Indicadores Semânticos na Engenharia de Prompts: Uma Abordagem Explicável para Detecção de Fake News

Camilla B. Quincozes[1], Diego Molinos[1], Rafael D. Araújo[1], Silvio E. Quincozes[1,2]

[1]PPGCO – Universidade Federal de Uberlândia (UFU)

[2]Horizon AI Labs/PPGES – Universidade Federal do Pampa (UNIPAMPA)

{camillaquincozes, rafael.araujo, diego.molinos}@ufu.br

Abstract. *The detection of fake news has benefited from the use of Large Language Models (LLMs). However, the decisions made by these models lack explainability. This work aims to fill this gap by constructing indicators (red flags) from news through prompt engineering, which enable explainable decision-making. As a result, a dataset was created with 16 red flags concerning true and false news. Experiments with the Random Forest classifier and the SHAP tool revealed an F1-Score of 95.38% in the explainable detection of fake news.*

Resumo. *A detecção de fake news tem se beneficiado do uso de Modelos de Linguagem de Grande Escala (LLMs). No entanto, as decisões tomadas por esses modelos carecem de explicabilidade. Este trabalho busca suprir essa lacuna ao construir indicadores (red flags) a partir de notícias por meio da engenharia de prompts, os quais que possibilitam uma tomada de decisão explicável. Como resultado, foi criado o um dataset com 16 red flags acerca de notícias verdadeiras e falsas. Experimentos com o classificador Random Forest e a ferramenta SHAP revelaram um F1-Score de 95,38% na detecção explicável de fake news.*

1. Introdução

A propagação de notícias falsas, popularmente denominada *fake news*, em ambientes digitais tem se intensificado nos últimos anos, impulsionada pelo alcance massivo das redes sociais e pela dificuldade da população em reconhecer conteúdos manipulados [KaabOmeir et al. 2024, Kanashina et al. 2023]. Nesse cenário, os Grandes Modelos de Linguagem, do inglês, *Large Language Models* – (LLM) emergem como ferramentas promissoras para apoiar a análise automatizada de textos suspeitos [Papageorgiou et al. 2024].

Dados recentes revelam que o Brasil segue entre os países com menor capacidade de discernimento entre fatos e desinformação, com menos de 60% de acerto médio na identificação da veracidade de uma notícia por parte da população [OCDE 2024]. O relatório da *BBC/Yonder Consulting* indica que 25% dos usuários não se sentem confiantes para identificar conteúdos falsos [BBC News 2024]. Ainda, o levantamento da *Meta* mostra que 29% dos brasileiros já acreditaram em *fake news* [Meta & Broadminded 2025].

Diante desses desafios, iniciativas vêm surgindo na literatura para a identificação automatizada de conteúdos enganosos. Alguns estudos baseiam-se em métodos tradicionais e bioinspirados, como SVM e *Random Forest*, enquanto outros se apoiam de forma limitada na utilização de LLMs, explorando apenas a capacidade de classificação com base em variações de *prompt* [Baarir and Djeffal 2021, Kong et al. 2020, Khanam et al. 2021,

Özbay and Alatas 2019, Aslam et al. 2021, Hu et al. 2024]. No entanto, observa-se uma lacuna na explicabilidade das decisões tomadas por esses métodos automatizados.

Este trabalho trata da aplicação de LLMs na identificação de padrões linguísticos recorrentes em notícias falsas (*red flags*), com foco na explicabilidade e na categorização semântica desses indícios. Para tanto, foram utilizados quatro LLMs para identificar *red flags* a partir de um conjunto de *fake news* analisadas. Ademais, foi proposto um *prompt* estruturado que visa analisar notícias a fim de avaliar a presença das *red flags* identificadas previamente usando uma LLM selecionada por seu desempenho superior em experimentos exploratórios. Por fim, com base nessa metodologia, foi construído o *dataset* rotulado *SBsegFakeNews2025* [1]. Este dataset contém as seguintes colunas: i) texto da notícia; ii) pontuações para cada *red flag* (16 colunas); a coluna "*temperatura*", computada a partir das pontuações das *red flags*; e, por fim, a classe (*i.e., fake news* ou normal).

