característicos de desinformação. O uso de um modelo de linguagem avançado como o LLaMA oferece ao FACTUAL a capacidade de compreender nuances e ambiguidade no conteúdo textual, aumentando a precisão da detecção de *fake news* em cenários complexos.

O Middleware (Rótulo 3, Figura 4.1) desempenha um papel central na integração entre a camada de dados e os modelos de IA. Esse componente é responsável por tarefas como o carregamento de dados do banco de dados, a geração de prompts apropriados para o modelo de IA, a análise dos resultados obtidos e o cálculo de confiança, o que permite ao sistema avaliar a veracidade das informações analisadas. Além disso, o middleware grava os resultados no banco de dados para posterior consulta, fechando o ciclo de análise do FACTUAL. por fim, a camada de Frameworks de Execução (Rótulo 4, Figura 4.1) é desenvolvida para otimizar o desempenho do FACTUAL, utilizando frameworks assíncronos como Asyncio e AioSQLite para garantir a execução paralela das tarefas. Esses frameworks permitem que o FACTUAL processe múltiplos pedidos de análise simultaneamente, sem sobrecarregar o sistema, garantindo eficiência e escalabilidade. Os resultados são organizados em Pandas DataFrames, facilitando a visualização e a classificação dos dados, além de fornecer justificativas claras para as conclusões sobre a veracidade das informações analisadas.

4.2 CAMADA DE BANCO DE DADOS NO FACTUAL

A camada de banco de dados do FACTUAL é responsável pela estruturação e armazenamento das informações necessárias para o processo de detecção de *fake news*. O sistema utiliza um banco de dados SQLite, escolhido pela sua leveza e integração com outras camadas do sistema. Dois componentes principais compõem essa camada: (i) a tabela “fakenews_compara”, Figura 4.2(a); e (ii) a tabela de resultados, Figura 4.2(b). A tabela “fakenews_compara” armazena os dados de entrada, que são informações textuais e outras variáveis coletadas de fontes que potencialmente disseminam *fake news*. Essa tabela mantém um registro das informações que serão analisadas pelo modelo de linguagem, proporcionando a base para o início do processo de detecção.

Após a análise realizada pelos modelos, os resultados são armazenados na tabela de resultados, que contém as saídas geradas pelo sistema após a detecção de *fake news*. Nessa tabela, são gravados dados como a classificação das notícias (verdadeiras ou falsas), as justificativas fornecidas pelos modelos e o nível de confiança atribuído a cada classificação. Esses registros são utilizados para a posterior visualização e análise, garantindo que o sistema mantenha um histórico das avaliações e suas justificativas.

A escolha por uma base de dados SQLite está atrelada à necessidade de um sistema leve que permita

	id	title	text	origin	url	label	publisher_name	publisher_site
1	1066	TV alemã exibiu elefantes ao falar sobre queimadas na Amazônia	TV alemã exibiu elefantes ao falar sobre queimadas na Amazônia	@direitadaopressao	https://politica.estadao.com.br/blogs/estadao-verifica/tv-alema-exibiu-elefantes-ao-falar-sobre-queimadas-na-amazonia/	0	Política | Notícias do governo, STF e Congresso Estadão	politica.estadao.com.br
2	101	Na Câmara, ministro da Educação erra sobre pesquisa científica e vagas em creches | Agência Lupa	A gente vê que a pré-escola, os aninhos antes da 1ª série, a gente ainda não universalizou	Abraham Weintraub	https://piaui.folha.uol.com.br/lupa/2019/05/17/camara-weintraub-educacao/	0	UOL	piaui.folha.uol.com.br
3	167	Na Câmara, ministro da Educação erra sobre pesquisa científica e vagas em creches | Agência Lupa	Quando a gente olha pro começo do [ensino] fundamental, a gente vê que, praticamente, já foi universalizado	Abraham Weintraub	https://piaui.folha.uol.com.br/lupa/2019/05/17/camara-weintraub-educacao/	0	UOL	piaui.folha.uol.com.br
4	312	Na Câmara, ministro da Educação erra sobre pesquisa científica e vagas em creches | Agência Lupa	O que que a gente vê no ensino médio? Ele não está universalizado	Abraham Weintraub	https://piaui.folha.uol.com.br/lupa/2019/05/17/camara-weintraub-educacao/	0	UOL	piaui.folha.uol.com.br
5	453	Weintraub erra sobre ranking de universidades, gastos do MEC e criação da aspirina	O analfabeto hoje é uma pessoa, de um modo geral, estatisticamente falando, velha	Abraham Weintraub	https://lupa.uol.com.br/jornalismo/2019/08/03/weintraub-universidade-aspirina	0	Lupa - UOL	lupa.uol.com.br
6	461	Segurança, renda per capita, educação: erros do ministro Weintraub antes de assumir o MEC	Metade dos estupros cometidos no Brasil são contra criança. (...) Não é o pai, nem o padrasto. É o amigo [quem mais comete estupro contra crianças]	Abraham Weintraub	https://lupa.uol.com.br/jornalismo/2019/04/10/weintraub-erros-educacao-mec	0	Lupa - UOL	lupa.uol.com.br
7	969	Weintraub erra sobre ranking de universidades, gastos do MEC e criação da aspirina	80% (...) das pessoas que fazem o ensino superior no Brasil estão em instituições privadas	Abraham Weintraub	https://lupa.uol.com.br/jornalismo/2019/08/03/weintraub-universidade-aspirina	0	Lupa - UOL	lupa.uol.com.br
8	627	Os dados equivocados usados na defesa do senador afastado Aécio Neves	Fui afastado do cargo para o qual fui eleito por mais de 7 milhões de mineiros	Aécio Neves	https://lupa.uol.com.br/jornalismo/2017/05/26/aecio-neves-defesa-jbs	0	Lupa - UOL	lupa.uol.com.br
9	257	Na CBN, Aldo Rebelo exagera ao falar sobre desemprego e agricultura	Quando [a Câmara] renovou muito [sua composição, o índice] foi 60%. Quando renovou pouco, 40%	Aldo Rebelo	https://lupa.uol.com.br/jornalismo/2018/05/31/aldo-rebelo-cbn	0	Lupa - UOL	lupa.uol.com.br
10	343	Molon exagera gastos de propaganda da Prefeitura do Rio no ano passado	O asfalto no Elevado no Joá durou 10 dias	Alessandro Molon	https://lupa.uol.com.br/jornalismo/2016/09/21/molon-exagera-gastos-de-publicidade-da-prefeitura-do-rio-em-2015	0	Lupa - UOL	lupa.uol.com.br

(a) Tabela comparação.

	id	title	text	url	publisher_name	publisher_site	original_label	response_llm	confidence	predicted_label
1	25577	do que "calote de Vieira" no Novo Banco. Confirma-se?	menos do que "calote de Vieira" no Novo Banco. Confirma-se?	rsi-dos-ciganos-custa-menos-do-que-calote-de-vieira-no-novo-banco-confirma-se	Poligrafo - SAPO	poligrafo.sapo.pt	0	Em resumo, sem mais informações ou evidências concretas, não há razão para confiar na veracidade da afirmação feita por Mayan sobre os custos do "RSI dos ciganos" e do "calote de Vieira" no Novo Banco.	95.0	-1
2	24639	Crivella toma posse usando dados equivocados sobre eleição e PIB	[Queremos priorizar] Nossas mais de 1.500 unidades educacionais	https://lupa.uol.com.br/jornalismo/2017/01/03/crivella-posse	Lupa - UOL	lupa.uol.com.br	0	haja alguma dúvida sobre a precisão das informações apresentadas, há suficientes evidências para sugerir que o evento descrito no texto pode ter ocorrido, mas é importante abordar as alegações com cautela e buscar mais informações antes de concluir.	95.0	0
3	26692	Imigrantes contribuem mais do que aquilo que recebem da Segurança Social?	Imigrantes contribuem mais do que aquilo que recebem da Segurança Social?	https://poligrafo.sapo.pt/fact-check/imigrantes-contribuem-mais-do-que-aquilo-que-recebem-da-seguranca-social	Poligrafo - SAPO	poligrafo.sapo.pt	0	evidências concretas e à ambiguidade da pergunta. Portanto, , indicando que há uma possibilidade de que a afirmação seja baseada em suposições ou estereótipos, mas não há evidências suficientes para confirmar ou negar completamente.	95.0	-1
4	24231	saiba quais dados eram falsos no debate da Band	R$ 600 [do Auxílio Brasil] não está na LDO mandada para o Congresso [para 2023]"	https://lupa.uol.com.br/jornalismo/2022/08/28/checagem-debate-band-presidenciaveis-2022	Lupa - UOL	lupa.uol.com.br	0	Em resumo, considerando a fonte confiável e a análise detalhada dos dados, , indicando que o texto é uma notícia de jornalismo investigativo bem fundamentada.	95.0	0
5	24869	Marta diz que não tem 'um tostão', mas arrecadou R$ 2,4 milhões	Eu votei pela perda de direitos políticos	https://lupa.uol.com.br/jornalismo/2016/09/07/marta-diz-que-nao-tem-um-tostao-mas-arrecadou-r-24-milhoes-em-3-semanas	Lupa - UOL	lupa.uol.com.br	0	comprove que Marta tenha arrecadado esse valor em 3 semanas. Diante desses pontos, acredito que o texto seja uma notícia falsa ou exagerada. A falta de evidências adicionais e a ambiguidade da data de publicação levam-me a essa conclusão.	95.0	-1
6	25557	Faro espera há 20 anos pela construção do novo hospital?	Faro espera há 20 anos pela construção do novo hospital?	https://poligrafo.sapo.pt/fact-check/faro-espera-ha-20-anos-pela-construcao-do-novo-hospital	Poligrafo - SAPO	poligrafo.sapo.pt	0	que não há nenhuma evidência de que Faro esteja esperando há 20 anos pela construção do novo hospital. Portanto, considerando a falta de evidências concretas e a incoerência com os fatos reais, concordo plenamente com a classificação original como "Fake news".	95.0	1
7	26001	Cristina Rodrigues defende fim das touradas mas vai ser assessora do Chega?	Cristina Rodrigues defende fim das touradas mas vai ser assessora do Chega?	https://poligrafo.sapo.pt/fact-check/cristina-rodrigues-defende-fim-das-touradas-mas-vai-ser-assessora-do-chega	Poligrafo - SAPO	poligrafo.sapo.pt	0	declarações de imprensa ou press-releases. Em resumo, embora a primeira parte do texto seja uma declaração feita por Cristina Rodrigues, a segunda parte do texto não é suportada por evidências concretas e parece ser uma informação falsa. Por isso, eu classifico o texto como Fake news com 0% de confiança.	95.0	1
8	25835	filho de Carlos César vai tornar-se líder parlamentar do PS Açores?	"Tudo em família!" O filho de Carlos César vai tornar-se líder parlamentar do PS Açores?	https://poligrafo.sapo.pt/fact-check/tudo-em-familia-o-filho-de-carlos-cesar-vai-tornar-se-lider-parlamentar-do-ps-acores	Poligrafo - SAPO	poligrafo.sapo.pt	0	evidências disponíveis sugerem que o texto é mais uma notícia sensacionalista do que um relato factual. Portanto, concordo que a classificação original seja incorreta e que o texto seja considerado "fake news".	95.0	1
9	24958	dados sobre vacinação e testagem; veja checagem	"Nunca, em tão pouco tempo, tivemos vacinas eficazes para combater uma doença viral como a Covid"	https://lupa.uol.com.br/jornalismo/2021/05/21/cpi-da-covid-queiroga-vacina	Lupa - UOL	lupa.uol.com.br	0	baseada em uma análise rigorosa das evidências disponíveis. A afirmação feita por Queiroga é claramente uma distorção da realidade, e sua classificação como "fato" sugere uma falta de crítica e análise cuidadosa dos dados.	95.0	1

(b) Tabela de Resultados.

