APLICAÇÃO PARA ANÁLISE E DETECÇÃO DE NOTÍCIAS FALSAS E MANIPULAÇÃO DE MÍDIA

Mateus Santos de Matos[1], Matheus Leandro Ferreira[2]

Resumo: A disseminação de *fake news* representa um dos principais desafios da era digital, comprometendo a confiabilidade das informações compartilhadas em ambientes online. Apesar da existência de ferramentas para verificação de textos, links ou imagens de forma isolada, há carência de soluções que integrem essas modalidades em uma única aplicação de forma prática e acessível. Este trabalho tem como objetivo identificar a veracidade e a confiabilidade de informações divulgadas na internet por meio de uma aplicação web baseada na análise multimodal de textos, links e imagens. A metodologia utilizou inteligência artificial, Processamento de Linguagem Natural e mecanismos de busca. A solução foi avaliada por meio das métricas de acurácia, precisão, *recall* e F1-*score* em cada modalidade, obtendo resultados expressivos, com destaque para a análise textual, que alcançou acurácia e F1-*score* de 0,95. Os resultados demonstram o potencial da inteligência artificial como ferramenta de apoio à verificação de informações e evidenciam a viabilidade da abordagem proposta, integrando diferentes formatos de conteúdo digital em um único ambiente e contribuindo para pesquisas na área de detecção automática de desinformação.

Palavras-chave: fake news; inteligência artificial; verificação de informações; processamento de linguagem natural; análise multimodal.

ABSTRACT: The spread of fake news represents one of the major challenges of the digital era, compromising the reliability of information shared in online environments. Despite the existence of tools for verifying text, links, or images individually, there is a lack of solutions that integrate these modalities into a single, practical, and accessible application. This study aims to identify the veracity and reliability of information published on the Internet through a web application based on the multimodal analysis of text, links, and images. The methodology employed Artificial Intelligence, Natural Lan-

[1]Curso de Ciência da Computação, Universidade do Extremo Sul Catarinense (Unesc), Criciúma - Santa Catarina - Brasil. mcmateus2011@gmail.com

[2]Orientador, Curso de Ciência da Computação, Universidade do Extremo Sul Catarinense (Unesc), Criciúma - Santa Catarina - Brasil. mlf@unesc.net

guage Processing, and search mechanisms. The solution was evaluated using the metrics of accuracy, precision, recall, and F1-score for each modality, achieving significant results, particularly in textual analysis, which reached an accuracy and F1-score of 0.95. The results demonstrate the potential of Artificial Intelligence as a supporting tool for information verification and highlight the feasibility of the proposed approach by integrating different digital content formats into a single environment, thereby contributing to research in the field of automated disinformation detection.

Keywords: fake news; artificial intelligence; information verification; natural language processing; multimodal analysis.

1 INTRODUÇÃO

A disseminação de desinformação na internet tornou-se um dos principais desafios contemporâneos, afetando diretamente a confiança social, a saúde pública e os processos democráticos. O avanço da internet e das redes sociais ampliou significativamente a velocidade de circulação das informações, permitindo que conteúdos verdadeiros e falsos alcancem milhões de pessoas em poucos instantes (Barboza; Servidoni, 2021). As chamadas *fake news*, caracterizadas pela divulgação de informações falsas ou deliberadamente enganosas, passaram a representar uma preocupação crescente para governos, organizações e instituições científicas (Exeter, 2025; Bovet; Makse, 2019).

Embora a desinformação acompanhe a história da humanidade há séculos, sua propagação foi potencializada pelos meios digitais e pelas plataformas de mídia social (Abad, 2019; Anderson et al., 2021). Os impactos desse fenômeno são observados em diferentes contextos sociais, especialmente nas áreas política e da saúde. Em processos eleitorais, notícias falsas podem influenciar a formação da opinião pública e comprometer a qualidade do debate democrático (Allcott; Gentzkow, 2017; Jerit; Zhao, 2020).

Na saúde pública, a pandemia de COVID-19 evidenciou como a circulação de teorias conspiratórias, tratamentos sem comprovação científica e informações pseudocientíficas pode ampliar cenários de crise e dificultar a adoção de medidas preventivas pela população (Naeem; Bhatti; Khan, 2021; WHO, 2022). Esse cenário contribuiu para o fortalecimento da chamada infodemia, caracterizada pelo excesso de informações verdadeiras e falsas circulando simultaneamente, dificultando a identificação de

fontes confiáveis (WHO, 2025; Mahlous, 2024).

As redes sociais desempenham papel central nesse processo. Plataformas como Facebook, Instagram, TikTok, YouTube, WhatsApp e X/Twitter tornaram-se importantes meios de acesso à informação, mas também favoreceram a disseminação de conteúdos enganosos devido à facilidade de compartilhamento e ao alcance de seus mecanismos de recomendação (Tarigan et al., 2023; Chandra; Anita, 2022). Os algoritmos dessas plataformas priorizam conteúdos com maior potencial de engajamento, reforçando crenças pré-existentes e contribuindo para a formação de câmaras de eco e bolhas informacionais (Narayanan, 2023; Cinus et al., 2022; Wang, 2023). Como consequência, conteúdos sensacionalistas ou emocionalmente apelativos tendem a receber maior visibilidade, independentemente de sua veracidade.

Além da desinformação textual, os avanços da inteligência artificial ampliaram as possibilidades de manipulação de conteúdos multimídia. Técnicas de geração sintética permitem criar imagens, áudios e vídeos altamente realistas, conhecidos como *deepfakes*, dificultando a distinção entre conteúdos autênticos e fabricados (Gupta; Kaushik; Kaur, 2025). O uso dessas tecnologias tem sido observado em campanhas políticas, fraudes digitais e manipulação da imagem de figuras públicas, aumentando a complexidade dos processos de verificação de informações (Al-Khazraji et al., 2023; Debroy; Hemmige, 2024).

Diante desse cenário, a checagem de fatos (*fact-checking*) consolidou-se como uma das principais estratégias de combate à desinformação (Graves; Amazeen, 2019). Organizações como Lupa, Aos Fatos e a International Fact-Checking Network desempenham papel relevante na validação de conteúdos e na promoção da educação midiática (Lupa, 2025; AosFatos, 2025; IFCN, 2025).

Entretanto, a verificação manual apresenta limitações relacionadas ao tempo de análise, à necessidade de especialistas e ao grande volume de informações produzidas diariamente (Graves, 2017; Quelle et al., 2025). Nesse contexto, a inteligência artificial tem se destacado como alternativa promissora para automatizar processos de verificação, utilizando técnicas de Processamento de Linguagem Natural (PLN), aprendizado de máquina, aprendizado profundo e visão computacional para identificar padrões associados à desinformação (Guetterman et al., 2018; Saeidnia et al., 2025).

Na análise textual, algoritmos como Support Vector Machine

(SVM), Naive Bayes, Random Forest e redes neurais profundas vêm apresentando resultados relevantes na classificação automática de notícias falsas (Nasteski, 2017; Sarker, 2021; Zhang et al., 2024). Paralelamente, técnicas baseadas em redes neurais convolucionais (CNNs) e busca reversa de imagens têm sido empregadas para identificar manipulações visuais, incluindo práticas de *copy-move*, *splicing* e *deepfakes* (Nguyen et al., 2022; Zanardelli et al., 2023; Gaillard; Egyed-Zsigmond, 2017; Basarslan; Kayaalp et al., 2020).