Como prova de conceito da utilidade do *dataset* proposto, foram conduzidos experimentos por meio do classificador *Random Forest*, o qual alcançou uma F1-Score de 95,38%. Para além da classificação, foi realizada uma análise de interpretabilidade através da ferramenta de Inteligência Artificial Explicável (XAI) denominada SHAP (*SHapley Additive exPlanations*)[2]. Portanto, este trabalho contribui para a explicação estruturada das decisões tomadas por modelos automatizados, além de disponibilizar o *dataset SBsegFakeNews2025* como recurso reutilizável para pesquisas futuras.

2. Trabalhos Relacionados

Diferentes abordagens vêm sendo exploradas para a automação da detecção de *fake news*. Alguns estudos utilizam técnicas tradicionais de aprendizado de máquina, como TF-IDF (*Term Frequency–Inverse Document Frequency*) e SVM (*Support Vector Machine*) [Baarir and Djeffal 2021]. Outros empregam métodos tradicionais de aprendizado profundo, como LSTM e algoritmos de otimização bioinspirados [Özbay and Alatas 2019, Baarir and Djeffal 2021]. Ainda, há aqueles que exploram a tecnologia *Transformer* ou arquiteturas híbridas com técnicas de Processamento de Linguagem Natural [Aslam et al. 2021, Raza and Ding 2022].

Com o avanço dos LLMs, surgiram também propostas que buscam empregar essas ferramentas para auxiliar ou automatizar a tarefa de verificação de veracidade textual. Em [Khivasara et al. 2020], o LLM GPT-2 foi utilizado com o objetivo de verificar o conteúdo de uma notícia, se o mesmo havia sido gerado por inteligência artificial, e não, propriamente, para avaliar sua veracidade. Já estudos mais recentes, tais como [Anjos 2023] e [Hu et al. 2024], utilizam LLMs modernas como o ChatGPT e Bard para realizar ou comparar classificações de notícias falsas, avaliando o potencial e as limitações LLMs.

Portanto, observa-se que, embora a maioria dos estudos realize tarefas de classificação, há uma ausência de explicabilidade e de avaliação de indicadores semânticos. Além disso, os trabalhos analisados, em sua grande maioria, apresentam abordagens envolvendo apenas um LLM. Em contraste, esta proposta é voltada à análise compreensiva de indicadores semânticos com enfoque explicável para a detecção de *fake*

[1]O dataset SBsegFakeNews2025 está disponível em:`https://github.com/camillabdt/fakenews2025`.

[2]SHAP Tool. Available at: `https://shap.readthedocs.io/en/latest/`

news, fundamentado na extração e organização de padrões linguísticos recorrentes em *fake news* a partir da análise de LLMs modernas e de técnicas de engenharia de *Prompt*.

3. Caracterização de Fake News

3.1. Materiais e Métodos

De acordo com [Wazlawick 2009], esta pesquisa caracteriza-se como exploratória, experimental e quantitativa. Trata-se de um estudo que investiga aspectos pouco explorados no uso de LLMs aplicados à detecção de *fake news*, com foco na extração, categorização e explicabilidade de padrões discursivos. Destarte, os materiais e método de pesquisa são listados a seguir:

1. **Seleção de notícias:** O *dataset* ISOT Fake News Dataset[3] foi escolhido como material para representar a base rotulada de notícias verdadeiras e *fake-news*. No total, são 21.417 artigos autênticos e 23.481 falsos. Um dos aspectos distintivos desse *dataset* é que, embora tenha passado por etapas de limpeza e padronização, manteve intencionalmente erros de gramática e pontuação nos textos classificados como falsos. Isso o torna especialmente relevante para análises baseadas em linguagem natural, já que esses desvios formais podem funcionar como indícios úteis na detecção de falsificações textuais.
2. **Mapeamento de Red Flags:** O método adotado baseia-se no emprego de quatro LLMs para investigar padrões linguísticos recorrentes em textos identificados como notícias falsas: Qwen (versão *qwen-qwq-32b*), DeepSeek (versão *deepseek-r1-distill-llama-70b*), Gemma (versão *gemma2-9b-it*), e LLaMA (versão *llama-3.1-8b-instant*). Cada modelo analisou subconjuntos distintos de amostras do dataset, permitindo capturar uma variedade de *red flags*. A análise foi conduzida até o ponto em que novas amostras deixaram de revelar *red flags* inéditas, indicando uma convergência interpretativa por parte de cada modelo. Em seguida, a fim de remover-se redundâncias e ambiguidades, o ChatGPT foi empregado como modelo adicional, desempenhando o papel de agente de síntese neutro.

3.2. Análise das Red Flags Mapeadas

A execução do procedimento descrito na Seção 3.1 resultou em 16 *red flags*, sintetizadas na Tabela 1 e discutidas a seguir.

A presença de "exagero" ou "sensacionalismo" se manifesta em declarações que utilizam hipérboles ou superlativos, como no caso de uma espécie de água-viva descrita como "a mais venenosa já registrada no mundo", sem respaldo técnico. Esse tipo de construção busca intensificar a reação emocional do leitor, especialmente quando combinada com *apelos à urgência ou ao medo*, como a expressão "LIVE FEED" em letras maiúsculas, que cria uma sensação de emergência e pressiona o leitor a reagir.

Outra marca recorrente é a *falta de fontes confiáveis*, especialmente quando as informações provêm de veículos como "Truth Feed" ou "Milo.com", que não são reconhecidos por critérios jornalísticos rigorosos. Esse recurso muitas vezes se entrelaça com o *uso de fontes duvidosas*, nas quais opiniões são tratadas como fatos. É comum que artigos opinativos sejam apresentados como reportagens objetivas, confundindo o leitor

[3] `www.kaggle.com/datasets/emineyetm/fakenews-detection-datasets`

Tabela 1. Lista consolidada de Red flags.

Red Flag Consolidada	Deepseek	Qwen	Llama	Gemma
Exagero/Sensacionalismo	✓	✓	✓	✓
Falta de Fontes Confiáveis	✓	✓	✓	✓
Linguagem Emocional/Apelativa	✓	✓	✓	✓
Dados Imprecisos ou Vagos	✓	✓	✓	✓
Viés/Narrativa Tendenciosa	✓	✓	✓	✓
Falta de Contexto	✓	✓	✓	✓
Generalizações/Stereótipos	✓	✓	✓	✓
Apelo à Urgência/Medo	✓	✓	✓	✓
Uso de Fontes Duvidosas	–	✓	✓	✓
Contradições Lógicas	–	✓	–	–
Erros Gramaticais/Formais	–	✓	–	✓
Seletividade Factual	✓	–	–	–
Acusação/Responsabilização	✓	–	–	✓
Simplificação Excessiva	✓	–	–	–
Apelo a Teorias da Conspiração	–	✓	–	–
Agenda Política Explícita	–	–	✓	✓

quanto à natureza da informação. A *linguagem emocional ou pejorativa* também é amplamente utilizada, com termos como "Gun Nuts of America" ou "scumbag" que apelam diretamente ao julgamento do leitor, substituindo argumentação por insultos. Em muitos casos, essa linguagem é acompanhada de *viés ou narrativa tendenciosa*, como quando opositores políticos são descritos como "extremistas de direita" ou quando há *agenda política explícita*, como em frases que defendem a expulsão de um partido político do poder. Também são comuns os *dados imprecisos ou vagos*, como no exemplo de "514 misdemeanors", apresentados sem contextualização ou fonte. Essas imprecisões muitas vezes se combinam com *generalizações e estereótipos*, como alegações de que "todos os artistas brancos foram excluídos", ou com *afirmações não comprovadas*, como culpar diretamente um grupo por decisões judiciais ou institucionais sem apresentar provas.