Figura 4.2: Estrutura de tabelas do banco de dados do FACTUAL.

o rápido acesso aos dados e que lide com operações de escrita e leitura simultâneas, sem comprometer o desempenho. Essa solução também facilita a integração com frameworks assíncronos, como o AioSQLite, que permitem o processamento de múltiplas requisições simultâneas, tornando o FACTUAL escalável e apto a processar grandes volumes de dados de maneira eficiente.

4.2.1 Processo de Classificação de Notícias na Camada de Modelos e Frameworks de IA

A camada de modelos e frameworks de IA é responsável pelo processamento e análise das informações textuais no FACTUAL. O principal componente dessa camada é o modelo de linguagem LLaMA 3.2, que realiza a classificação das notícias enviadas pelo usuário, determinando se são verdadeiras ou falsas. O Algoritmo 1 demonstra como a detecção de *fake news* é realizada por meio da análise textual, identificando padrões no conteúdo. O primeiro passo no processo é a inicialização do FACTUAL, em que o modelo LLaMA 3.2 e o framework Ollama são carregados e conectados. Essa etapa é essencial para preparar o ambiente de execução, garantindo que o sistema esteja pronto para processar as notícias. Uma vez configurado, o FACTUAL pode começar a receber as notícias a serem analisadas.

O código a seguir implementa um sistema automatizado para a análise e classificação de notícias, com

o objetivo de identificar *fake news* utilizando o modelo de Inteligência Artificial LLaMA 3.2. O processo é dividido em várias etapas, cada uma representada por uma `procedure` responsável por realizar uma tarefa específica, como a análise de notícias, cálculo de confiança e classificação das notícias. Conforme demonstrado no Algoritmo 1

Algoritmo 1 Camada de Modelos e Frameworks de IA no Sistema FACTUAL

```
1:  procedure INICIAR SISTEMA FACTUAL
2:      Iniciar sistema
3:      Carregar modelo LLaMA 3.2
4:      Carregar framework Ollama
5:      Conectar LLaMA 3.2 ao framework Ollama
6:  end procedure
7:  procedure ANALISAR NOTÍCIAS(noticias)
8:      for all noticia in noticias do
9:          resultado_analise ← Ollama.enviar_prompt_para_modelo(LLaMA, noticia)
10:         probabilidade_fake ← resultado_analise["probabilidade_fake_news"]
11:         nivel_confianca ← calcular_nivel_confianca(resultado_analise)
12:         if probabilidade_fake > limiar_fake_news then
13:             classificar_como_fake(noticia, nivel_confianca)
14:         else
15:             classificar_como_veridica(noticia, nivel_confianca)
16:         end if
17:         armazenar_resultado(noticia, probabilidade_fake, nivel_confianca)
18:     end for
19: end procedure
20: procedure CALCULAR NÍVEL DE CONFIANÇA(resultado_analise)
21:     confianca ← resultado_analise["confianca_total"]
22:     return confianca
23: end procedure
24: procedure CLASSIFICAR COMO FAKE(noticia, nivel_confianca)
25:     registrar_no_sistema(noticia, "fake", nivel_confianca)
26:     exibir_mensagem("Notícia classificada como fake. Nível de confiança: "+ nivel_confianca)
27: end procedure
28: procedure CLASSIFICAR COMO VERÍDICA(noticia, nivel_confianca)
29:     registrar_no_sistema(noticia, "verídica", nivel_confianca)
30:     exibir_mensagem("Notícia classificada como verídica. Nível de confiança: "+ nivel_confianca)
31: end procedure
32: procedure ENVIAR PROMPT PARA MODELO (FRAMEWORK OLLAMA)(modelo, prompt)
33:     resposta ← modelo.processar(prompt)
34:     return resposta
35: end procedure
```

4.2.1.1 Explicação do Algoritmo 1 de Análise e Classificação de Notícias para Detecção de *Fake News*

Abaixo está o código do sistema, que é estruturado em diversas `procedures`, cada uma com uma função essencial para o processo de análise de notícias. Essas `procedures` trabalham de maneira cola-

borativa para alcançar o objetivo principal do sistema:

1. **INICIAR SISTEMA FACTUAL:** Esta `procedure` é responsável por iniciar o sistema. Ela carrega o modelo de IA LLaMA 3.2 e o framework Ollama, essenciais para o funcionamento do sistema. Além disso, a `procedure` também estabelece a conexão entre o modelo e o framework, garantindo que o modelo de IA possa processar as notícias posteriormente. Esse passo é crucial para preparar a infraestrutura de análise de notícias.

2. **ANALISAR NOTÍCIAS:** A `procedure` recebe uma lista de notícias e começa a processá-las uma por uma. Para cada notícia, o texto é enviado para o modelo de IA usando a `procedure` ENVIAR PROMPT PARA MODELO. O modelo, então, analisa o conteúdo da notícia e calcula a probabilidade de a notícia ser falsa *fake news*. Essa análise é seguida pelo cálculo do nível de confiança, que indica o quão preciso é o modelo ao identificar a veracidade da notícia. Isso envolve passar a notícia pelo modelo LLaMA 3.2, que é treinado para detectar padrões e avaliar a veracidade da informação.

3. **CALCULAR NÍVEL DE CONFIANÇA:** Após o modelo realizar a análise, essa `procedure` retorna o nível de confiança da previsão feita. O nível de confiança é uma métrica que indica a certeza do modelo em relação à classificação da notícia. Ele pode ser usado para decidir se a notícia será considerada falsa ou verdadeira. Um nível de confiança mais alto significa que o modelo tem mais certeza em sua classificação, enquanto um nível de confiança baixo sugere uma maior incerteza.

4. **CLASSIFICAR COMO FAKE:** Caso a probabilidade de a notícia ser falsa seja maior que um valor preestabelecido (o limiar de *fake news*), essa `procedure` é executada. Ela registra a notícia como *fake news* no sistema e exibe uma mensagem indicando que a notícia foi classificada como falsa, junto com o nível de confiança calculado. Isso ajuda a informar ao usuário ou sistema de monitoramento que a notícia em questão é altamente suspeita de ser falsa.

5. **CLASSIFICAR COMO VERÍDICA:** Se a análise do modelo indicar que a notícia é verdadeira, essa `procedure` é acionada. Ela registra a notícia como verídica no sistema e exibe uma mensagem indicando que a notícia foi classificada como verdadeira, também acompanhada do nível de confiança. Essa classificação ajuda a validar informações que o modelo considera confiáveis.

6. **ENVIAR PROMPT PARA MODELO:** Essa `procedure` é responsável por enviar o texto da notícia para o modelo LLaMA 3.2, o qual processa o conteúdo e retorna uma análise. A análise inclui a probabilidade de *fake news* (quanto é provável que a notícia seja falsa) e o nível de confiança, que representa a precisão da análise realizada pelo modelo. Essa comunicação entre o framework Ollama e o modelo LLaMA é fundamental para o funcionamento do sistema de análise e classificação de notícias.

Essas `procedures` formam o núcleo do algoritmo, trabalhando de maneira sequencial e colaborativa para avaliar e classificar as notícias com base na probabilidade de serem *fake news*. O sistema ajuda a automatizar o processo de verificação de notícias, tornando-o mais rápido e eficiente, especialmente quando lidamos com grandes volumes de informações.

O código modulariza a análise de notícias em diferentes `procedures`, garantindo organização e separação de responsabilidades. O fluxo principal ocorre em ANALISAR NOTÍCIAS, que chama outras `procedures` para processar cada notícia, avaliar sua veracidade e armazenar os resultados. O uso do modelo LLaMA 3.2 integrado ao framework Ollama permite a análise automatizada de *fake news* com base em inteligência artificial.

O processo de análise de notícias ocorre de maneira iterativa, passando por cada notícia individualmente. O texto da notícia é enviado ao modelo LLaMA 3.2, por meio do framework Ollama, que processa o conteúdo e gera uma resposta contendo a probabilidade de que a notícia seja falsa. A partir dessa resposta, o sistema calcula o nível de confiança da análise, utilizando as informações fornecidas pelo modelo para gerar uma métrica que reflete a certeza do sistema em sua classificação.

O cálculo do nível de confiança no FACTUAL é baseado na análise do conteúdo textual retornado pelo modelo LLaMA 3.2. Essa análise utiliza uma abordagem heurística que considera palavras-chave e expressões presentes na resposta gerada pelo modelo. A função responsável (Linha 20, Algoritmo 1) por esse cálculo ajusta dinamicamente o nível de confiança, avaliando tanto o tom da resposta quanto a presença de informações que indiquem certeza ou incerteza. Conforme demonstrada na Tabela 4.1

Tabela 4.1: Tabela de Ajustes no Nível de Confiança

Passo	**Condição**	**Ação/Resultado**
1	Confiança Padrão Inicial	O sistema começa com um nível de confiança de 75%.
2	Aumento de Confiança por Palavras Indicativas de Certeza	Se a resposta contém palavras como *"claramente"*, *"indiscutivelmente"*, *"sem dúvida"*, o nível de confiança aumenta para 95%.
3	Redução de Confiança por Palavras Indicativas de Incerteza	Se o texto contém palavras como *"não há evidências suficientes"*, *"incerteza"*, o nível de confiança é reduzido para 60%.
4	Verificação de Fontes Confiáveis	A presença de fontes confiáveis aumenta a confiança em 10 pontos; fontes suspeitas reduzem a confiança em 10 pontos.
5	Garantia de Limites	O nível de confiança final é garantido para permanecer entre 0% e 100%.