Com o objetivo de compreender o estado da arte relacionado à detecção de desinformação e manipulação de mídia, foram analisadas pesquisas correlatas relevantes para esta investigação. Monteiro et al. (2018) desenvolveram o Fake.Br Corpus, um dos primeiros conjuntos de dados brasileiros voltados à detecção automática de notícias falsas, obtendo resultados de até 96% de acurácia em experimentos de aprendizado de máquina.

Chavarro et al. (2023), propuseram o FakeTrueBR, um conjunto de dados de 3.582 notícias em português, balanceado entre falsas e verdadeiras, coletadas entre 2017 e 2023 e alinhadas semanticamente por meio do modelo Sentence-BERT. O software foi avaliado com técnicas de representação textual como BoW e TF-IDF combinadas com classificadores tradicionais, obtendo F1-*score* de 0,945 com o classificador MLP.

Os autores Sallami e Aïmeur (2025) criaram o Aletheia, uma extensão para o navegador Google Chrome que utiliza Recuperação Aumentada por Geração (RAG) e Modelos de Linguagem de Grande Escala (LLMs). O objetivo geral da pesquisa foi detectar *fake news* e fornecer explicações baseadas em evidências recuperadas via busca na web, classificando notícias como verdadeiras, falsas ou com informações insuficientes, atingindo F1-*score* de 0,85 no *dataset* PolitiFact.

Apesar dos avanços observados na literatura, ainda existe uma lacuna relacionada à integração de diferentes mecanismos de verificação em uma única plataforma. Grande parte das soluções concentra-se exclusivamente na análise textual, na checagem de fatos ou na identificação de manipulações visuais, além de depender de bases de dados específicas e contextos restritos de aplicação (Hangloo; Arora, 2021; Ali et al., 2022). Dessa forma, torna-se relevante investigar abordagens multimodais capazes de combinar diferentes fontes de evidência e ampliar a confiabilidade dos resultados obtidos.

A relevância desta pesquisa está associada à necessidade de

disponibilizar ferramentas acessíveis para auxiliar usuários na avaliação da confiabilidade de conteúdos digitais, contribuindo para a redução dos impactos causados pela desinformação e pela manipulação de mídia (Rani; Das; Bhardwaj, 2022; Nascimento et al., 2022; Fátima; Miranda, 2022).

De acordo com o supracitado, este trabalho tem como objetivo geral identificar a veracidade e a confiabilidade de informações divulgadas na internet por meio de uma aplicação que realize a análise de dados provenientes de links, textos e imagens. Para alcançar esse objetivo geral, foram definidos os seguintes objetivos específicos: Definir os métodos para busca, análise e comparação de dados no sistema; Classificar o nível de confiabilidade das tecnologias estudadas; Aplicar os métodos em uma aplicação web para analisar informações e indicar sua veracidade; Verificar a acurácia dos resultados obtidos na aplicação; Avaliar as contribuições e limitações da aplicação em relação às tecnologias recentes.

Para atender aos objetivos propostos, a solução desenvolvida integra técnicas de Processamento de Linguagem Natural, mecanismos de busca e comparação de conteúdos, análise de imagens e consultas a bases externas de fact-checking. Dessa forma, busca fornecer aos usuários um parecer fundamentado sobre a autenticidade das informações analisadas.

2 MATERIAIS E MÉTODOS

A presente pesquisa caracteriza-se como aplicada, exploratória, descritiva e experimental, adotando abordagem mista, com integração de procedimentos quantitativos e qualitativos.

A natureza aplicada decorre do desenvolvimento de uma aplicação web destinada à verificação automática da veracidade de conteúdos digitais. O caráter exploratório está relacionado à investigação do uso de modelos de inteligência artificial generativa na identificação de possíveis desinformações em diferentes formatos de entrada, enquanto o aspecto descritivo busca registrar e analisar o comportamento do sistema durante os testes realizados.

O método adotado foi o dedutivo, partindo de fundamentos teóricos consolidados sobre desinformação, *fact-checking*, inteligência artificial e análise de conteúdo digital. Com base nesses referenciais, foi desenvolvida uma solução computacional capaz de analisar textos, links e imagens, permitindo posteriormente verificar se os resultados obtidos corroboram com os encontrados descritos na literatura.

A estratégia metodológica utilizada foi a experimentação com-

putacional. A aplicação desenvolvida foi submetida a diferentes cenários de uso, possibilitando observar seu comportamento, registrar os resultados produzidos e avaliar sua capacidade de fornecer respostas consistentes para os conteúdos analisados.

2.1 AMOSTRAS UTILIZADAS NA PESQUISA

Para a validação experimental, foram selecionadas um conjunto de dados composto por 60 amostras balanceadas entre 10 conteúdos verdadeiras e 10 falsos para cada modalidade.

Os textos analisados, classificados como verdadeiros ou falsos, foram obtidos prioritariamente em sites de notícias, portais de informação e perfis de redes sociais brasileiras. Da mesma forma, as URLs utilizadas nos experimentos foram selecionadas a partir dessas mesmas fontes, buscando representar diferentes tipos de conteúdos compartilhados em ambientes digitais.

Em relação às imagens selecionadas como verdadeiras, não foram estabelecidas restrições quanto à sua procedência. O conjunto de dados incluiu fotografias registradas por dispositivos móveis, imagens disponíveis na internet, publicadas em sites de notícias e compartilhadas em redes sociais nacionais e internacionais, de modo a representar diferentes contextos e cenários.

