CONIC SEMESP

25º Congresso Nacional de Iniciação Científica

TÍTULO: ALETHEIA: DETECÇÃO DE FAKE NEWS COM MODELOS DE LINGUAGEM QUE RACIOCINAM

CATEGORIA: EM ANDAMENTO

ÁREA: CIÊNCIAS EXATAS, DA TERRA E AGRÁRIAS

SUBÁREA: Computação e Informática

INSTITUIÇÃO: UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ - UTFPR

AUTOR(ES): IGOR ARAUJO DE MATTOS

ORIENTADOR(ES): ROBSON PARMEZAN BONIDIA

Realização: SEMESP

IES parceiras: PUC CAMPINAS

CATEGORIA EM ANDAMENTO

1. RESUMO

A rápida propagação da desinformação na *internet*, sobretudo pelas redes sociais e aplicativos de mensagens, intensifica a necessidade de tecnologias que auxiliem na verificação da veracidade de conteúdos compartilhados. Este trabalho propõe elaborar uma aplicação baseada em modelos de linguagem de última geração, integrada à abordagem de raciocínio automatizado e à de explicabilidade, conforme a realidade linguística e sociocultural do Brasil. Nomeada Aletheia, que em grego se refere à revelação da verdade, a aplicação funcionará como um sistema transparente e acessível à detecção de *fake news*, priorizando a interpretabilidade das decisões tomadas pelo modelo. Para isso, foi recorrido a um conjunto robusto de dados rotulados em português, oriundos de diversas fontes jornalísticas e repositórios de *fact-checking*. Como aplicação final, pretende-se desenvolver um protótipo funcional em forma de bot para plataformas como Telegram ou WhatsApp, permitindo acessibilidade da solução a um público abrangente, principalmente em comunidade marginalizadas. Este trabalho visa acima de tudo promover a confiança e o engajamento crítico do usuário, reforçando a importância de ferramentas responsáveis e interpretáveis no combate às *fake news*.

2. INTRODUÇÃO

A proliferação da desinformação online representa um desafio crítico para o bem-estar social, a confiança pública e os processos democráticos. Notícias falsas possuem o potencial de incitar o medo, manipular a opinião pública e influenciar decisões cruciais. As *fake news* se espalham em uma velocidade sem precedentes, e um único indivíduo pode alcançar uma audiência tão vasta quanto os veículos midiáticos tradicionais, por exemplo, Fox News, CNN ou The New York Times (Aïmeur; Brassard, 2023; Flamino et al., 2023).

Além disso, os recentes avanços em Inteligência Artificial (IA) Generativa, especialmente os Modelos de Linguagem de Grande Escala (*Large Language Models* – LLMs), apresentam uma espécie de paradoxo: se, de um lado, eles podem

gerar conteúdos extremamente realistas, mas falsos, aumentando a dificuldade de detecção da desinformação (Sun et al., 2024); por outro, disponibilizam técnicas avançadas para a detecção, a identificação e a mitigação do avanço das *fake news*, convertendo-se em parte da solução para o próprio problema que contribuem para produzir (Kareem; Abbas, 2023).

Sites de verificação de fatos, como PolitiFact e Snopes, utilizam profissionais especializados para realizar a checagem manual das informações, porém a rápida propagação da desinformação torna esse processo extremamente trabalhoso, demorado e pouco escalável. Como resposta a esse desafio, nos últimos anos, foram desenvolvidos modelos baseados em IA para automatizar a verificação de fatos e aprimorar a detecção de notícias falsas (Liao et al., 2023).

Nesse contexto, o presente trabalho se justifica pela necessidade de aprimorar os métodos de verificação de fatos, combinando LLMs com técnicas *reasoning*, a fim de aumentar a transparência e a interpretabilidade dos modelos de detecção. A pesquisa pretende contribuir tanto com o progresso acadêmico quanto com o desenvolvimento de soluções práticas que possam ser aplicadas por jornalistas, pesquisadores e cidadãos comuns no combate à desinformação. Além disso, este estudo busca oferecer um sistema que auxilie na escolha de fontes confiáveis, promovendo um ambiente digital mais seguro e informativo.