Com base na probabilidade retornada pelo modelo LLaMA 3.2, a notícia é classificada como fake, verídica ou indeterminada. O sistema compara a probabilidade de ser fake com um limiar previamente definido. Se a probabilidade ultrapassar esse valor, a notícia é marcada como fake; caso contrário, ela é classificada como verídica ou indeterminada. Além da classificação, o sistema armazena o resultado da análise, o nível de confiança associado à decisão e gera uma justificativa clara ao usuário. O framework Ollama facilita essa interação, fornecendo a interface para o envio dos prompts ao modelo e a recepção dos resultados. Esses resultados, que incluem a classificação e justificativa, são enfileirados em uma tarefa assíncrona, permitindo que o sistema processe múltiplas requisições de forma eficiente.

4.3 FLUXO DE PROCESSAMENTO E INTEGRAÇÃO DO MIDDLEWARE/DASHBOARD NO FACTUAL

A camada de Middleware/Dashboard do FACTUAL gerencia o fluxo de dados entre as diferentes partes do sistema, facilitando a interação entre o usuário, o modelo de IA e o banco de dados. A principal função dessa camada é carregar os dados, gerar prompts para o modelo LLaMA 3.2, processar os resultados e retornar as respostas ao usuário. O diagrama de sequência apresentado na Figura 4.3 apresenta o fluxo de processamento que ocorre dentro do FACTUAL. O processo começa com o usuário submetendo uma notícia, que é então carregada da tabela "fakenews_compara" no banco de dados SQLite. As notícias incluem informações como título, texto, URL e editor. Essas informações são organizadas e preparadas para o envio ao modelo LLaMA 3.2 por meio da geração de prompts. O middleware organiza e estrutura essas informações, gerando o prompt que será enviado ao modelo de IA para análise.

Após a geração do prompt, o modelo LLaMA 3.2 realiza a classificação da notícia como verdadeira (fato) ou falsa (fake) ou indeterminado e gera uma justificativa para a classificação realizada. Essa resposta é então enfileirada em uma tarefa assíncrona, utilizando o framework Asyncio. A capacidade de processar múltiplas tarefas de forma assíncrona é fundamental para que o FACTUAL consiga lidar com um fluxo constante de requisições, conforme apresentado na Figura 4.3.

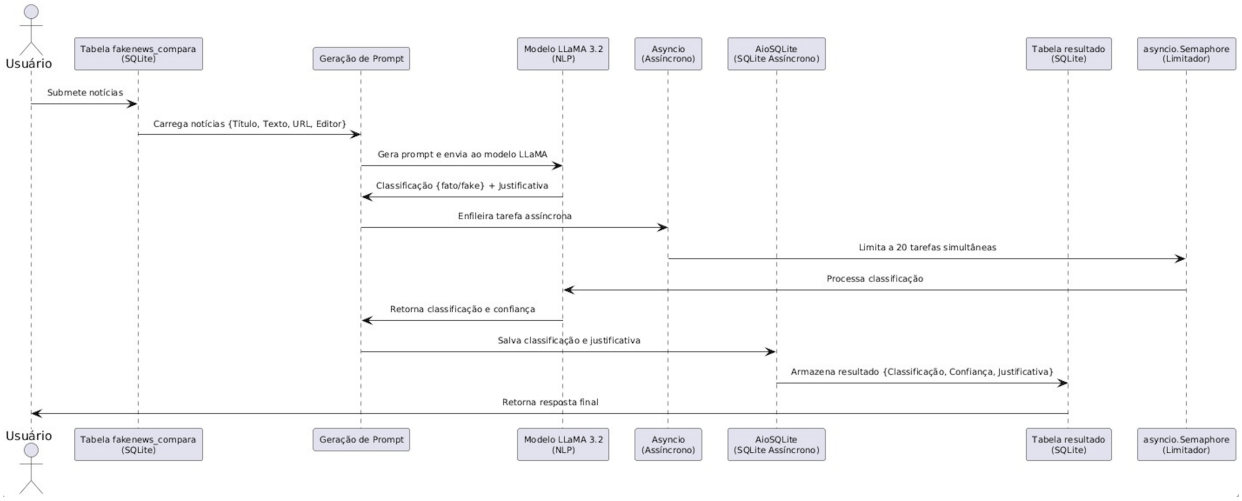


Figura 4.3: Diagrama de sequência para classificação de notícias

A seguir, o middleware recupera a classificação gerada pelo modelo, juntamente com a confiança (i.e., probabilidade associada à predição), e armazena esses dados na tabela de resultados, retornando a resposta final ao usuário. O sistema faz uso do AioSQLite, uma extensão assíncrona para o SQLite, que habilita operações de I/O não bloqueantes, fundamentais para otimizar a latência e throughput em ambientes com alta concorrência. Essa abordagem permite que múltiplas requisições de leitura e escrita no banco de dados sejam processadas simultaneamente, maximizando a eficiência ao lidar com grandes volumes de dados.

A tabela de resultados no banco de dados armazena informações, tais como a classe prevista, a confiança associada à predição, e uma justificativa gerada pelo modelo. O uso de índices no banco de dados é configurado para otimizar a consulta desses dados, garantindo que o histórico de análises possa ser acessado rapidamente, mesmo em cenários de escalabilidade. Além disso, o sistema implementa um asyn-

cio.Semaphore, um semáforo assíncrono que limita o número de corrotinas concorrentes que acessam o banco de dados, prevenindo sobrecargas no sistema e controlando o uso dos recursos computacionais de maneira eficiente.

4.4 EXECUÇÃO ASSÍNCRONA E ESCALABILIDADE NA CAMADA DE FRAMEWORKS DE EXECUÇÃO DO FACTUAL

A Camada de Frameworks de Execução do FACTUAL é responsável por otimizar o processamento e garantir a escalabilidade do sistema, gerenciando as operações de forma assíncrona. Esta camada utiliza frameworks como Asyncio e AioSQLite para processar múltiplas tarefas simultaneamente, sem sobrecarregar o sistema, e lidar com grandes volumes de dados que é uma das característica desta pesquia.

O Asyncio é um framework de execução assíncrona que permite o processamento de várias requisições em paralelo. No FACTUAL, o Asyncio é utilizado para gerenciar a execução de tarefas simultâneas, conforme apresentado no diagrama da Figura 4.3. Isso é necessário para garantir que o modelo de IA possa analisar diversas notícias de maneira eficiente, mesmo em momentos de alta demanda.

Além de processar múltiplas tarefas em paralelo, a camada de execução também faz uso do AioSQLite, uma extensão assíncrona do banco de dados SQLite. O AioSQLite permite que o FACTUAL interaja com o banco de dados de forma eficiente, armazenando os resultados da classificação de *fake news* em tempo real. Esse framework assegura que a gravação de dados, como a classificação das notícias, a confiança calculada e as justificativas, ocorra sem interrupções no fluxo de trabalho do sistema.

É válido salientar que semáforo assíncrono controla o número de tarefas que podem ser executadas simultaneamente. Conforme mostrado na Figura 4.3, o asyncio.Semaphore limita o número de tarefas, evitando a sobrecarga dos recursos do sistema e garantindo que as operações sejam processadas dentro da capacidade do servidor. Isso permite que o FACTUAL mantenha um equilíbrio entre desempenho e eficiência no processamento de grandes volumes de notícias.

4.5 CONSIDERAÇÕES FINAIS

Neste capítulo, foi apresentada a FACTUAL, um sistema baseado em LLMs para a detecção e mitigação de *fake news*. O desenvolvimento do FACTUAL foi estruturado em torno de componentes integrados com o objetivo de oferecer uma solução eficiente e escalável para a análise de grandes volumes de dados textuais. A solução propõe uma abordagem para lidar com os desafios impostos pela desinformação, combinando técnicas de processamento de linguagem natural com estratégias otimizadas de integração e armazenamento de dados. Além disso, foi apresentado o fluxo de processamento das notícias, desde sua entrada no sistema até a classificação e geração de justificativas claras e fundamentadas para os usuários.

Os principais diferenciais técnicos do FACTUAL incluem a utilização de frameworks assíncronos, como Asyncio e AioSQLite, para garantir escalabilidade e alta concorrência, bem como a implementação de um mecanismo dinâmico do nível de confiança, que reflete a precisão e a credibilidade do modelo na

análise dos conteúdos. As considerações apresentadas neste capítulo estabelecem as bases para a avaliação do desempenho do FACTUAL, que será explorada no próximo capítulo.

5 AVALIAÇÃO DE DESEMPENHO

Neste capítulo, são apresentados os resultados obtidos na avaliação de desempenho do sistema FACTUAL. O objetivo deste capítulo é validar a eficácia do sistema proposto na detecção e mitigação de *fake news*, utilizando métricas quantitativas e qualitativas que evidenciem sua precisão, robustez e confiabilidade. A análise foi realizada com base em datasets amplamente reconhecidos e em dados coletados de portais de notícias, permitindo uma avaliação abrangente do desempenho do modelo em cenários reais.

5.1 CONFIGURAÇÃO PARA OS EXPERIMENTOS

Para avaliar o desempenho do FACTUAL, foram utilizados dois datasets amplamente reconhecidos pela comunidade acadêmica no campo da detecção de *fake news*: (i) FakeBR; e (ii) *fake news*. O FakeBR é um dataset brasileiro que contém notícias reais e falsas compartilhadas em redes sociais. Ele é especialmente importante para avaliar o desempenho do FACTUAL no contexto brasileiro, pois captura as nuances culturais e linguísticas do país. Com um total de 7.200 notícias, o FakeBR apresenta uma distribuição equilibrada entre notícias verdadeiras e falsas, abrangendo temas variados, tais como política, saúde e entretenimento. Além disso, sua composição inclui textos de diferentes tamanhos e estilos de escrita, garantindo que o modelo seja avaliado em um espectro representativo de conteúdos textuais. Já o *fake news* é uma coleção internacional de notícias rotuladas como “fake” ou “real”, utilizada em pesquisas acadêmicas devido à sua riqueza de dados. O *fake news* inclui mais de 20.000 notícias, com informações complementares sobre o contexto de compartilhamento, como popularidade e tempo de disseminação. Ele permite testar o FACTUAL em um cenário mais global, avaliando sua capacidade de generalizar entre diferentes linguagens, contextos culturais e padrões de desinformação.

Além dos datasets mencionados, é válido mencionar que foram coletadas informações adicionais diretamente de portais de notícias confiáveis e suspeitos, com o objetivo de aumentar a diversidade e a complexidade das amostras. Essa coleta incluiu textos curtos e longos, cobrindo uma ampla gama de tópicos tais como política, economia e saúde pública. As notícias foram classificadas de acordo com seu conteúdo (real ou falso), contexto de disseminação e características específicas de cada portfólio de notícias.

5.1.1 DETALHAMENTO DA AMOSTRA DE DADOS

A amostra de dados utilizada para os experimentos do FACTUAL foi composta por dados provenientes de fontes estruturadas, semi-estruturadas e não estruturadas. Cada formato de dado possui características e desafios próprios no que diz respeito ao processamento e análise.