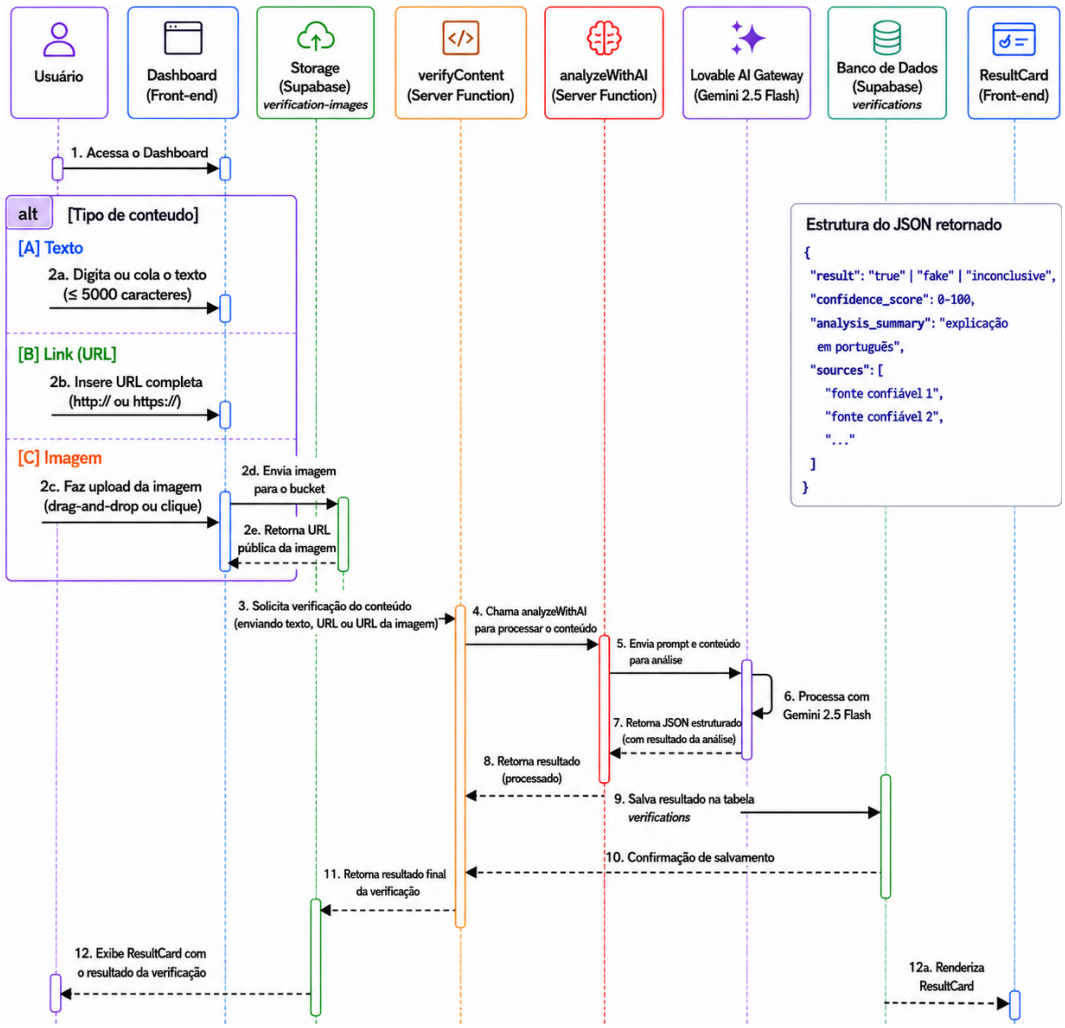
Para a avaliação de imagens manipuladas, foram utilizadas imagens alteradas obtidas na internet, provenientes de redes sociais, bancos de dados especializados em imagens falsas e conteúdos gerados por modelos de inteligência artificial. Essa diversidade de fontes permitiu avaliar a aplicação em cenários distintos e mais próximos das situações encontradas no uso cotidiano.

2.2 ARQUITETURA DA SOLUÇÃO

Foi desenvolvida uma aplicação web para verificação automática de conteúdos digitais, capaz de processar textos, URLs e imagens submetidos pelos usuários. A arquitetura da solução segue o modelo cliente-servidor, sendo composta por uma camada de apresentação (*front-end*), uma camada de processamento das requisições e uma camada de persistência de dados.

A Figura 1 apresenta a arquitetura geral da aplicação. Inicialmente, o usuário acessa o *dashboard* do sistema e seleciona o tipo de conteúdo que deseja verificar.

Figura 1 - Diagrama de Sequência da Verificação de Conteúdo



Fonte: Do autor.

O usuário seleciona a modalidade de verificação, sendo texto, URL ou imagem e envia o conteúdo para o sistema. Inicialmente, a aplicação realiza a validação dos dados de entrada, verificando se as informações fornecidas estão no formato esperado. No caso de imagens, o arquivo é armazenado temporariamente em um bucket do Supabase Storage, que gera uma URL pública utilizada durante o processo de análise.

Após a validação, a requisição é encaminhada ao modelo Gemini 2.5 Flash por meio do Lovable AI Gateway, juntamente com um *prompt* que orienta o modelo a realizar a verificação do conteúdo. O modelo realiza a análise do conteúdo e retorna uma resposta estruturada em formato JSON, contendo a classificação da informação como verdadeira, falsa ou inconclusiva, um índice de confiança, uma justificativa textual e as referências utilizadas durante a análise.

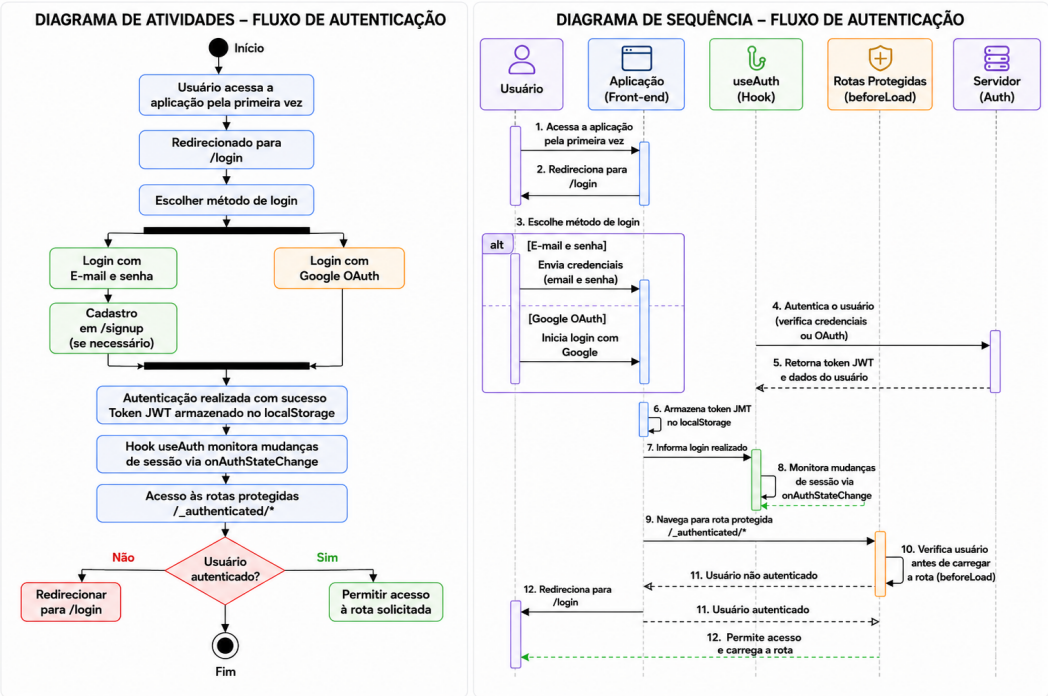
Em seguida, a aplicação realiza uma nova validação da resposta recebida, garantindo que os dados estejam no formato esperado antes de processá-los. Posteriormente essa etapa, o resultado da verificação é armazenado no banco de dados, associado ao usuário autenticado, e apresentado na interface da aplicação por meio de um card que exibe o veredito da análise, o índice de confiança, a justificativa e as fontes utilizadas. Finalizado o processo, a verificação permanece disponível no histórico do usuário para consultas futuras.

2.3 PROCESSO DE AUTENTICAÇÃO

O controle de acesso foi implementado utilizando os serviços de autenticação do Supabase Auth. O sistema permite autenticação por meio de credenciais tradicionais (e-mail e senha) ou utilizando contas Google através do protocolo OAuth 2.0.

Conforme ilustrado na Figura 2, quando o usuário acessa a aplicação pela primeira vez, é redirecionado para a página de autenticação. Após a validação das credenciais, o Supabase gera um token JWT utilizado para gerenciamento da sessão.

Figura 2 - Diagramas de Ativadede e Sequência do Fluxo de Autenticação



Fonte: Do autor.