3. OBJETIVOS

O objetivo desta pesquisa é desenvolver e implementar um modelo de Processamento de Linguagem Natural capaz de mitigar a desinformação online, detectando a possibilidade de notícias serem falsas ou reais. Para isso, serão analisados os principais LLMs em combinação com técnicas de raciocínio automatizado e explicabilidade (XAI/*reasoning*), visando avaliar a porcentagem de veracidade de uma determinada afirmação ou artigo. Além disso, busca-se incentivar a disseminação responsável de informações, contribuindo para a redução da propagação de notícias falsas e para o aumento da confiabilidade das análises.

4. METODOLOGIA

A metodologia está organizada em quatro etapas principais. (1) Primeiro, é realizada a aquisição de notícias rotulados como verdadeiras ou falsas, provenientes de fontes jornalísticas e plataformas de checagem de fatos de língua portuguesa. A etapa busca assegurar a qualidade, diversidade temática e representatividade linguística, em diferentes domínios (como política, saúde e ciência).

Em seguida, (2) técnicas de Engenharia de *Prompts* serão analisadas, incluindo abordagens *zero-shot*, *few-shot* e *chain-of-thought*, visando extrair raciocínios claros e justificativas interpretáveis por parte do modelo. Na etapa de (3) Configuração Experimental e Arquitetura *Ensemble* foram selecionados alguns LLMs, como Llama 3, Gemini 1.5, DeepSeek, Phi 4, Qwen 3 e Gemma 3, focando no desempenho, compatibilidade com o português e capacidade de explicabilidade.

Para aumentar a precisão e a robustez, será implementada uma Arquitetura *Ensemble*. Essa abordagem combinará múltiplos LLMs por meio de estratégias de votação e raciocínio estruturado, explorando os pontos fortes de cada modelo para reduzir vieses e melhorar a generalização. As métricas de desempenho e de explicabilidade serão usadas para validar a eficácia da abordagem.

Por fim, uma aplicação para o usuário final será disponibilizada, com base nos modelos e estratégias mais promissores, em formato de bot para plataformas como Telegram ou WhatsApp. O grande objetivo é oferecer uma ferramenta acessível, confiável e interpretável para identificação de possíveis notícias falsas.

5. DESENVOLVIMENTO

O desenvolvimento do projeto encontra-se em andamento. Até o momento, foram selecionados e preparados os conjuntos de dados, com destaque para a base FakeTrue.Br (Chavarro et al., 2023), com cerca de 3.582 notícias diversas em português, publicadas entre 2017 e 2023 (1.791 verdadeiras e 1.791 falsas), e o Dataset-Fake-news-and-True-news (Rodrigues, 2025), que contém 65.204 notícias, sendo 32.602 verdadeiras e 32.602 falsas. Essa escolha permitiu trabalhar com dados atualizados e relevantes para a realidade linguística e cultural do Brasil.

Além disso, a estrutura de código para o agente já foi implementada e uma série de testes de estratégias de *prompting* e a avaliação comparativa de LLMs com

foco em *Reasoning* está sendo conduzida. Com a conclusão desta fase de testes, os próximos passos incluem a análise quantitativa e qualitativa das predições dos modelos. Em seguida, será disponibilizado o protótipo funcional, validado e otimizado para a luta contra a desinformação.

6. RESULTADOS PRELIMINARES

Na etapa de desenvolvimento e testes das estratégias de *prompting* da pesquisa, foram definidos seis *prompts* baseados em técnicas de Engenharia de *Prompts*. Três deles utilizam o conhecimento interno dos modelos, enquanto os outros três recorrem ao conhecimento externo, coletado por meio de APIs e incorporado como contexto via RAG (*Retrieval-Augmented Generation*). As APIs utilizadas para a coleta das fontes externas foram *DuckDuckGo*, *GDELT* e *Google Fact Check*. Utilizou-se três APIs como precaução, para garantir que, caso alguma não retornasse informações, as demais pudessem suprir essa necessidade.