- **Estruturados**: O dataset FakeBR e o *fake news* contêm informações bem definidas, como o rótulo (fake ou real) de cada notícia, data de publicação e, no caso do *fake news*, informações adicionais sobre o compartilhamento (popularidade e tempo de disseminação). Esses dados são estruturados

em tabelas, com campos bem definidos e relações claras entre as diferentes entidades.

- **Semi-Estruturado**: Algumas das notícias coletadas a partir de portais confiáveis e suspeitos podem apresentar dados semi-estruturados, como metadados extraídos de HTML, etiquetas de categoria (politica, saúde, etc.), e informações relacionadas a comentários ou interações dos usuários. Essas informações não estão tão rigidamente estruturadas como nos dados totalmente estruturados, mas ainda podem ser interpretadas e organizadas para análise, como por exemplo, tags associadas a cada notícia.

- **Não Estruturados**: A maioria das notícias, sejam elas do FakeBR, *fake news* ou de fontes adicionais, são textos não estruturados. Essas notícias são compostas por conteúdos textuais em forma de parágrafos, títulos, subtítulos, e informações de conteúdo, sem uma organização formal definida. O processamento de dados não estruturados exige técnicas de processamento de linguagem natural (PLN) para análise, como tokenização, extração de entidades, e análise de sentimentos, que são utilizadas para transformar essas informações em dados úteis para os experimentos de detecção de *fake news*.

A técnica de hold-out foi utilizada para dividir o dataset consolidado em três partes: 60% para treinamento, 30% para teste e 10% para validação. Essa técnica foi escolhida devido à sua simplicidade, eficiência e adequação a conjuntos de dados grandes, como os utilizados nesta pesquisa.

A implementação do FACTUAL foi realizada utilizando os seguintes conjuntos de bibliotecas e ferramentas: Python 3.11 para a programação, SQLite para gerenciamento de banco de dados, LLaMa 3.2 como o modelo de linguagem principal, Asyncio para processamento assíncrono, Hub Ollama para integração de modelos, e PyTorch para construção e treinamento do modelo de aprendizado de máquina.

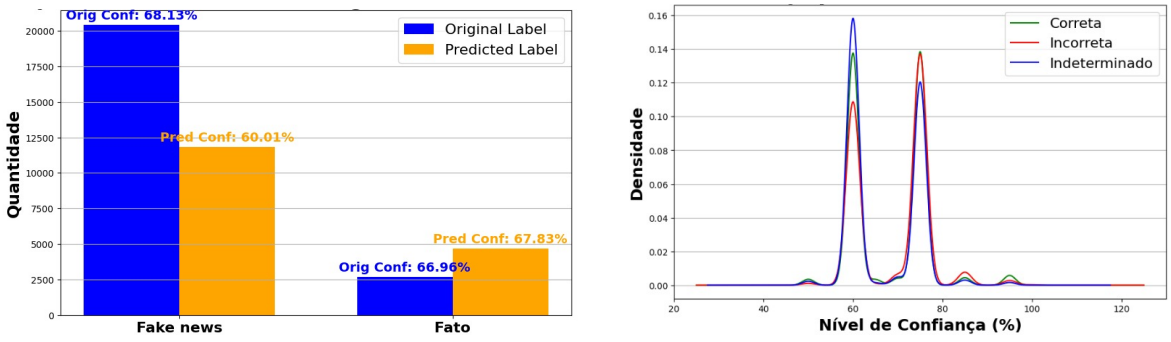
Os experimentos foram realizados em um servidor com o sistema operacional Ubuntu 23.04, utilizando o Kernel 6.2.0-39-generic. A máquina é equipada com um processador Intel Core i7-10700F, operando a uma frequência de 2,90 GHz, com arquitetura x86_64 e um total de 16 núcleos físicos, capazes de executar 16 threads simultaneamente devido ao suporte de dois threads por núcleo. O servidor dispõe de 32 GB de memória RAM e um volume de armazenamento configurado com LVM (Logical Volume Manager). Além disso, o cluster de GPU é composto por 6 GPUs NVIDIA GeForce RTX 3060 Ti, cada uma com 8 GB de memória dedicada. Essas GPUs possuem um número significativo de CUDA cores, que são unidades de processamento paralelas responsáveis por executar operações simultâneas em grandes volumes de dados. O número de CUDA cores nas GPUs GeForce RTX 3060 Ti acelera significativamente o processamento, especialmente em tarefas de aprendizado de máquina e inteligência artificial, como o treinamento de modelos do FACTUAL. Cada GPU tem a capacidade de realizar múltiplas operações em paralelo, o que melhora a eficiência e reduz o tempo necessário para concluir experimentos complexos.

Essa configuração robusta de GPUs, com a capacidade de processamento paralelo das unidades CUDA, proporciona um ambiente ideal para experimentos que exigem alto poder computacional. A versão do Python empregada foi a 3.11, garantindo compatibilidade com as bibliotecas e frameworks utilizados no desenvolvimento do FACTUAL.

5.2 ANÁLISE DE NÍVEIS DE CONFIANÇA E TENDÊNCIA PREDITIVA DO FACTUAL

A Figura 5.2 apresenta dois aspectos importantes do desempenho do FACTUAL: a quantidade de amostras classificadas (Figura 5.1(a)) e a densidade dos níveis de confiança (Figura 5.1(b)) para os diferentes rótulos analisados. Na Figura 5.1(a), observa-se a comparação entre os rótulos originais (Original Label) e os rótulos previstos pelo modelo (Predicted Label) para as categorias *fake news* e "Fato". Para o rótulo *fake news*, a confiança média atribuída às amostras nos rótulos originais foi de 68,13%, enquanto nas previsões foi ligeiramente menor, atingindo 60,01%. Para o rótulo "Fato", a confiança média nos rótulos originais foi de 66,96%, enquanto a confiança das previsões apresentou uma leve melhora, atingindo 67,83%. A confiança representa o grau de certeza do modelo em relação à classificação realizada, indicando a probabilidade de que a previsão esteja correta. Essa variação na confiança média entre os rótulos originais e previstos evidencia a consistência do modelo, embora indique uma leve tendência de ajuste mais conservador na detecção de *fake news*.

Na Figura 5.1(b), é apresentada a densidade dos níveis de confiança para três categorias de previsão: Correta, Incorreta e Indeterminada. As curvas indicam a distribuição de confiança nas previsões, com picos principais em torno de 60% e 90%. Esses picos sugerem que o modelo tende a ser mais confiante nas classificações em torno desses intervalos, refletindo sua capacidade de diferenciar padrões claros em *fake news* e fatos. No entanto, a proximidade das curvas para previsões incorretas e indeterminadas sugere que, em alguns casos, o modelo enfrenta dificuldades em separar essas categorias, o que pode indicar a necessidade de ajustes nos limiares de confiança ou maior refinamento do treinamento para melhorar a separabilidade entre previsões corretas e errôneas. Esses resultados demonstram que o FACTUAL mantém um desempenho consistente, mas destacam áreas de potencial aprimoramento, como a redução das previsões classificadas como indeterminadas e ajustes nos limiares de confiança para aumentar a precisão geral



(a) Quantidade de amostras.

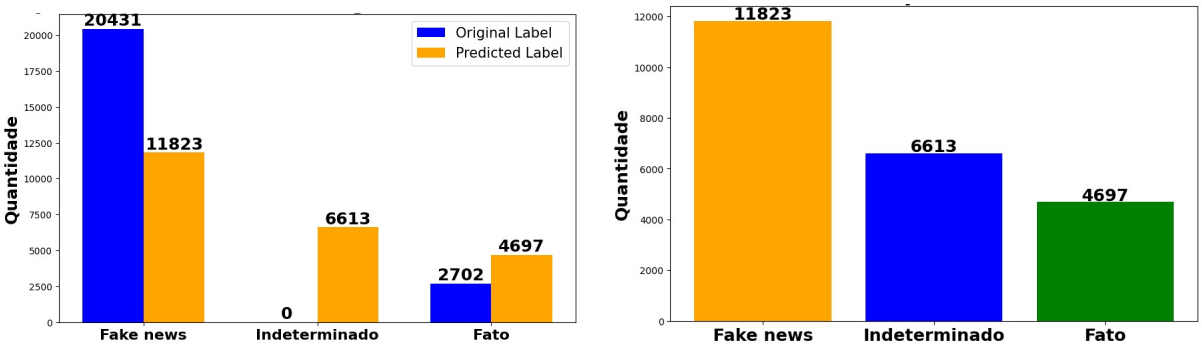
(b) Densidade.

Figura 5.1: Desempenho do FACTUAL para as métricas quantidade de amostras e densidade.

A Figura 5.2 apresenta uma análise do desempenho do sistema FACTUAL com relação aos rótulos *fake news*, Indeterminado, e Fato. São exibidos dois aspectos importantes: a comparação entre os rótulos originais e previstos (Figura 5.2(b)) e a distribuição final dos rótulos previstos pelo FACTUAL (Figura 5.2(b)). Na Figura 5.2(b), observa-se que, para o rótulo *fake news*, o sistema identificou 20.431 instâncias no conjunto de rótulos originais, enquanto previu 11.823 para a mesma categoria. Isso indica uma redução

significativa nas instâncias classificadas como *fake news* nas previsões, possivelmente refletindo uma tendência mais conservadora do modelo em atribuir notícias à categoria de desinformação. No caso do rótulo Fato, o FACTUAL originalmente classificou 2.702 instâncias, mas previu 4.697, evidenciando uma superestimação de notícias como verdadeiras nas previsões. Para o rótulo Indeterminado, o FACTUAL não possui instâncias no conjunto original (valor 0), mas gerou 6.613 previsões nessa categoria, indicando que um número considerável de notícias foi classificado com incerteza.

A Figura 5.2(b) complementa a análise ao exibir a distribuição dos rótulos previstos. O FACTUAL apresenta uma predominância de previsões para *fake news* (11.823 instâncias), seguida pela categoria *Indeterminad* (6.613 instâncias) e, por fim, Fato (4.697 instâncias). Essa distribuição destaca duas tendências: (i) a diminuição da quantidade de instâncias previstas como *fake news* em relação aos rótulos originais e o aumento significativo de previsões na categoria Indeterminado. A introdução da categoria Indeterminado sugere que o sistema enfrentou dificuldades em classificar com alta confiança algumas notícias, refletindo incertezas no processo preditivo. Esses resultados indicam que, embora o FACTUAL demonstre boa capacidade de detecção de *fake news*, ajustes nos parâmetros do modelo podem ser necessários para melhorar a separabilidade entre as categorias e reduzir a proporção de previsões indeterminadas. O aumento das instâncias previstas como "Fato" também sugere uma necessidade de refinamento para mitigar a tendência de superestimar a veracidade de certas notícias.



(a) Comparação entre os rótulos.

(b) Distribuição dos rótulos.