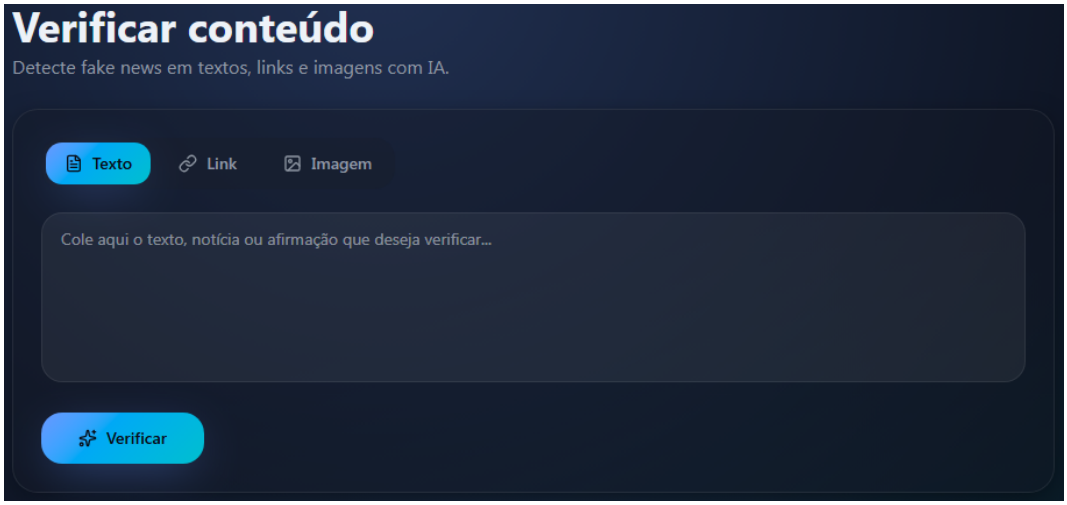
O monitoramento contínuo do estado da autenticação é realizado por meio do hook useAuth, responsável por acompanhar alterações

de sessão e controlar o acesso às rotas protegidas da aplicação. Usuários autenticados podem acessar os recursos internos do sistema, enquanto usuários não autenticados são redirecionados automaticamente para a tela de login.

2.4 PROCESSO DE VERIFICAÇÃO DE CONTEÚDO

O processo de verificação inicia-se após o envio do conteúdo pelo usuário. Conforme mostra a Figura 3, da tela inicial, a aplicação suporta três tipos de entrada, texto, link (URL) e imagem.

Figura 3 - Tela Inicial



Fonte: Do autor.

Quando o conteúdo é submetido, uma função de processamento recebe os dados e encaminha a solicitação para o módulo de análise baseado em inteligência artificial.

A análise é realizada pelo modelo Gemini 2.5 Flash, que utiliza capacidades multimodais para interpretar conteúdos textuais, endereços eletrônicos e imagens. Após o processamento das informações fornecidas pelo usuário, o modelo gera uma resposta estruturada contendo a classificação do conteúdo como verdadeiro, falso ou inconclusivo, acompanhada de um índice de confiança que indica o grau de certeza da avaliação realizada. Sua resposta também retorna uma justificativa detalhada para a classificação atribuída e apresenta as fontes utilizadas durante o processo de análise, permitindo maior transparência e rastreabilidade dos resultados obtidos.

Após a obtenção da resposta, os resultados são armazenados na base de dados da aplicação para fins de consulta posterior e histórico

de verificações.

É possível conferir o código do funcionamento da pesquisa conforme demonstrado na Figura 4, do código da função de busca com IA.

Figura 4 - Código da função de busca com IA

```
async function analyzeWithAI(payload: {
  input_type: "text" | "url" | "image";
  content?: string | null;
  image_url?: string | null;
}): Promise<AnalysisResult> {
  const LOVABLE_API_KEY = process.env.LOVABLE_API_KEY;
  if (!LOVABLE_API_KEY) throw new Error("LOVABLE_API_KEY ausente");

  const systemPrompt = `Você é um especialista em fact-checking. Analise o conteúdo fornecido e determine se é:
- "true" (informação verdadeira e verificada)
- "fake" (provável fake news, desinformação)
- "inconclusive" (sem informação suficiente)

Retorne SEMPRE um JSON válido no formato:
{
  "result": "true" | "fake" | "inconclusive",
  "confidence_score": 0-100,
  "analysis_summary": "explicação concisa em português (2-4 frases)",
  "sources": [{ "title": "nome da fonte", "url": "https://..." }]
}
```

Fonte: Do autor.

O módulo de verificação automática de informações foi implementado por meio da função `analyzeWithAI`, responsável por intermediar a comunicação entre a aplicação e o modelo de linguagem utilizado para a análise de *fact-checking*. A função recebe um objeto contendo o tipo de conteúdo a ser verificado de texto, link ou imagem juntamente com o respectivo conteúdo ou URL correspondente. Antes de efetuar a chamada à API externa, o sistema realiza a validação da chave de autenticação, garantindo que a requisição só seja processada caso as credenciais necessárias estejam corretamente configuradas no ambiente da aplicação, o que evita falhas de segurança e chamadas não autorizadas ao serviço de IA.

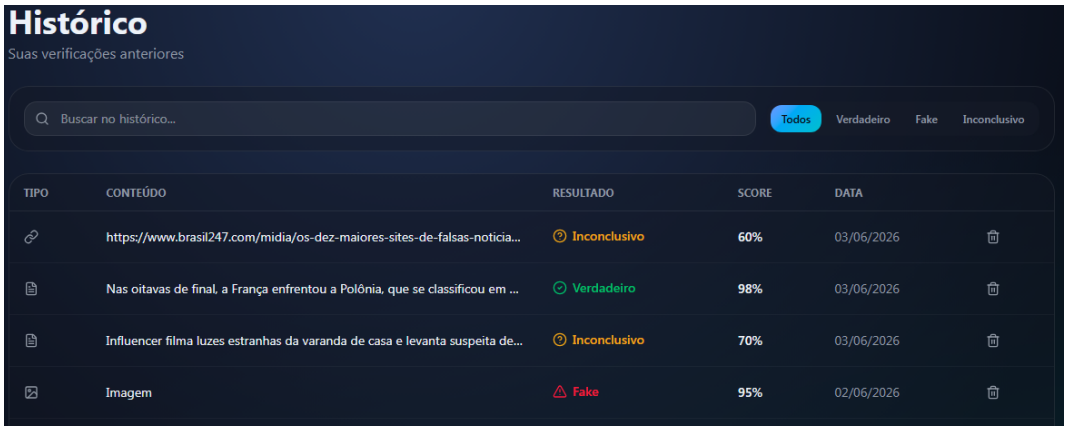
Para orientar o comportamento do modelo, foi definido um *prompt* de sistema que atribui à IA o papel de especialista em checagem de fatos, instruindo-a a classificar o conteúdo analisado em uma das três categorias possíveis, verdadeiro, falso ou inconclusivo, quando não há evidências suficientes para uma conclusão. Além da classificação, o *prompt* exige que a resposta seja retornada em um formato estruturado JSON, contendo o grau de confiança da análise, um resumo explicativo em linguagem natural e as fontes utilizadas como base para a verificação.

2.5 HISTÓRICO DE CONSULTA

Além das funcionalidades de verificação de conteúdo, a aplicação disponibiliza uma aba denominada Histórico, na qual o usuário pode

consultar todas as verificações realizadas anteriormente em sua conta, conforme demostra a Figura 5.