Atualmente, o trabalho encontra-se na fase de avaliação comparativa, com foco em *reasoning* e desempenho. Os modelos utilizados até o momento nesta fase foram *Falcon3-10B-Instruct*, *Mistral-7B-Instruct-v0.2-GPTQ*, *Llama-3.1-8B-Instruct*, *Phi-4*, *DeepSeek-V2-Lite-Chat* e *Gemma-3-27b-pt*. Futuramente, pretende-se incluir outros modelos, como *Sabiá-7b*, *Qwen3-32B*, *Gemini-1.5-Flash* e *GPT-3.5-Turbo*.

Durante essa fase, foram realizados testes iniciais com 50 notícias verdadeiras e 50 falsas, utilizando os seis modelos mencionados e as seis diferentes abordagens de *prompting*, tanto com conhecimento interno quanto externo. Observou-se que os modelos cujos resultados ficaram mais próximos dos rótulos reais foram *Llama-3.1-8B-Instruct* (~90% de acerto)*, Falcon3-10B-Instruct* (~85% de acerto) e *Mistral-7B-Instruct-v0.2-GPTQ* (~80% de acerto). Para viabilizar essa análise, foi desenvolvida uma função que extrai manualmente, das respostas dos modelos, a classificação e a probabilidade de certeza atribuída à notícia ser “fake” ou “real”.

A pesquisa segue em contínua fase de desenvolvimento. No qual, o foco atual é testar os modelos em um conjunto mais expressivo de notícias, incrementar os modelos propostos, aplicar a abordagem *Ensemble* com os modelos que

apresentarem os melhores resultados, tanto em assertividade quanto em qualidade das respostas, e, por fim, desenvolver um protótipo funcional da aplicação.

7. FONTES CONSULTADAS

AïMEUR, S. A. E.; BRASSARD, G. Fake news, disinformation and misinformation in social media: a review. **Social Network Analysis and Mining**, v. 13, p. 30, 2023. ISSN 1869-5469. Disponível em: https://doi.org/10.1007/s13278-023-01028-5.

CHAVARRO, J. et al. Faketruebr: Um corpus brasileiro de notícias falsas. In: ANAIS DA XVIII ESCOLA REGIONAL DE BANCO DE DADOS. 2023, Porto Alegre, RS, Brasil. **Anais [...]** Porto Alegre, RS, Brasil: SBC, 2023. p. 108–117. ISSN 2595-413X. Disponível em: https://sol.sbc.org.br/index.php/erbd/article/view/24352.

FLAMINO, J. et al. Political polarization of news media and influencers on twitter in the 2016 and 2020 US presidential elections. **Nature Human Behaviour**, v. 7, p. 904–916, 2023. Disponível em: https://doi.org/10.1038/s41562-023-01550-8.

KAREEM, W.; ABBAS, N. Fighting lies with intelligence: Using large language models and chain of thoughts technique to combat fake news. In: BRAMER, M.; STAHL, F. (Ed.). ARTIFICIAL INTELLIGENCE XL. 2023, Cham. **Anais [...]** Cham: Springer Nature Switzerland, 2023. p. 253–258. ISBN 978-3-031-47994-6.

LIAO, H. et al. Muser: A multi-step evidence retrieval enhancement framework for fake news detection. In: PROCEEDINGS OF THE 29TH ACM SIGKDD CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING. 2023. **Anais [...]** ACM, 2023. p. 4461–4472. Disponível em: http://dx.doi.org/10.1145/3580305.3599873.

RODRIGUES, M. D. L. **Dataset Fake News and True News Brazil**. 2025. Repositório GitHub. Licença CC BY 4.0. Disponível em: https://github.com/michelleluzrodrigues/Dataset-Fake-news-and-True-news.

SUN, Y. et al. **Exploring the Deceptive Power of LLM-Generated Fake News: A Study of Real-World Detection Challenges**, 2024. Disponível em: https://arxiv.org/abs/2403.18249.