Figura 5.2: Desempenho do FACTUAL para os rótulos*fake news*, Indeterminado e Fato.

5.3 VALIDAÇÃO DA EFICÁCIA DO FACTUAL NA DETECÇÃO DE *FAKE NEWS* E FATOS

Esta seção apresenta os resultados da validação do FACTUAL por meio de um formulário composto por 40 perguntas, cujo objetivo foi avaliar a eficácia do FACTUAL em interpretar e justificar classificações relacionadas à identificação de desinformação. O formulário foi aplicado a 106 participantes, e suas perguntas podem ser consultadas no Apêndice 6.1.

A Figura 5.3 apresenta o percentual de acertos e erros obtidos no formulário em relação ao FACTUAL. Os resultados revelam variações significativas no desempenho dos participantes ao longo das perguntas. Questões com maior percentual de acertos sugerem que o FACTUAL forneceu explicações claras e co-

erentes, auxiliando os usuários na validação de informações. Este é um dos principais pontos fortes do modelo: (i) sua capacidade de gerar justificativas compreensíveis para as classificações realizadas. Por outro lado, perguntas com maior percentual de erros podem estar associadas a cenários mais complexos, em que os desafios estão relacionados à interpretação das respostas do sistema ou à dificuldade intrínseca das questões, como aquelas envolvendo informações parcialmente verdadeiras ou *fake news* mais sofisticadas.

É importante destacar que o bom desempenho em determinadas perguntas reforça a capacidade do FACTUAL de generalizar para múltiplos contextos e domínios heterogêneos, um dos principais desafios enfrentados por sistemas baseados em LLMs. No entanto, a presença de erros em outras questões evidencia lacunas que precisam ser abordadas, como o aprimoramento da robustez do modelo para identificar nuances em *fake news* e a adaptação das justificativas para atender a diferentes perfis de usuários, especialmente aqueles sem formação técnica.

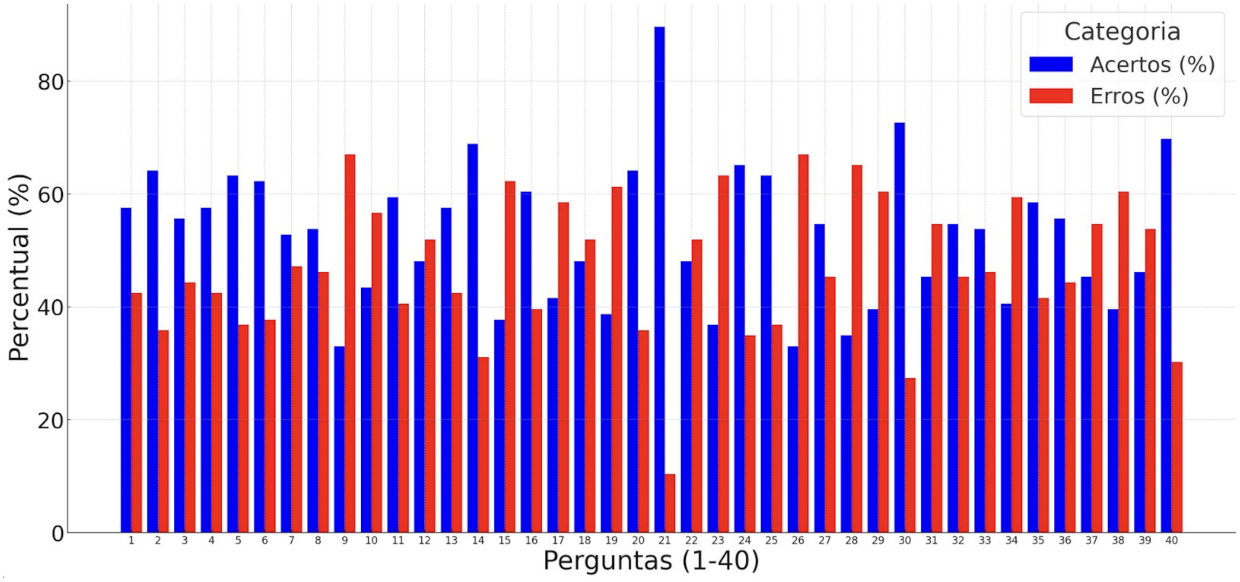
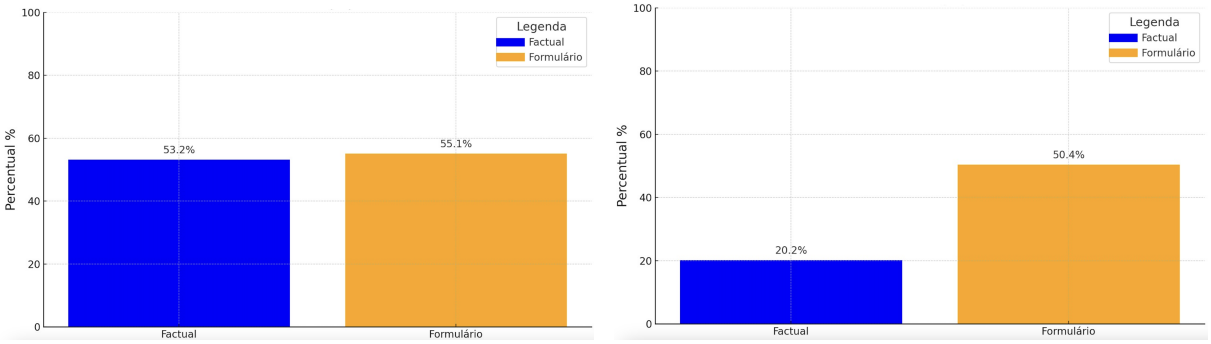


Figura 5.3: Percentual de Acertos e Erros do Formulário em Relação ao FACTUAL.

A Figura 5.4 apresenta uma comparação percentual entre o desempenho do FACTUAL e o formulário utilizado para validação, em relação à detecção de *fake news* (Figura 5.4(a)) e fatos verdadeiros (Figura 5.4(b)). Na detecção de *fake news* (Figura 5.4(a)), o FACTUAL alcançou 53,2% de erros, enquanto os participantes do formulário obtiveram 55,1% de erros. Em outras palavras, o FACTUAL errou menos em relação ao desempenho humano, reforçando sua consistência na identificação de conteúdos falsos. Essa consistência é particularmente relevante em cenários nos quais a interpretação de padrões linguísticos associados a *fake news* apresenta maior grau de subjetividade.

Já na detecção de fatos (Figura 5.4(b)), o FACTUAL apresentou um desempenho significativamente superior, com uma taxa de erro de apenas 20,2%, enquanto os participantes erraram em média 50,4%. Esse resultado reflete um ponto positivo do modelo: sua menor taxa de erros em relação aos participantes humanos na identificação de informações verdadeiras. Esse desempenho sugere que o FACTUAL demonstra maior precisão em evitar classificações incorretas de fatos verdadeiros, evidenciando sua consistência ao lidar com cenários que demandam maior rigor na validação de informações. Essa característica pode ser especialmente valiosa em contextos sensíveis, em que a categorização equivocada de fatos pode gerar

desinformação adicional.



(a) Fake. (b) Fato.

Figura 5.4: Desempenho do FACTUAL em comparação com o formulário

5.4 CUSTO COMPUTACIONAL X ESCALABILIDADE DA SOLUÇÃO

A relação entre custo computacional e escalabilidade de uma solução é um fator crítico em projetos que envolvem o processamento de grandes volumes de dados, especialmente em sistemas que utilizam GPUs e CUDA cores. A escalabilidade refere-se à capacidade de um sistema ou solução de lidar com aumentos na carga de trabalho, seja em termos de volume de dados ou número de operações simultâneas, sem que haja uma degradação significativa do desempenho. Já o custo computacional está associado aos recursos necessários para a execução dessas tarefas, como tempo de processamento, consumo de energia e uso de memória.

5.4.1 Escalabilidade da Solução

A escalabilidade de uma solução depende da arquitetura de hardware e do modelo de computação empregados. No caso de GPUs com CUDA cores, a escalabilidade é garantida pela capacidade de processar múltiplas operações de forma paralela, permitindo que o sistema aumente o desempenho à medida que mais recursos (como mais GPUs ou núcleos) sejam adicionados. Esse tipo de arquitetura é especialmente vantajoso em cenários que exigem treinamento de modelos complexos de aprendizado de máquina ou processamento de grandes volumes de dados.

O uso de múltiplas GPUs, como no caso de clusters de GPU, também contribui para a escalabilidade. À medida que o número de GPUs aumenta, o sistema pode distribuir a carga de trabalho entre elas, mantendo o desempenho mesmo com o crescimento do volume de dados. Essa característica é fundamental em tarefas como o treinamento de redes neurais profundas, onde o número de parâmetros é grande e o poder de processamento das GPUs é aproveitado ao máximo.

5.4.2 Custo Computacional

O custo computacional de uma solução envolve diversos aspectos, entre eles:

- **Tempo de Processamento:** O tempo necessário para realizar uma tarefa, como o treinamento de um modelo de aprendizado de máquina, pode ser elevado dependendo da complexidade do modelo e do volume de dados. A utilização de GPUs com CUDA cores pode reduzir significativamente esse tempo, mas a eficiência do código também é um fator determinante para o desempenho.
- **Consumo de Energia:** As GPUs, especialmente em grandes clusters, podem consumir uma quantidade significativa de energia, impactando diretamente os custos operacionais. Esse aspecto deve ser considerado ao planejar a escalabilidade de uma solução, pois o aumento no número de GPUs pode resultar em um aumento proporcional no consumo de energia.
- **Uso de Memória:** Soluções que lidam com grandes volumes de dados ou modelos com muitos parâmetros podem exigir grandes quantidades de memória. Embora as GPUs com memória dedicada (como as 8 GB da RTX 3060 Ti) possam oferecer um bom desempenho, é essencial monitorar o uso de memória para evitar gargalos que impactem o processamento.
- **Custo de Infraestrutura:** Em ambientes que utilizam clusters de GPU, o custo de aquisição, manutenção do hardware e a gestão da infraestrutura também devem ser levados em consideração. Além disso, o custo associado ao uso de infraestrutura de nuvem pode influenciar no custo total de operação.

5.4.3 Equilíbrio entre Custo Computacional e Escalabilidade

Encontrar o equilíbrio entre custo computacional e escalabilidade é essencial para otimizar o desempenho da solução sem aumentar excessivamente os custos. Em muitos casos, a utilização de técnicas de paralelização, como o processamento distribuído em múltiplas GPUs, pode melhorar a escalabilidade sem um aumento linear no custo computacional. Além disso, otimizações no código, como a utilização de bibliotecas de baixo nível que aproveitam de forma eficiente o poder das GPUs, podem contribuir para reduzir o tempo de processamento sem aumentar de forma significativa o consumo de recursos.