Figura 5 - Aba Histórico



Fonte: Do autor.

A funcionalidade de histórico permite ao usuário consultar todas as verificações realizadas anteriormente em sua conta. É possível realizar pesquisas por meio do título ou conteúdo da verificação, assim como seus resultados podem ser filtrados de acordo com sua classificação, sendo elas verdadeiro, falso ou inconclusivo.

Para cada registro, são exibidas informações como o tipo de conteúdo analisado, o conteúdo verificado, o resultado da análise acompanhado do respectivo índice de confiança e a data em que a verificação foi realizada. Adicionalmente, a interface disponibiliza a opção de exclusão individual dos registros, possibilitando ao usuário gerenciar seu histórico de consultas de forma simples e organizada.

2.6 TECNOLOGIAS UTILIZADAS

O desenvolvimento da aplicação foi realizado utilizando o framework TanStack Start, baseado no React 19, que fornece suporte à renderização híbrida por meio das estratégias Server-Side Rendering (SSR) e Static Site Generation (SSG), contribuindo para melhorias de desempenho e otimização da experiência do usuário.

No processo de compilação e empacotamento da aplicação foi realizado utilizando o Vite 7, ferramenta responsável por disponibilizar um ambiente de desenvolvimento moderno, com inicialização rápida e atualização eficiente dos módulos durante a execução do projeto.

A camada de interface foi desenvolvida utilizando React 19 associado ao TanStack Router para gerenciamento das rotas da aplicação. O

sistema de roteamento baseado em arquivos permitiu a organização estruturada das páginas e o controle de acesso às rotas protegidas da aplicação.

A estilização foi realizada por meio do Tailwind CSS v4, complementado pela utilização de Design Tokens em CSS para padronização visual da interface. Essa abordagem possibilitou a manutenção consistente de cores, tipografia, espaçamentos e demais elementos visuais ao longo de toda a aplicação.

Os componentes de interface foram implementados utilizando a biblioteca shadcn/ui, que disponibiliza componentes reutilizáveis e personalizáveis, permitindo a construção de interfaces modernas e consistentes alinhadas à identidade visual do sistema. Os ícones utilizados na interface foram fornecidos pela biblioteca Lucide React, empregada para a representação visual de ações, funcionalidades e estados do sistema, contribuindo para a usabilidade da aplicação.

Notificações exibidas ao usuário foram implementadas utilizando a biblioteca Sonner, responsável pela apresentação de mensagens de feedback relacionadas a operações realizadas no sistema, como autenticação, envio de conteúdo para análise, sucesso de operações e tratamento de erros.

Na comunicação assíncrona entre a interface e os serviços foi implementada utilizando TanStack Query, permitindo o gerenciamento eficiente de requisições, cache, sincronização e atualização dos dados consumidos pela aplicação.

Para autenticação, armazenamento de arquivos e persistência de dados, foi utilizada a plataforma Supabase, integrada ao ambiente Lovable Cloud. A autenticação dos usuários foi realizada por meio do Supabase Auth, utilizando os métodos de login por e-mail e autenticação federada através do Google OAuth.

O armazenamento permanente das informações foi realizado em banco de dados PostgreSQL disponibilizado pela infraestrutura do Supabase, garantindo a persistência e a integridade dos dados utilizados pela aplicação. Já o armazenamento de imagens enviadas pelos usuários foi realizado por meio do Supabase Storage, serviço responsável pelo gerenciamento de arquivos utilizados durante o processo de verificação de conteúdo multimodal. Por fim, a análise dos conteúdos foi realizada por meio do modelo de inteligência artificial Gemini 2.5 Flash, acessado através do Lovable AI Gateway, responsável pela intermediação das requisições entre a aplicação e o modelo generativo. Essa arquitetura permitiu a análise

de conteúdos textuais, endereços eletrônicos e imagens submetidos pelos usuários.

3 RESULTADOS E DISCUSSÃO

Na presente seção são apresentados os resultados obtidos com a aplicação VerificaAI, uma ferramenta de verificação de conteúdos falsos. O objetivo é validar a eficácia e o desempenho da aplicação por meio de testes automatizados com diferentes fontes de dados, categorizadas em três tipos diferentes texto, links e imagens.

3.1 MÉTRICAS DE AVALIAÇÃO

O desempenho da aplicação foi avaliado por meio da matriz de confusão, uma estrutura que permite comparar os valores reais com os valores preditos pelo modelo.

A partir dessa matriz, foram calculados quatro métricas de desempenho: acurácia, precisão, *recall* e F1-*score*. A acurácia representa a proporção total de classificações corretas realizadas pelo modelo. A precisão mede a proporção de notícias classificadas como falsas que realmente eram falsas. O *recall* indica a capacidade do modelo de identificar corretamente as notícias falsas presentes no conjunto de dados. Já o F1-*score* corresponde à média entre precisão e *recall*, fornecendo uma métrica especialmente útil quando se deseja avaliar simultaneamente a capacidade do modelo de identificar notícias falsas e de evitar classificações incorretas.

Vale destacar que a ferramenta pode retornar três resultados distintos de análise, fake, verdadeiro ou inconclusivo. Como o conjunto de teste é composto exclusivamente por amostras previamente rotuladas como falsas ou verdadeiras, os resultados “inconclusivos” foram considerados erros de classificação. Assim, foram contabilizados como Falsos Negativos quando a amostra real era falsa e como Falsos Positivos quando a amostra real era verdadeira. Essa abordagem foi adotada porque o objetivo da aplicação consiste em fornecer uma classificação definitiva para cada notícia analisada, sendo as respostas inconclusivas interpretadas como falhas no processo de classificação para o cálculo das métricas de desempenho.

Na Tabela 1 é possível observar os resultados obtidos com os 60 conjuntos de dados utilizados, sendo organizados dentro das métricas para os cálculos da matriz de confusão.

Tabela 1 - Resultados das métricas de avaliação por conteúdo

Conteúdo	VP	VN	FP	FN	Acurácia	Precisão	Recall	F1-Score
Texto	10	9	1	0	0,95	0,91	1,00	0,95
Link	6	10	0	4	0,80	1,00	0,60	0,75
Imagem	8	8	2	2	0,80	0,80	0,80	0,80

Fonte: Do autor.

Para o cálculo das métricas de avaliação, as notícias falsas foram definidas como a classe positiva, uma vez que o objetivo principal da aplicação é identificar conteúdos potencialmente falsos. Com essa definição, os quadrantes da matriz de confusão assumem o seguinte significado: os Verdadeiros Positivos (VP) representam as notícias falsas corretamente identificadas pelo modelo; os Verdadeiros Negativos (VN) correspondem às notícias verdadeiras corretamente classificadas; os Falsos Positivos (FP) indicam notícias verdadeiras erroneamente classificadas como falsas; e os Falsos Negativos (FN) representam notícias falsas que o modelo falhou em detectar.

3.2 CONDUÇÃO DE EXPERIMENTOS E PERFORMANCE DA APLICAÇÃO

A Tabela 2, demostra o desempenho da verificação por cada tipo de conteúdo diferente, podendo ser observada e comparada cada um dos resultados obtidos, para obtenção das métricas de performance. Os resultados obtidos a partir da avaliação da aplicação foram organizados por tipo de conteúdo texto, link e imagem, utilizando as métricas derivadas da matriz de confusão: acurácia, precisão, *recall* e F1-*score*.