A *falta de contexto* é observada quando informações são apresentadas de forma isolada, como no caso de um "golpe na Turquia" sem menção à data, local ou envolvidos. Essa omissão de detalhes se relaciona diretamente com a *seletividade factual*, na qual apenas partes convenientes da realidade são reportadas, criando uma ilusão de verdade parcial. Alguns textos contêm *contradições lógicas*, como no uso do termo "fascist antifa", no qual o adjetivo contradiz o sujeito. Além disso, há construções que recorrem à *simplificação excessiva*, reduzindo eventos complexos a slogans ou frases de efeito. Isso frequentemente está ligado ao *apelo a teorias da conspiração*, como alegações de silenciamento político ou manipulação cultural sem qualquer evidência. Por fim, a presença de *erros gramaticais ou formais*, como "Turkey s coup", revela falta de revisão ou pressa na publicação, o que é comum em conteúdos fraudulentos. Esses erros, embora sutis, contribuem para enfraquecer a credibilidade e podem servir como sinal de alerta adicional.

4. Processamento de Notícias e Geração de Dataset

Com base na caracterização de *fake news* apresentada na Seção 3, foi implementado um mecanismo de avaliação por meio de um *prompt* estruturado. Esse mecanismo permite analisar cada notícia em função das *red flags* previamente definidas, atribuindo uma pontuação de 0 a 10 conforme a presença e a intensidade das *red flags*.

As quatro LLMs utilizadas na etapa de caracterização de *fake news* foram novamente avaliadas quanto à sua consistência na atribuição de pontuações às *red flags*. Durante essa análise, o modelo *DeepSeek* demonstrou desempenho superior, especialmente em termos de coerência interna e baixa variação entre execuções repetidas. Por essa razão, foi escolhida como o modelo principal para aplicação do *prompt* sobre o conjunto de notícias e geração do *dataset* numérico.

Com a LLM selecionada, procedeu-se à construção do *SBSegFakeNews2025*, um novo conjunto de dados público e estruturado, composto por mil notícias (sendo 500 verdadeiras e 500 falsas), previamente balanceadas para evitar viés de classe. Cada notícia foi processada com base no *prompt* definido, resultando em uma linha vetorial com 16 valores numéricos correspondentes às pontuações atribuídas para cada *red flag*. Além dessas pontuações, foi adicionada uma coluna chamada *temperatura*, calculada como a soma total dos valores atribuídos a cada amostra, refletindo a intensidade geral da linguagem persuasiva presente. A última coluna do *dataset* indicava a classe da notícia (0 para verdadeira e 1 para falsa), totalizando 18 colunas e originando uma matriz de dados de dimensão 1000×18.

5. Experimentação e Resultados

Esta seção apresenta os principais resultados obtidos com a aplicação do classificador *Random Forest* sobre o conjunto de dados proposto. Os dados foram analisados sob duas perspectivas complementares: a performance quantitativa do modelo, com base em métricas tradicionais de classificação supervisionada, e a interpretabilidade dos resultados, utilizando a abordagem SHAP para identificar os atributos de maior impacto.