Ao planejar uma solução computacional escalável, é necessário considerar não apenas a capacidade do sistema de aumentar o desempenho com a adição de recursos, mas também os custos envolvidos. A escolha de uma arquitetura adequada, juntamente com a otimização do código e o gerenciamento eficiente dos recursos, pode proporcionar uma solução escalável que equilibre de forma eficaz o desempenho e o custo computacional.

5.5 CONSIDERAÇÕES FINAIS

Os resultados obtidos na avaliação do FACTUAL mostram que o sistema apresenta consistência e confiança na análise de notícias, mesmo ao lidar com cenários desafiadores. No entanto, algumas discrepâncias

foram identificadas, evidenciando áreas que podem ser ajustadas para melhorar o desempenho do modelo. Entre os pontos observados estão: a redução nas previsões para o rótulo *fake news* em comparação aos rótulos originais, o aumento de previsões na categoria Indeterminado, indicando incertezas em algumas classificações, e a tendência do sistema em superestimar o rótulo Fato, o que sugere a necessidade de ajustes para equilibrar melhor as previsões entre as categorias.

Salienta-se que a validação do FACTUAL por meio de um formulário composto por 40 perguntas aplicadas a 106 participantes reforçou sua capacidade de gerar explicações compreensíveis e coerentes para as classificações realizadas. O formulário revelou que o FACTUAL apresentou uma menor taxa de erros em relação aos participantes humanos, especialmente na detecção de *fake news*. Contudo, algumas limitações foram evidenciadas, como a dificuldade do modelo em diferenciar cenários mais complexos ou informações parcialmente verdadeiras, o que reforça a necessidade de ajustes para melhorar a precisão em tais contextos. Essa análise destacou a importância de estratégias voltadas para a adaptação das justificativas fornecidas pelo modelo, a fim de atender a diferentes perfis de usuários.

Essas observações destacam a importância de ajustes nos limiares de confiança, maior treinamento em dados representativos e alterações nos parâmetros do modelo, especialmente para lidar com categorias mais desafiadoras e reduzir a quantidade de previsões indeterminadas. Além disso, o aumento nas previsões de "Fato" em relação aos rótulos originais indica que o FACTUAL poderia se beneficiar de estratégias para evitar a superestimação de notícias como verdadeiras.

A avaliação do FACTUAL indica que o sistema é funcional no combate à desinformação, apresentando resultados consistentes em termos de níveis de confiança e classificações. O sistema demonstra potencial como ferramenta escalável e confiável, capaz de lidar com grandes volumes de dados e fornecer análises baseadas em evidências. Essas análises fornecem as bases para melhorias futuras, com o objetivo de aumentar sua precisão e aplicabilidade em cenários reais.

6 CONCLUSÃO E TRABALHOS FUTUROS

A disseminação de *fake news* tem causado impactos em diversos setores da sociedade, tais como saúde pública, processos eleitorais e plataformas de mídia social. Esse fenômeno foi intensificado pela popularização das redes sociais e pela ausência de mecanismos regulamentares eficazes, o que tem comprometido a confiança na informação e gerado consequências sociais, políticas e econômicas graves. A problemática abordada nesta dissertação consistiu na necessidade de desenvolver um sistema eficiente, escalável e transparente para identificar e mitigar a desinformação, especialmente considerando os desafios impostos pela ambiguidade textual, a velocidade de disseminação e a escassez de dados rotulados de alta qualidade.

Nesse contexto, o objetivo principal desta dissertação foi propor, implementar e avaliar o FACTUAL (*Fake News Analysis with Confidence and Trust Using AI and LLM*), um sistema baseado em LLMs voltado para identificar e mitigar a *fake news*. O FACTUAL foi modelado para não apenas identificar de maneira eficiênte informações falsas, mas também fornecer uma análise criteriosa da confiabilidade das fontes, utilizando um mecanismo dinâmico para atribuição de graus de confiança às previsões realizadas. Tal abordagem objetiva promover maior transparência e interpretabilidade nos resultados, contribuindo para decisões informadas e fundamentadas.

Os resultados obtidos demonstraram que o FACTUAL alcançou os objetivos propostos. O sistema foi capaz de identificar padrões de desinformação com alta precisão, utilizando o modelo LLaMA 3.2 integrado ao framework Ollama. Além disso, a implementação de frameworks assíncronos, como Asyncio e AioSQLite, possibilitou que o sistema processasse grandes volumes de dados textuais de maneira escalável e eficiente. O mecanismo dinâmico de cálculo de confiança mostrou-se eficaz ao fornecer classificações fundamentadas e claras, atendendo à necessidade de maior confiabilidade e transparência em sistemas de detecção de *fake news*.

A dissertação apresenta contribuições científicas e tecnológicas significativas, entre elas, destacando-se: (i) o desenvolvimento do sistema FACTUAL modelado para lidar com desafios de análise textual, maximizando a precisão na detecção de *fake news* e oferecendo uma solução escalável e eficiente para análises em tempo real; (ii) a implementação de frameworks assíncronos, que otimizam a integração com bancos de dados e permitem a execução paralela de tarefas, aumentando a escalabilidade e eficiência no processamento de grandes volumes de dados; (iii) a modelagem de uma categorização para descrever padrões de desinformação, tais como desinformação deliberada, informações falsas não intencionais e rumores, contribuindo para uma compreensão mais profunda do fenômeno e o refinamento das técnicas de detecção; e (iv) a modelagem de um mecanismo dinâmico de cálculo de confiança, que avalia palavras-chave e expressões indicativas de certeza ou incerteza nas respostas do modelo, ajustando dinamicamente o nível de confiabilidade das classificações e proporcionando maior clareza aos usuários. Essas contribuições não apenas avançam o estado da arte na detecção de *fake news*, mas também fornecem uma base para futuras investigações e aplicações em larga escala.

6.1 TRABALHOS FUTUROS

As contribuições apresentadas nesta dissertação abrem uma série de oportunidades para pesquisas futuras que podem ampliar e aprimorar as capacidades do FACTUAL. Entre as principais direções para trabalhos futuros, destacam-se:

- Generalização para Múltiplos Domínios e Idiomas: Adaptar o FACTUAL para operar em diferentes contextos e línguas, incluindo aquelas com recursos linguísticos limitados. Essa expansão pode ser alcançada por meio de técnicas de aprendizado por transferência e treinamentos específicos em datasets multilinguísticos e multidomínios, aumentando a aplicabilidade do sistema em cenários globais.
- Incorporação de Análise Multimodal: Estender as funcionalidades do FACTUAL para incluir dados não textuais, tais como imagens, vídeos e áudios, explorando abordagens multimodais. A combinação de técnicas de PLN com redes neurais convolucionais para análise multimodal pode tornar o sistema mais eficiente e abrangente.
- Redução de Custos Computacionais: Investigar estratégias de compressão de modelos e aprendizado eficiente para diminuir os custos computacionais associados à execução de LLMs. Essa abordagem tornará o FACTUAL mais acessível e aplicável em dispositivos com recursos limitados, tais como smartphones e sistemas embarcados.
- Integração com Plataformas de Mídia Social e Governamentais: Explorar a integração do FACTUAL com sistemas de monitoramento em tempo real de plataformas de mídia social e órgãos governamentais, oferecendo suporte para ações rápidas de mitigação de desinformação em larga escala.
- Aprimoramento do Mecanismo de Cálculo de Confiança: Refinar o algoritmo dinâmico de cálculo de confiança, incorporando métricas adicionais que avaliem não apenas o conteúdo textual, mas também aspectos contextuais e históricos, aumentando a precisão das decisões tomadas pelo sistema.

REFERÊNCIAS BIBLIOGRÁFICAS

1 PAPAGEORGIOU, E.; CHRONIS, C.; VARLAMIS, I.; HIMEUR, Y. A survey on the use of large language models (llms) in fake news. *Future Internet*, MDPI, v. 16, n. 8, p. 298, 2024.

2 ALGHAMDI, J.; LUO, S.; LIN, Y. A comprehensive survey on machine learning approaches for fake news detection. *Multimedia Tools and Applications*, Springer, v. 83, n. 17, p. 51009–51067, 2024.

3 JUNIOR, J. L. S.; GRAEML, A. R. Fake news e seus impactos nas organizações. 2021.

4 ALNABHAN, M. Q.; BRANCO, P. Fake news detection using deep learning: A systematic literature review. *IEEE Access*, IEEE, 2024.

5 BERNARDI, A. J. B. *Fake news e as eleições de 2018 no Brasil: como diminuir a desinformação?* [S.l.]: Editora Appris, 2021.

6 iG São Paulo. *Fake news marcaram as eleições de 2018; relembre as 10 mais emblemáticas*. 2018. Acessado em: 18 out. 2024. Disponível em: <https://ultimosegundo.ig.com.br/politica/2018-10-29/10-fake-news-das-eleicoes.html>.

7 SHARMA, K.; QIAN, F.; JIANG, H.; RUCHANSKY, N.; ZHANG, M.; LIU, Y. Combating fake news: A survey on identification and mitigation techniques. *ACM Transactions on Intelligent Systems and Technology (TIST)*, ACM New York, NY, USA, v. 10, n. 3, p. 1–42, 2019.

8 Brasil. *Lei n° 12.965, de 23 de abril de 2014 - Marco Civil da Internet*. 2014. Acesso em: 28 mar. 2025. Disponível em: <http://www.planalto.gov.br/ccivil_03/leis/l12965.htm>.

9 ZHANG, Q.; GUO, Z.; ZHU, Y.; VIJAYAKUMAR, P.; CASTIGLIONE, A.; GUPTA, B. B. A deep learning-based fast fake news detection model for cyber-physical social services. *Pattern Recognition Letters*, Elsevier, v. 168, p. 31–38, 2023.

10 WU, J.; GUO, J.; HOOI, B. Fake news in sheep's clothing: Robust fake news detection against llm-empowered style attacks. In: *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. [S.l.: s.n.], 2024. p. 3367–3378.

11 SUN, Y.; HE, J.; CUI, L.; LEI, S.; LU, C.-T. Exploring the deceptive power of llm-generated fake news: A study of real-world detection challenges. *arXiv preprint arXiv:2403.18249*, 2024.

12 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. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. [S.l.: s.n.], 2024. v. 38, n. 20, p. 22105–22113.

13 WANG, D.; ZHANG, W.; WU, W.; GUO, X. Soft-label for multi-domain fake news detection. *IEEE Access*, IEEE, 2023.

14 KALIYAR, R. K.; GOSWAMI, A.; NARANG, P. Fakebert: Fake news detection in social media with a bert-based deep learning approach. *Multimedia tools and applications*, Springer, v. 80, n. 8, p. 11765–11788, 2021.

15 SEGURA-BEDMAR, I.; ALONSO-BARTOLOME, S. Multimodal fake news detection. *Information*, MDPI, v. 13, n. 6, p. 284, 2022.