Tabela 2 - Desempenho da verificação por tipo de conteúdo

Tipo de Conteúdo	Acurácia	Precisão	Recall	F1-Score
Texto	0,95	0,91	1,00	0,95
Link	0,80	1,00	0,60	0,75
Imagem	0,80	0,80	0,80	0,80

Fonte: Do autor.

O modelo apresentou seu melhor desempenho geral na classificação de conteúdo em formato de texto, atingindo acurácia de 0,95 e F1-*score* de 0,95, valores que indicam um desempenho sólido e equilibrado. O *recall* de 1,0 revela que os textos retirados de *fake news* foram corretamente identificadas pelo modelo. A precisão de 0,91, ligeiramente inferior, indica que uma pequena parcela dos trechos retirados das notícias verdadeiras foi classificada incorretamente como falsa, resultado diretamente relacionado

ao caso inconclusivo registrado durante os testes, que foi tratado como erro de classificação.

Nos textos verdadeiros, o veredito retornado pela ferramenta foi consistente com a classificação esperada. Já nos textos falsos, a aplicação foi capaz de identificar a desinformação e apresentar justificativas que corrigiam ou contextualizavam as informações incorretas presentes no conteúdo analisado.

Na avaliação de conteúdo em formato de link, o modelo demonstrou comportamento assimétrico entre as métricas. A precisão atingiu valor máximo de 1,0, sinalizando que nenhuma notícia verdadeira foi classificada incorretamente como falsa nessa modalidade. No entanto, o *recall* de 0,6 indica que uma parcela relevante das *fake news* não foi detectada, com 4 notícias falsas retornando resultado inconclusivo e sendo tratadas como erro de classificação, esse comportamento reflete uma necessidade de aprimoramentos no modelo na detecção de links falsos, priorizando uma maior taxa de detecção de links de notícias falsas. O F1-*score* de 0,75 confirma esse desequilíbrio entre precisão e *recall*.

Para conteúdo em formato de imagem, o modelo apresentou desempenho uniforme entre todas as métricas, com acurácia, precisão, *recall* e F1-*score* todos iguais a 0,8. Esse equilíbrio sugere que os erros de classificação foram distribuídos de forma homogênea entre falsos positivos e falsos negativos, sem tendência de viés para nenhuma das classes e que o modelo possui consistência e se comporta de maneira previsível na análise de imagens.

De modo geral, o modelo apresentou desempenho preliminar nas três modalidades avaliadas, com destaque para a classificação de textos. A queda de desempenho observada em links e imagens revelam melhorias a serem feitas na ferramenta.

No caso dos links, a alta precisão associada ao baixo *recall* sugere que o modelo tende a ser conservador, classificando como falso apenas quando há maior grau de certeza, comportamento que reduz os falsos alarmes, mas permite que algumas *fake news* passassem despercebidas.

Para imagens, o modelo apresentou desempenho uniforme entre todas as métricas, com valores iguais a 0,8, o que demonstra uma classificação equilibrada entre as classes sem tendência de viés.

3.3 OBSERVAÇÕES ADICIONAIS APÓS VEREDITO DAS ANÁLISES

Nos textos classificados como verdadeiros, o veredito retornado pela ferramenta apresentou concordância com a classificação esperada, demonstrando consistência na avaliação dos conteúdos analisados.

Já nos textos classificados como falsos, a aplicação foi capaz de identificar a presença de desinformação e apresentar justificativas que contextualizavam as informações incorretas presentes no conteúdo analisado. Além disso, as explicações fornecidas contribuíram para uma melhor compreensão dos motivos que levaram à classificação do texto como falso.

Observou-se que a ferramenta utiliza a confiabilidade da fonte como principal critério de avaliação. Links provenientes de portais jornalísticos e perfis oficiais tendem a ser classificados como verdadeiros, enquanto conteúdos oriundos de fontes desconhecidas ou de credibilidade duvidosa geralmente recebem classificação falsa ou inconclusiva.

Além disso, conteúdos que possuíam vídeo não puderam ser avaliados de forma adequada, pois a aplicação não realiza interpretação audiovisual.

Durante a seleção das amostras, verificou-se que grande parte das notícias falsas estava associada às redes sociais, principalmente em formatos de vídeo e imagem acompanhados de texto.

Na análise de imagens, a aplicação demonstrou capacidade de identificar sinais de manipulação digital, elementos visuais suspeitos e padrões característicos de conteúdos gerados por inteligência artificial. Imagens contendo montagens evidentes, informações descontextualizadas ou características visuais inconsistentes foram classificadas corretamente como falsas.

Em imagens geradas por IA que retratam situações cotidianas, como atividades domésticas, esportivas ou de lazer, frequentemente receberam classificação verdadeira ou inconclusiva, devido à ausência de elementos que indicassem desinformação.

3.4 COMPARAÇÃO COM OS TRABALHOS CORRELATOS

Após coleta de resultados, foi realizada uma análise comparativa entre a ferramenta devolvida e os trabalhos correlatos de Monteiro et al. (2018), Chavarro et al. (2023) e Sallami e Aïmeur (2025). Com o intuito de compreender melhor as características da solução proposta e compará-la visualmente com as abordagens já existentes na literatura, a Tabela 3 apresenta uma comparação entre o trabalho desenvolvido e os estudos

correlatos.

Tabela 3 - Comparativo entre o trabalho proposto e correlatos

Critério	Monteiro et al.(Fake.Br, 2018)	Chavarro et al.(FakeTrueBR, 2023)	Sallami e Aïmeur (Aletheia, 2025)	Este Trabalho
Solução	Dataset e plataforma de consulta online	Dataset atualizado para treinamento e avaliação	Extensão para navegador	Aplicação web
Dados Utilizados	7.200 notícias(3.600 falsas e 3.600 verdadeiras)	3.582 notícias(1.791 falsas e 1.791 verdadeiras)	Conjunto Polifact e fontes obtidas via busca na web	Conjunto de 60 dados de textos, links e imagens divididos igualmente entre verdadeiro e falso
Tecnologias Principais	SVM e Bag-of-Words	TF-IDF e Sentence-BERT	LLM, RAG e Google Fact Check	Gemini 2.5 Flash
Modalidades Analisadas	Texto	Texto	Texto	Texto, Links e Imagens
Principal Contribuição	Primeiro dataset brasileiro de fake news para pesquisa e classificação automática	Dataset atualizado com alinhamento semântico entre notícias falsas e verdadeiras	Integração de verificação, explicação e participação comunitária em uma única ferramenta	Verificação multimodal com score de confiança e veredito

Fonte: Adaptado de Monteiro et al. (2018), Chavarro et al. (2023), Sallami e Aïmeur (2025).

Os trabalhos correlatos analisados apresentam abordagens distintas para o combate à desinformação. Os estudos de Monteiro et al. (2018) e Chavarro et al. (2023). concentram-se na construção de *datasets* e na avaliação de modelos de classificação automática, contribuindo para o desenvolvimento de bases de dados que servem de suporte para pesquisas e aplicações futuras na área de detecção de *fake news*. Já o trabalho de Sallami e Aïmeur (2025) aproxima-se da proposta desta pesquisa ao apresentar uma aplicação voltada ao usuário final, utilizando modelos de linguagem e mecanismos de busca para justificar suas classificações.