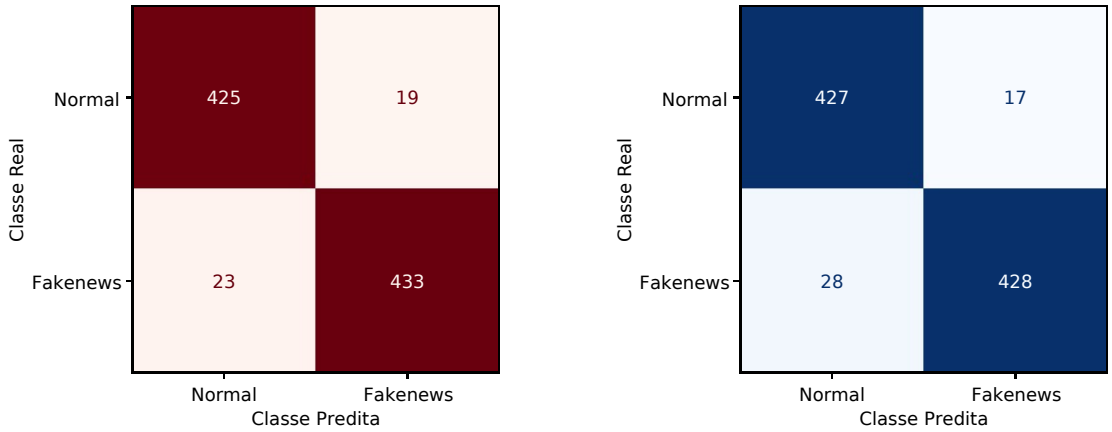
5.1. Classificação de Fakenews usando Red Flags

A Figura 1(a) apresenta a matriz de confusão resultante da classificação de notícias pelo modelo *Random Forest*. No total, foram corretamente classificadas 425 notícias verdadeiras (*True Negatives - TN*) e 433 notícias falsas (*True Positives - TP*). Houve 19 falsos positivos (notícias verdadeiras classificadas como falsas) e 23 falsos negativos (notícias falsas classificadas como verdadeiras). Para fins comparativos, a Figura 1(b) exibe a matriz de confusão sem a presença da *temperatura* computada, onde podem ser observadas variações sutis nos resultados. Com os dados da Figura 1(a), os principais indicadores de desempenho do modelo usando todas as *features* disponíveis foram calculados a seguir: acurácia de 95,33%; Precisão de 95,80%; Revocação de 94,96%; F1-Score de 95,38%.

Esses resultados indicam um desempenho robusto do classificador, com equilíbrio entre precisão e recall, o que é fundamental em cenários onde tanto falsos positivos quanto falsos negativos têm impacto relevante. A Seção 5.2 explica esses resultados.

5.2. Análise de Importância e Frequência de Red Flags

A análise de explicabilidade foi conduzida sobre o conjunto de teste, o qual continha 90% dos dados, uma vez que o modelo foi treinado com os 10% restantes, com foco na classe `fakenews`. As Figuras 2(a) e 2(b) apresentam uma comparação entre a frequência de ocorrência de diferentes *red flags* em notícias e a importância média dessas mesmas *red flags* para a predição de *fake news*, conforme atribuído pelo modelo de explicabilidade SHAP. Observa-se uma diferença significativa entre os dois critérios. Por exemplo, a



(a) Todas as *red flags*.

(b) Sem o uso da *red flag* "temperatura".

Figura 1. Matriz de confusão para classificação das notícias (90% das amostras).

Linguagem Emocional/Pejorativa aparece com alta frequência em ambas as classes de notícias (falsas e verdadeiras) na Figura 2(b), mas tem uma contribuição relativamente baixa para a decisão do modelo na Figura 2(a). Isso sugere que sua presença, embora comum, não é discriminativa o suficiente para diferenciar conteúdos falsos dos verdadeiros.

Por outro lado, *Seletividade Factual*, *Simplificação Excessiva* e *Agenda Política Explícita*, embora menos frequentes (Figura 2(b)), apresentam maior impacto médio no modelo (Figura 2(a)), sendo mais relevantes na detecção da desinformação.