16 RAJA, E.; SONI, B.; BORGOHAIN, S. K. Fake news detection in dravidian languages using transfer learning with adaptive finetuning. *Engineering Applications of Artificial Intelligence*, Elsevier, v. 126, p. 106877, 2023.

17 DOU, Y.; SHU, K.; XIA, C.; YU, P. S.; SUN, L. User preference-aware fake news detection. In: *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*. [S.l.: s.n.], 2021. p. 2051–2055.

18 SANTOS, V. S. dos; RODRIGUES, P. H. B.; FILHO, G. P. R.; CANEDO, E. D.; PALMA, F. d. A. M.; MENDONÇA, F. L. L. de. Combate à desinformação com factual: Uma solução baseada em modelos de linguagem de grande escala.

19 SANTOS, V. S.; SERRANO, A. L. M.; BECKMAN, M. K.; PRACIANO, B. J. G.; FILHO, G. P. R.; CANEDO, E. D. UtilizaÇÃo de inteligencia artificial para verificar notÍcias falsas com a utilizaÇÃo do chatgpt.

20 CANEDO, E. D.; SOARES, L.; SILVA, G. R. S.; SANTOS, V. S. D.; MENDES, F. F. Do you see what happens around you? men's perceptions of gender inequality in software engineering. In: *Proceedings of the XXXVII Brazilian Symposium on Software Engineering*. [S.l.: s.n.], 2023. p. 464–474.

21 RABELO, B. S.; MOURA, L.; SANTOS, V. S. dos; FILHO, F. L. de C.; JR, R. T. de S.; MENDONÇA, F. L. L. de. Plataforma iot para predição de falhas em congeladores através de modelo auto-regressivo integrado de média movel (arima).

22 BRISOLA, A.; BEZERRA, A. C. Desinformação e circulação de "fake news": distinções, diagnóstico e reação. In: *XIX Encontro Nacional de Pesquisa em Ciencia da Informação (XIX ENANCIB)*. [S.l.: s.n.], 2018.

23 PENNYCOOK, G.; RAND, D. G. The psychology of fake news. *Trends in cognitive sciences*, Elsevier, v. 25, n. 5, p. 388–402, 2021.

24 SANTOS, R. L.; MONTEIRO, R. A.; PARDO, T. A. The fake. br corpus-a corpus of fake news for brazilian portuguese. In: *Latin American and Iberian Languages Open Corpora Forum (OpenCor)*. [S.l.: s.n.], 2018. p. 1–2.

25 MORONI, J. Possíveis impactos de fake news na percepção-ação coletiva. *Complexitas–revista de Filosofia temática*, v. 3, n. 1, p. 130–160, 2018.

26 DANTAS, L. F. S.; DECCACHE-MAIA, E. Divulgação científica no combate às fake news em tempos de covid-19. *Research, Society and Development*, v. 9, n. 7, p. e797974776–e797974776, 2020.

27 RUSSELL, S.; NORVIG, P. *Artificial intelligence: a modern approach. Hoboken*. [S.l.]: NJ: Pearson, 2020.

28 BRAGA, I. Revisão sistemática dos modelos de inteligência artificial generativa llm. In: *XIII Seminario Hispano-Brasileño de Investigación en Información, Documentación y Sociedad 2024*. [S.l.: s.n.], 2024.

29 BARBOSA, L. M.; PORTES, L. A. F. A inteligência artificial. *Revista Tecnologia Educacional [on line], Rio de Janeiro*, n. 236, p. 16–27, 2023.

30 MAHESH, B. Machine learning algorithms-a review. *International Journal of Science and Research (IJSR).[Internet]*, v. 9, n. 1, p. 381–386, 2020.

31 CHAVES, G. M.; ANDRADE, I. B. Identificaçao da estrutura de github wikis a partir de aprendizado de máquina.

32 KANG, Y.; CAI, Z.; TAN, C.-W.; HUANG, Q.; LIU, H. Natural language processing (nlp) in management research: A literature review. *Journal of Management Analytics*, Taylor & Francis, v. 7, n. 2, p. 139–172, 2020.

33 COSTA, F. C. Métodos e aplicações de processamento de linguagem natural. 002, 2021.

34 OLIVIERI, G. A. d. P. Governança global da inteligência artificial no combate à desinformação: o impacto do ato de ia da união europeia. Universidade Estadual Paulista (Unesp), 2024.

35 SCOTT, K.; SHAW, G. *O futuro da inteligência artificial: de ameaça a recurso*. [S.l.]: HARLEQUIN, 2023.

36 LAZZARESCHI, N.; GUERRA, C. M. F.; NAKAOKA, M. Y. T. O impacto das tecnologias gpt no futuro do trabalho. *Caderno Eletrônico de Ciências Sociais*, v. 11, n. 2, p. 38–51, 2023.

37 ROCHA, C. J. da; PORTO, L. V.; ABAURRE, H. E. Discriminação algorítmica no trabalho digital. *Revista de Direitos Humanos e Desenvolvimento Social*, v. 1, p. 1–21, 2020.

38 FREITAS, J.; FREITAS, T. B. Direito e inteligência artificial. *Belo Horizonte: Fórum*, 2020.

39 MORAES, A. L. Z. de; BARBOSA, L. V. F.; GROSSI, V. C. D. D. *Inteligência artificial e direitos humanos: aportes para um marco regulatório no Brasil*. [S.l.]: Editora Dialética, 2022.

40 FILHO, R. de S. M.; VIEIRA, R. de S. Estratégia brasileira de inteligência artificial (ebia): Avanços na cidadania a partir das aplicações no poder público. *Anais do Seminário Internacional em Direitos Humanos e Sociedade*, v. 6, 2024.

41 FELIX, H. M. A.; MEDEIROS, O. D. de. Inteligência artificial e teoria do risco no projeto de lei nº 2.338/2023. *RECIMA21-Revista Científica Multidisciplinar-ISSN 2675-6218*, v. 4, n. 11, p. e4114406–e4114406, 2023.

42 TIRONI, L. F. Governança global–ocde, regulação, normas técnicas e tecnologia digital. Instituto de Pesquisa Econômica Aplicada (Ipea), 2024.

43 MENDES, L. S. Transparência e privacidade: violação e proteção da informação pessoal na sociedade de consumo. 2010.

44 FERREIRA, R. C. V.; GARCIA, G. H. M.; BRASIL, D. R. O surgimento do chat gpt e a insegurança sobre o futuro dos trabalhos acadêmicos. *Cadernos de dereito actual*, n. 21, p. 130–143, 2023.

45 GRINBERG, N.; JOSEPH, K.; FRIEDLAND, L.; SWIRE-THOMPSON, B.; LAZER, D. Fake news on twitter during the 2016 us presidential election. *Science*, American Association for the Advancement of Science, v. 363, n. 6425, p. 374–378, 2019.

46 GUPTA, M.; DENNEHY, D.; PARRA, C. M.; MÄNTYMÄKI, M.; DWIVEDI, Y. K. Fake news believability: The effects of political beliefs and espoused cultural values. *Information & Management*, Elsevier, v. 60, n. 2, p. 103745, 2023.

47 PYTON, T. *PyDev.* [S.l.]: Disponível em: <https://www.pydev.org/> . Acessado em: 10 mar. 2023, 2023.

48 SILVA, R. O. da; SILVA, I. R. S. Linguagem de programação python. *Tecnologias em Projeção*, v. 10, n. 1, p. 55–71, 2019.

49 BARBOSA, J.; VIEIRA, J. P. A.; SANTOS, R.; JUNIOR, G. V. M.; MUNIZ, M. d. S.; MOURA, R. S. Introdução ao processamento de linguagem natural usando python. *III Escola Regional de Informatica do Piauí*, v. 1, p. 336–360, 2017.

50 MA, Z.; LUO, M.; GUO, H.; ZENG, Z.; HAO, Y.; ZHAO, X. Event-radar: Event-driven multi-view learning for multimodal fake news detection. In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. [S.l.: s.n.], 2024. p. 5809–5821.

51 HAN, L.; ZHANG, X.; ZHOU, Z.; LIU, Y. A multifaceted reasoning network for explainable fake news detection. *Information Processing & Management*, Elsevier, v. 61, n. 6, p. 103822, 2024.

52 SMITH, L.; THOMPSON, J. Cybersecurity in fake news detection: A deep learning approach to prevent digital manipulation. *Journal of Cybersecurity Research*, v. 45, n. 2, p. 113–126, 2023.

53 JONES, T.; ROBERTS, A. Mitigating cyber threats in fake news detection using machine learning techniques. *Cybersecurity and Threat Intelligence*, v. 11, n. 1, p. 88–101, 2023.

54 LIU, Y.; ZHANG, Q. Cybersecurity measures for fake news detection: Preventing manipulation in digital news spaces. *Cybersecurity and Privacy Journal*, v. 5, n. 4, p. 214–229, 2022.

55 PATEL, R.; KUMAR, A. Blockchain and cryptography for fake news detection: Enhancing cybersecurity in digital media. *Journal of Information Security and Cryptography*, v. 40, n. 2, p. 58–72, 2024.

56 LEE, S.; CHOI, M. Cyberattacks as a means to spread fake news: A countermeasure using artificial intelligence. *International Journal of Cybersecurity and Digital Forensics*, v. 9, n. 3, p. 227–239, 2024.

57 BROWN, T.; WILLIAMS, S. Defensive strategies against fake news in social networks: A cybersecurity perspective. *Journal of Digital Security*, v. 31, n. 1, p. 54–66, 2022.

58 MARTIN, G.; THOMPSON, C. Digital threats and fake news: Security measures to protect against information manipulation. *Cybersecurity and Digital Forensics Journal*, v. 6, n. 2, p. 102–115, 2023.

59 TAYLOR, P.; EDWARDS, R. Validation of news sources: Cybersecurity in the verification of digital media content. *Journal of Information Privacy and Security*, v. 19, n. 3, p. 78–90, 2023.

60 WANG, H.; ZHAO, J. Ensuring privacy in fake news detection systems: A cybersecurity approach using anonymization techniques. *Journal of Data Protection and Privacy*, v. 7, n. 4, p. 320–332, 2023.

61 AMARASINGHE, M.; KOTTEGODA, S.; ARACHCHI, A. L.; MURAMUDALIGE, S.; BANDARA, H. D.; AZEEZ, A. Cloud-based driver monitoring and vehicle diagnostic with obd2 telematics. In: IEEE. *2015 fifteenth international conference on advances in ICT for emerging regions (ICTer)*. [S.l.], 2015. p. 243–249.

62 ASTARITA, V.; GUIDO, G.; MONGELLI, D.; GIOFRÈ, V. P. A co-operative methodology to estimate car fuel consumption by using smartphone sensors. *Transport*, v. 30, n. 3, p. 307–311, 2015.