Em comparação, o presente trabalho diferencia-se por adotar uma abordagem multimodal, permitindo a análise de textos, links e imagens em uma única plataforma. Além de fornecer um veredito sobre o conteúdo analisado, a aplicação apresenta justificativas e níveis de confiança

para os resultados obtidos. Outra característica relevante é sua acessibilidade, dispensando a instalação de extensões, o treinamento de modelos ou a configuração de plugins adicionais. Os resultados obtidos na análise textual também demonstraram desempenho comparável aos principais trabalhos da literatura, alcançando F1-*score* de 0,95 na análise de textos.

3.5 LIMITAÇÕES

Apesar dos resultados preliminares obtidos durante a avaliação da aplicação, algumas limitações devem ser consideradas. A primeira refere-se à quantidade reduzida de amostras utilizadas nos testes, o que pode limitar a representatividade dos resultados alcançados. Dessa forma, a utilização de uma base de dados mais ampla e diversificada poderia fornecer uma avaliação mais robusta e precisa do desempenho da ferramenta.

Outra limitação está relacionada à dependência da base de conhecimento utilizada pelo modelo de inteligência artificial. Como a aplicação realiza suas verificações com base em informações disponíveis até o ano de 2024, conteúdos mais recentes podem apresentar justificativas imprecisas ou vereditos incorretos devido à ausência de informações atualizadas durante o processo de análise. Além disso, algumas referências utilizadas na verificação podem tornar-se indisponíveis ao longo do tempo, uma vez que notícias podem ser removidas, atualizadas ou transferidas para outros endereços pelos próprios sites de origem. Essa restrição afeta principalmente a verificação de textos, links e imagens associados a eventos recentes.

Além disso, a aplicação não realiza interpretação direta de conteúdos audiovisuais, como vídeos, limitando sua capacidade de análise em plataformas nas quais esse formato é predominante. Considerando o crescimento da disseminação de desinformação por meio de vídeos e conteúdo multimídia, a incorporação de mecanismos específicos para análise audiovisual representa uma oportunidade para trabalhos futuros.

4 CONCLUSÃO

Para sanar o problema relacionada a desinformação de diferentes tipos de conteúdos na internet, foi desenvolvida uma ferramenta com foco na acessibilidade e na facilidade de uso, disponibilizando uma interface simples e intuitiva para a realização das análises com o apoio da inteligência artificial. Além do veredito sobre o conteúdo analisado, a aplicação apresenta um *score* de confiança e justificativas textuais que auxiliam na

interpretação dos resultados. Adicionalmente, foi implementado um mecanismo de histórico de consultas, permitindo o acompanhamento das análises realizadas pelos usuários.

A avaliação da aplicação foi conduzida por meio da análise de 60 amostras de conteúdo, sendo 20 de cada modalidade de texto, URL e imagem, distribuídas igualmente entre conteúdos verdadeiros e falsos. Com base nos resultados obtidos foram calculadas as métricas de acurácia, precisão, *recall* e F1-*score* utilizando a matriz de confusão, possibilitando avaliar o desempenho da aplicação e a confiabilidade dos resultados produzidos.

Os resultados obtidos por meio dessas métricas demonstraram um desempenho preliminar da aplicação no geral, com destaque para a modalidade de análise textual, que alcançou valores de 0,95 para a acurácia e F1-*score*, evidenciando o potencial da abordagem proposta para auxiliar na identificação de possíveis conteúdos falsos.

Dessa forma, o objetivo geral deste trabalho foi alcançado ao desenvolver uma aplicação web capaz de verificar conteúdos em formato de texto, URL e imagem por meio da utilização de inteligência artificial. Além disso, os objetivos específicos foram atendidos ao definir os métodos de processamento e análise das informações, aplicar esses métodos em uma aplicação web, avaliar o desempenho por meio de métricas de classificação e identificar as principais contribuições e limitações da solução desenvolvida.

Como contribuição para a área, este trabalho desenvolveu uma ferramenta capaz de centralizar a análise de textos, links e imagens em um único ambiente, ampliando as possibilidades de verificação de conteúdos digitais. Os resultados demonstram o potencial da inteligência artificial como ferramenta de apoio à identificação de possíveis desinformações e reforçam a viabilidade de sua aplicação em processos de verificação de informações.

No contexto acadêmico, a aplicação desenvolvida e os resultados apresentados podem servir como base para pesquisas futuras relacionadas à detecção automática de *fake news* e à verificação de diferentes formatos de conteúdo digital.

Apesar dos resultados obtidos, o trabalho apresenta algumas limitações. A quantidade reduzida de amostras utilizadas nos testes limita a representatividade dos resultados, sendo recomendada a utilização de bases de dados mais amplas em avaliações futuras. Além disso, a preci-

são das análises depende do conhecimento disponível no modelo de inteligência artificial, podendo haver limitações na verificação de conteúdos recentes. Por fim, a aplicação não possui suporte à análise de vídeos, restringindo a verificação a conteúdos em formato de texto, URL e imagem.

Como trabalhos futuros, pretende-se aperfeiçoar os mecanismos de análise por meio da utilização de bases de dados mais amplas e atualizadas, da adoção de técnicas avançadas de validação de informações e da incorporação de suporte à análise de vídeos e outros formatos multimídia.

Por fim, os resultados alcançados indicam que a abordagem proposta é viável para auxiliar na verificação de informações em ambientes digitais. Assim sendo, a pesquisa contribui tanto para o avanço das pesquisas sobre a detecção automática de conteúdos falsos quanto para o desenvolvimento de ferramentas acessíveis voltadas ao apoio à verificação de informações.

DECLARAÇÃO DE USO DE INTELIGÊNCIA ARTIFICIAL

Os autores declaram que ferramentas de Inteligência Artificial generativa foram utilizadas exclusivamente como apoio auxiliar à elaboração deste trabalho, incluindo assistência na revisão linguística, estruturação textual e organização de ideias. Nenhuma ferramenta de IA foi utilizada para substituição da autoria intelectual, análise científica, interpretação dos resultados ou tomada de decisões acadêmicas. A responsabilidade integral pelo conteúdo, originalidade, veracidade das informações e conformidade com os princípios de integridade científica permanece integralmente atribuída aos autores.

REFERÊNCIAS

ABAD, C. S. **La primera fake news de la historia.** Historia y comunicación social, v. 24, n. 2. 2019.

AL-KHAZRAJI, S. H. et al. **Impact of deepfake technology on social media: Detection, misinformation and societal implications**. The Eurasia Proceedings of Science Technology Engineering and Mathematics, ISRES Publishing, v. 23, n. 429-441, p. 2. 2023.

ALI, I. et al. **Fake news detection techniques on social media: A survey**. Wireless Communications and Mobile Computing, Wiley Online Library, v. 2022, n. 1, p. 6072084. 2022.

ALLCOTT, H.; GENTZKOW, M. **Social media and fake news in the 2016 election**. Journal of economic perspectives, American Economic

Association 2014 Broadway, Suite 305, Nashville, TN 37203-2418, v. 31, n. 2, p. 211–236. 2017.

ANDERSON, C. et al. **Propaganda, misinformation, and histories of media techniques**. Harvard Kennedy School Misinformation Review, v. 2021, n. 2, p. 1–7. 2021.