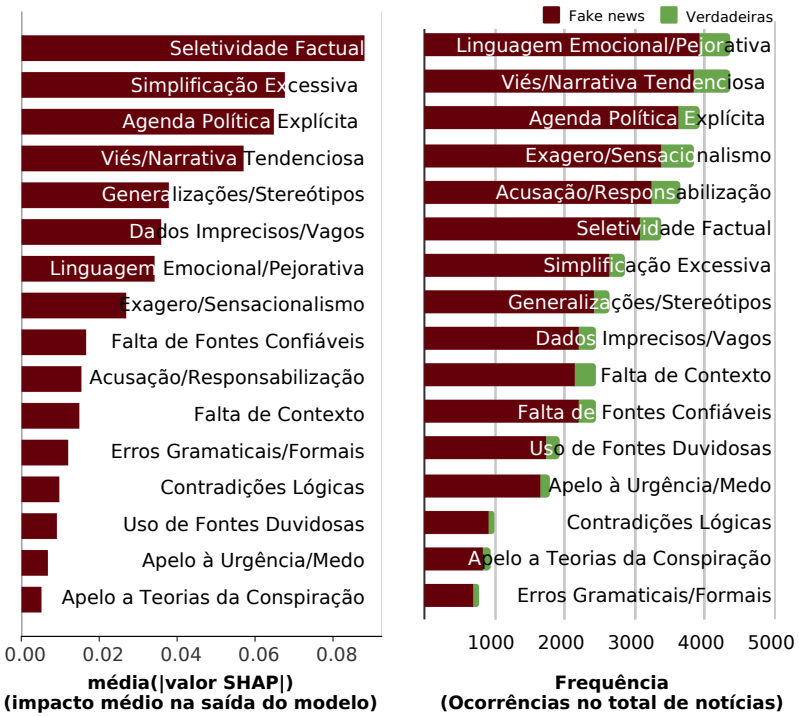
Portanto, a análise conjunta de todas as *red flags* é importante. Características mais sutis, porém semanticamente marcantes, podem ter papel fundamental na detecção automática de *fake news*. Foi observado em um experimento isolado que, se a *temperatura* for considerada, a mesma atinge um valor SHAP médio de 0.16 (o dobro do valor SHAP mais alto sem essa *feature*) e que todas as demais reduzem seus valores (sendo a mais alta com valor médio de 0.07). Portanto, a fim de concentrar esta análise nos indicadores linguísticos isolados, excepcionalmente neste gráfico, a *temperatura* foi removida.

5.3. Análise de Temperatura

Em geral, ao combinar-se as *red flags* propostas, obtém-se importantes informações acerca das *fake news*. Foi observado que notícias verdadeiras possuem uma *temperatura* (soma dessas *red flags*) média de 8,57 (variando de 0 a 88), enquanto as *fake news* possuem uma média de 77,03 (variando de 4 a 114). Portanto, a *temperatura* apresenta-se como o principal indicador de *fake news*, o qual soma 38.513 pontos no total de *fake news* analisadas. No entanto, esse mesmo indicador soma 4.286 em notícias verdadeiras, o que revela que muitas das *red flags* também estão presentes em notícias comuns e que isso, isoladamente, pode levar a interpretações incorretas da veracidade de uma notícia.

6. Conclusão e Trabalhos Futuros

Este trabalho abordou a lacuna existente na detecção de *fake news* por meio de modelos de linguagem, destacando a ausência de explicabilidade nas abordagens atuais. Como solução, foi proposta uma metodologia baseada em engenharia de *prompts* combinada



(a) Summary Bar Plot do SHAP.

(b) Pontuação total por *red flags* (1000 notícias, cada uma pontuando de 0 a 10).

Figura 2. Comparação de valores SHAP com a pontuação total de red flags.

com LLMs modernas para identificar e categorizar indícios linguísticos recorrentes — denominados *red flags*. Esses indicadores foram organizados semanticamente e utilizados para compor um novo *dataset* balanceado, o *SBSegFakeNews2025*, que permitiu o treinamento de classificadores supervisionados com foco em interpretabilidade.

Os experimentos realizados com o classificador *Random Forest* demonstraram alto desempenho, com F1-score de 95,38%, reforçando a efetividade dos *red flags* como atributos discriminativos. A análise com SHAP confirmou a importância de certos padrões discursivos, como a seletividade de fatos. Como principais contribuições, destacam-se a geração de um *dataset* interpretável e reutilizável, a integração estruturada entre categorização semântica e classificação supervisionada, e o uso de LLMs para promover decisões mais transparentes em ambientes de desinformação.