63 BARBOSA, V. H. F.; BESSA, G. M. A. Um estudo comparativo sobre as linguagens java e kotlin para o desenvolvimento de aplicativos android. *Caderno de Estudos em Sistemas de Informação*, v. 7, n. 1, 2022.

64 BIEHL, M. *API Architecture*. [S.l.]: API-University Press, 2015. v. 2.

65 BIEHL, M. *RESTful Api Design*. [S.l.]: API-University Press, 2016. v. 3.

APÊNDICES

FORMULÁRIO DE VALIDAÇÃO DO FACTUAL

Este anexo apresenta o formulário utilizado para validar o FACTUAL. O formulário é composto por 40 perguntas, divididas em afirmações, em que os participantes deveriam indicar se cada uma delas é "Fato" ou "Fake". Cada participante respondeu com base em sua percepção sobre a veracidade das afirmações apresentadas. As respostas coletadas foram analisadas para avaliar o desempenho do FACTUAL em relação ao julgamento humano. A seguir, as perguntas apresentadas:

1. - 25960 - Imagem de crianças (uma vacinada e outra não) que foram expostas à varíola é autêntica.
    - ( ) Fato
    - ( ) Fake
2. - 4611 - É falso que Lula disse que pretende acabar com o Pix caso seja eleito. Lula diz que um dos seus projetos é acabar com o PIX assim que assumir a presidência.
    - ( ) Fato
    - ( ) Fake
3. - 6811 - Nas redes sociais, Auxílio Brasil gera dúvidas sobre valor e quem pode receber o benefício. Programa que vai garantir a sobrevivência de cerca de 17 milhões de famílias que foram ainda mais prejudicadas com a pandemia e a crise do fique em casa.
    - ( ) Fato
    - ( ) Fake
4. - 4716 - É montagem foto que mostra atirador de Suzano usando camisa de Bolsonaro Bolsonaro fazendo história. Luiz Henrique de Castro, autor da chacina de Suzano.
    - ( ) Fato
    - ( ) Fake
5. - 26808 - Já existem "mais sargentos e oficiais do que praças" nas Forças Armadas de Portugal.
    - ( ) Fato
    - ( ) Fake
6. - 26371 - Coronavírus. Ainda devemos deixar os sapatos que usamos na rua à porta de casa.
    - ( ) Fato
    - ( ) Fake

7. - 14083 - Xingar o presidente é crime, mas dificilmente leva à prisão Xingar o presidente é crime?

    - ( ) Fato
    - ( ) Fake

8. - 26801 - Rei dos Países Baixos deslocou-se de bicicleta para evento oficial em Haia.

    - ( ) Fato
    - ( ) Fake

9. - 5831 - É falso que Joe Biden doou US$ 235 milhões para 'terroristas do Hamas' "Joe Biden Financiou Terroristas do Hamas Doando US$235 Milhões Para Faixa de Gaza e Cisjordânia".

    - ( ) Fato
    - ( ) Fake

10. - 15408 - Papa Francisco apelou às mulheres católicas para que se reproduzissem com muçulmanos.

    - ( ) Fato
    - ( ) Fake

11. - 25126 - É falso que presidente do Chile declarou 'guerra a comunistas' Chile deu o exemplo. Agora é a nossa vez. A corda arrebentou. Militares em ação no Chile. Forças Armadas do Chile ficam ao lado do presidente. Sebastian Piñera comunica: estamos em guerra contra os socialistas comunistas.

    - ( ) Fato
    - ( ) Fake

12. - 13159 - É falso que 'Estadão' tenha noticiado candidatura de Suzane Von Richtofen Suzane Von Richtofen se candidatará a vereadora pelo PT.

    - ( ) Fato
    - ( ) Fake

13. - 24774 - Mitos e verdades sobre novo surto de sarampo, doença 'altamente contagiosa'. O surto de sarampo está ligado à crise venezuelana.

    - ( ) Fato
    - ( ) Fake

14. - 6930 - É falso que bonecos em hospitais da França simulam pacientes com variante ômicron França - Os hospitais estão tão 'saturados' da variante ômicron que nem dá tempo de colocar os braços nos manequins para mostrar na TV.

    - ( ) Fato
    - ( ) Fake

15. - 25356 - A Índia vai superar a China como "nação mais populosa do mundo em 2023", realça-se nas redes sociais A Índia vai superar a China como "nação mais populosa do mundo em 2023".

    - ( ) Fato
    - ( ) Fake

16. - 24336 - Marina diz que a Rede foi o único partido a defender a Lava Jato na TV. Será? Se comparar com o PSOL, que tem mais de dez anos, tivemos [a Rede] desempenho muito bom [nas eleições de 2016].

    - ( ) Fato
    - ( ) Fake

17. - 26914 - Apontado como chefe de milícia, Adriano da Nóbrega foi chamado de “brilhante oficial” por Bolsonaro em 2005 O presidente Jair Bolsonaro já elogiou o ex-policial militar Adriano da Nóbrega, apontado como chefe de milícia.

    - ( ) Fato
    - ( ) Fake

18. - 5840 - É falso que Argentina bloqueou ‘todas as contas bancárias’ da população “Todas as contas bancárias bloqueadas! Viva o socialismo. Muito triste! Não souberam votar, votaram na esquerda maldita (...)”.

    - ( ) Fato
    - ( ) Fake

19. - 26572 - Portugal é o país da União Europeia com "mais emigrantes espalhados pelo mundo".

    - ( ) Fato
    - ( ) Fake

20. - 5881 - É falso que Bolsonaro ganha R$ 68 mil de aposentadoria. Bolsonaro recebe R$ 68.548,32 de aposentadoria.

    - ( ) Fato
    - ( ) Fake

21. - 12873 - Cabo Daciolo foi a lançamento de biografia de Karl Marx.

    - ( ) Fato
    - ( ) Fake

22. - 24330 - Lula usa dados falsos sobre voto de Fachin e erra sobre variante de Manaus. Eu era presidente da República, a gente vacinou 83 milhões de brasileiros em três meses.

- ( ) Fato
- ( ) Fake

23. - 24362 - Vacina contra HIV: conheça os mitos e as verdades sobre o imunizante em teste no Brasil. A fase de teste da vacina de HIV, deve durar uns anos.

    - ( ) Fato
    - ( ) Fake

24. - 4037 - É falso que ACM morreu em acidente aéreo após criticar Lula em discurso "ACM morreu em Julho de 2.007, em um 'acidente' (???) de helicóptero. Poucas semanas antes, havia feito um pronunciamento na tribuna do Senado, em plena vigência do 2º mandato de Lula. Tal pronunciamento até hoje não havia sido divulgado.".

    - ( ) Fato
    - ( ) Fake

25. - 26877 - Lituânia mudou nome de rua da Embaixada Russa para "Rua dos Heróis Ucranianos".

    - ( ) Fato
    - ( ) Fake

26. - 7079 - É falso que bandeira brasileira foi 'espalhada por Dubai' graças a Bolsonaro Dubai, espalha bandeira do Brasil pelo país.

    - ( ) Fato
    - ( ) Fake

27. - 26778 - Poluição atmosférica causou mais de 347 mil mortes na União Europeia em 2019.

    - ( ) Fato
    - ( ) Fake

28. - 14363 - Tuíte engana ao afirmar que vacinas usam células de fetos abortados. Conteúdo verificado: Texto do site Estudos Nacionais afirma que a vacina em testes no Brasil é produzida com células de fetos abortados. O artigo foi publicado originalmente no site Brasil Livre em junho. Mas viralizou na última semana depois de ser reproduzido pelo Estudos Nacionais e divulgado pelo Twitter.

    - ( ) Fato
    - ( ) Fake

29. - 6914 - Golpistas usam nome do Carrefour e prometem iPhone para roubar dados de usuários Hoje, 17 fevereiro, 2021, você foi escolhido para participar de nossa pesquisa. Você levará apenas um minuto e receberá um prêmio fantástico: um iPhone 12 Pro! Cada Quarta feira nós escolhemos aleatoriamente 50 usuários para dar a eles a chance de ganhar prêmios incríveis. O prêmio de hoje é um iPhone 12 Pro! Haverá 50 vencedores sortudos. Esta pesquisa visa melhorar a qualidade do

atendimento aos nossos usuários e sua participação será recompensada 100%. Você só tem 3 minutos e 37 segundos, para responder a esta pesquisa! Apresse-se, o número de prêmios disponíveis é limitado!

- ( ) Fato
- ( ) Fake

30. - 24779 - É falso que primeiro-ministro britânico Boris Johnson fingiu tomar vacina contra a Covid-19 E o primeiro ministro afirmando que o povão tem que v@chin@r! Só esqueceram de tirar a tampa da agulha .

    - ( ) Fato
    - ( ) Fake

31. - 12970 - Carta falsa de general exalta ditadura militar e faz críticas ao STF Carta ao Brasil do general Gilberto Pimentel critica STF e exalta ditadura.

    - ( ) Fato
    - ( ) Fake

32. - 14225 - Rede de supermercado não obrigou funcionários a comemorarem o golpe de 1964.

    - ( ) Fato
    - ( ) Fake

33. - 25286 - Contagem manual de voto impresso foi apresentada em comissão da Câmara. Nunca se falou de contagem manual em discussão sobre voto impresso.

    - ( ) Fato
    - ( ) Fake

34. - 7036 - É falsa informação de que advogado Cristiano Zanin é réu na Lava Jato Advogado de Lula vira réu por embolsar R$ 68 milhões da Fecomércio.

    - ( ) Fato
    - ( ) Fake

35. - 26906 - A bandeira brasileira foi projetada em uma montanha suíça em apoio no combate à COVID-19 Suíça projetou a bandeira brasileira na montanha Matterhorn em apoio à luta contra a pandemia da COVID-19.

    - ( ) Fato
    - ( ) Fake

36. - 13954 - Abrir registro do botijão pela metade não economiza gás e gera riscos Aprenda a economizar até 50% do seu gás com esta dica.

    - ( ) Fato

- ( ) Fake

37. - 26912 - Um cachorro foi colocado em quarentena em Hong Kong após dar positivo para o novo coronavírus Teste em cão dá positivo para o novo coronavírus.

    - ( ) Fato
    - ( ) Fake

38. - 5820 - Sete boatos e dois fatos sobre a reforma da previdência em redes sociais Tempo de carteira assinada. Aposentadoria atual: homens - 35 anos, mulheres - 30 anos; Reforma da Previdência: 40 anos.

    - ( ) Fato
    - ( ) Fake

39. - 6808 - É montagem foto de vagão de trem 'transportando Covid-19' "Covid19 sendo transportado... Mas ninguém conhecia essas siglas antes da contaminação ...." .

    - ( ) Fato
    - ( ) Fake

40. - 24407 - De olho nas verdades e mentiras que rondam o carnaval Analgésico para curar ressaca pode ser perigoso.

    - ( ) Fato
    - ( ) Fake