AOSFATOS. **Sobre o Aos Fatos**. 2025. Disponível em: <https://www.aosfatos.org/sobre-o-aos-fatos/>. Acesso em: 08 nov. 2025.

BARBOZA, E. D.; SERVIDONI, M. C. **O impacto das fake news na sociedade**. Revista Interface Tecnológica, v. 18, n. 1, p. 169–180. 2021.

BASARSLAN, M. S.; KAYAALP, F. et al. **Sentiment analysis with machine learning methods on social media**. Ediciones Universidad de Salamanca (España). 2020.

BOVET, A.; MAKSE, H. A. **Influence of fake news in twitter during the 2016 us presidential election**. Nature communications, Nature Publishing Group UK London, v. 10, n. 1, p. 7. 2019.

CHANDRA, K. S.; ANITA, D. G. **Social media: Positive and negative effects on journalism**. Journal of Education: Rabindra Bharati University, XXV, n. 3(IV), p. 82–92. ISSN 0972-7175. 2022.

CHAVARRO, J. P. et al. **Faketruebr: Um corpus brasileiro de notícias falsas**. In: SBC. **Escola Regional de Banco de Dados (ERBD)**. [S.l.], 2023. 2023. p. 108–117.

CINUS, F. et al. **The effect of people recommenders on echo chambers and polarization**. In: **Proceedings of the International AAAI Conference on Web and Social Media**. [S.l.: s.n.], 2022. 2022. v. 16, p. 90–101.

DEBROY, O.; HEMMIGE, B. D. **Psycho-social impact of deepfake content in entertainment media**. International journal for innovative research in multidisciplinary field, v. 10, n. 5, p. 235–249. 2024.

EXETER, U. of. **What is fake news?** 2025. Disponível em: <https://libguides.exeter.ac.uk/fakenews>. Acesso em: 14 out. 2025.

FÁTIMA, B. D.; MIRANDA, S. **Discurso de ódio, fake news e redes sociais: uma breve introdução**. Razón y Palabra, v. 26, n. 113, p. 12–16. 2022.

GAILLARD, M.; EGYED-ZSIGMOND, E. **Large scale reverse image search**. XXXVème Congrès INFORSID, v. 127. 2017.

GRAVES, L. **Anatomy of a fact check: Objective practice and the contested epistemology of fact checking**. Communication, culture & critique, Oxford University Press, v. 10, n. 3, p. 518–537. 2017.

GRAVES, L.; AMAZEEN, M. **Fact-checking as idea and practice in journalism**. Oxford University Press. 2019.

GUETTERMAN, T. C. et al. **Augmenting qualitative text analysis with natural language processing: methodological study**. Journal of medical Internet research, JMIR Publications Toronto, Canada, v. 20, n. 6, p. e231. 2018.

GUPTA, S.; KAUSHIK, A.; KAUR, P. **Deepfake overview: Generation, detection, risks and opportunities**. Detection, Risks and Opportunities (March 21, 2025). 2025.

HANGLOO, S.; ARORA, B. **Fake news detection tools and methods–a review**. arXiv preprint arXiv:2112.11185. 2021.

IFCN, I. F. N. **About – International Fact-Checking Network (IFCN)**. 2025. Disponível em: <https://www.poynter.org/ifcn/>. Acesso em: 08 nov. 2025.

JERIT, J.; ZHAO, Y. **Political misinformation**. Annual Review of Political Science, Annual Reviews, v. 23, n. 1, p. 77–94. 2020.

LUPA, A. **Sobre a Lupa**. 2025. Disponível em: <https://www.agencialupa.org/sobre-a-lupa/>. Acesso em: 08 nov. 2025.

MAHLOUS, A. R. **The impact of fake news on social media users during the covid-19 pandemic, health, political and religious conflicts: A deep look**. Int. J. Psychol. Relig, v. 5, p. 481–492. 2024.

MONTEIRO, R. A. et al. **Contributions to the study of fake news in portuguese: New corpus and automatic detection results**. In: SPRINGER. **International Conference on Computational Processing of the Portuguese Language**. [S.l.], 2018. 2018. p. 324–334.

NAEEM, S. B.; BHATTI, R.; KHAN, A. **An exploration of how fake news is taking over social media and putting public health at risk**. Health Information & Libraries Journal, Wiley Online Library, v. 38, n. 2, p. 143–149. 2021.

NARAYANAN, A. **Understanding social media recommendation algorithms**. 2023.

NASCIMENTO, I. J. B. D. et al. **Infodemics and health misinformation: a systematic review of reviews**. Bulletin of the World Health Organization, v. 100, n. 9, p. 544. 2022.

NASTESKI, V. **An overview of the supervised machine learning methods**. Horizons. b, v. 4, n. 51-62, p. 56. 2017.

NGUYEN, T. T. et al. **Deep learning for deepfakes creation and detection: A survey**. Computer Vision and Image Understanding, Elsevier, v. 223, p. 103525. 2022.

QUELLE, D. et al. **Lost in translation: using global fact-checks to measure multilingual misinformation prevalence, spread, and evolution**. EPJ Data Science, Springer Berlin Heidelberg, v. 14, n. 1, p. 22. 2025.

RANI, N.; DAS, P.; BHARDWAJ, A. K. **Rumor, misinformation among web: a contemporary review of rumor detection techniques during different web waves**. Concurrency and Computation: Practice and Experience, Wiley Online Library, v. 34, n. 1, p. e6479. 2022.

SAEIDNIA, H. R. et al. **Artificial intelligence in the battle against disinformation and misinformation: a systematic review of challenges and approaches**. Knowledge and Information Systems, Springer, v. 67, n. 4, p. 3139–3158. 2025.

SALLAMI, D.; AÏMEUR, E. **Aletheia: Detect, discuss, and stay informed on fake news**. In: **Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence**. [S.l.: s.n.], 2025. 2025. p. 11109–11113.

SARKER, I. H. **Machine learning: Algorithms, real-world applications and research directions**. SN computer science, Springer, v. 2, n. 3, p. 160. 2021.

TARIGAN, I. M. et al. **Understanding social media: Benefits of social media for individuals**. Jurnal Pendidikan Tambusai, v. 7, n. 1, p. 2317–2322. 2023.

WANG, X. **The impact of the application of algorithms in social media marketing on society**. Lecture Notes in Education Psychology and Public Media, v. 9, n. 1, p. 424–431. 2023.

WHO. **Infodemia**. 2025. Disponível em: <https://www.who.int/health-topics/infodemic#tab=tab_1>. Acesso em: 17 out. 2025.

WORLD HEALTH ORGANIZATION. **14.9 million excess deaths associated with the COVID-19 pandemic in 2020 and 2021**. 2022. Disponível em: <https://www.who.int/news/item/05-05-2022-14.9-million-excess-deaths-were-associated-with-the-covid-19-pandemic-in-2020-and-2021>. Acesso em: 1 set. 2025.

ZANARDELLI, M. et al. **Image forgery detection: a survey of recent deep-learning approaches**. Multimedia Tools and Applications, Springer, v. 82, n. 12, p. 17521–17566. 2023.

ZHANG, B. et al. **Review of nlp applications in the field of text sentiment analysis**. Journal of Industrial Engineering and Applied Science, v. 2, n. 3, p. 28–34. 2024.