Como trabalhos futuros, propõe-se expandir a abordagem para múltiplos idiomas e contextos culturais, testar modelos adicionais de classificação, como redes neurais explicáveis, e aplicar o método em cenários reais de triagem jornalística ou moderação automatizada. Além disso, pretende-se explorar a interação entre múltiplos *red flags* por meio de modelagens baseadas em grafos e incorporar *feedback* humano no processo de refinamento dos *prompts*, ampliando a robustez e aplicabilidade da solução.

Referências

Anjos, L. S. d. (2023). Análise experimental do desempenho de grandes modelos de linguagem na detecção de notícias falsas. In *Anais do Trabalho de Conclusão de*

Curso - Universidade Federal de Uberlândia, pages 1–10. Universidade Federal de Uberlândia.

Aslam, N., Ullah Khan, I., Alotaibi, F. S., Aldaej, L. A., and Aldubaikil, A. K. (2021). Fake detect: A deep learning ensemble model for fake news detection. *Complexity*.

Baarir, N. F. and Djeffal, A. (2021). Fake news detection using machine learning. *2020 Workshop on Human-Centric Smart Environments for Health and Well-being (IHSH)*.

BBC News (2024). Pesquisa aponta que 25% das pessoas não sabem identificar fake news. `https://www.bbc.com/news/articles/cd1d28xgjp2o`. Acesso em: 14 maio 2025.

Hu, B., Sheng, Q., Cao, J., Shi, Y., Li, Y., Wang, D., and Qi, P. (2024). Bad actor, good advisor: Exploring the role of large language models in fake news detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 22105–22113.

KaabOmeir, F., Khademizadeh, S., Seifadini, R., Balani, S. O., and Khazaneha, M. (2024). Overview of Misinformation and Disinformation Research from 1971 to 2022. *Journal of Scientometric Research*, 13(2):430–447.

Kanashina, O., Huertas García, R., and Jiménez-Zarco, A. I. (2023). Fake news and youngsters' decision journey: An evaluation of the influence of misinformation on social media. *Intangible Capital*, 19(4):534–554.

Khanam, Z., Alwasel, B. N., Sirafi, H., and Rashid, M. (2021). Fake news detection using machine learning approaches. In *IOP Conference Series: Materials Science and Engineering*, volume 1099, page 012040. IOP Publishing.

Khivasara, Y., Khare, Y., and Bhadane, T. (2020). Fake news detection system using web-extension. *2020 IEEE Pune Section International Conference*.

Kong, S. H., Tan, L. M., Gan, K. H., and Samsudin, N. H. (2020). Fake news detection using deep learning. In *2020 IEEE 10th symposium on computer applications & industrial electronics (ISCAIE)*, pages 102–107. IEEE.

Meta & Broadminded (2025). Relatório de percepção sobre fake news no brasil. `https://about.fb.com/pt-br/news/2025/01/pesquisa-meta-fake-news/`. Acesso em: 14 maio 2025.

OCDE (2024). The OECD Truth Quest Survey: Methodology and findings. *OECD Digital Economy Papers*, 369.

Papageorgiou, E., Chronis, C., Varlamis, I., and Himeur, Y. (2024). A survey on the use of large language models (llms) in fake news. *Future Internet*, 16(8):298.

Raza, S. and Ding, C. (2022). Fake news detection based on news content and social contexts: a transformer-based approach. *Int. Jrnl. of Data Science and Analytics*.

Wazlawick, R. S. (2009). *Metodologia de pesquisa para ciência da computação*, volume 2. Elsevier Rio de Janeiro.

Özbay, F. and Alatas, B. (2019). A novel approach for detection of fake news on social media using metaheuristic optimization algorithms. *Elektronika ir Elektrotechnika*.