APRENDIZADO DE MÁQUINA APLICADO À DETECÇÃO DE FAKE NEWS TEXTUAIS: UMA REVISÃO INTEGRATIVA DA LITERATURA

MACHINE LEARNING APPLIED TO THE DETECTION OF TEXTUAL FAKE NEWS: AN INTEGRATIVE LITERATURE REVIEW

Karina Nascimento Favareto
Victor Hugo Orlandi
Eliane Vendramini de Oliveira

Resumo

A disseminação de desinformação em ambientes digitais tornou-se um dos principais desafios da sociedade contemporânea. Nesse contexto, técnicas de Inteligência Artificial, Aprendizado de Máquina e Processamento de Linguagem Natural têm sido investigadas para a detecção automática de *fake news*. Este artigo apresenta uma revisão integrativa da literatura sobre modelos de classificação textual, com destaque para arquiteturas *Transformer* e BERT. Os resultados indicam que abordagens baseadas em *deep learning* apresentam desempenho superior aos métodos tradicionais. Conclui-se que essas técnicas representam uma alternativa promissora no combate à desinformação digital, embora ainda existam desafios relacionados aos dados, interpretação dos modelos e questões éticas.

Palavras-chave: *fake news*, aprendizado de máquina, processamento de linguagem natural, inteligência artificial, classificação de texto.

Abstract

The spread of misinformation in digital environments has become one of the main challenges of contemporary society. In this context, Artificial Intelligence, Machine Learning, and Natural Language Processing techniques have been widely investigated for the automatic detection of fake news. This article presents an integrative literature review on text classification models, with emphasis on Transformer-based architectures and BERT models. The results indicate that deep learning approaches achieve superior performance when compared to traditional methods. It is concluded that these techniques represent a promising alternative for combating digital misinformation, although challenges related to data quality, model interpretability, and ethical issues still remain.

Keywords: *fake news, machine learning, natural language processing, artificial intelligence, text classification.*

Aluno do curso de Tecnologia em Análise e Desenvolvimento de Sistemas, da Faculdade de Presidente Prudente. Email: victor.orlandi@aluno.cps.sp.gov.br.
Aluno do curso de Tecnologia em Análise e Desenvolvimento de Sistemas, da Faculdade de Presidente Prudente. Email: karina.favareto@aluno.cps.sp.gov.br.
Professora orientadora Dra. em Engenharia Elétrica, da Faculdade de Presidente Prudente. E-mail: eliane.oliveira02@cps.sp.gov.br.

1. INTRODUÇÃO

A circulação de informações sempre esteve presente na dinâmica social humana. A disseminação de acontecimentos cotidianos, seja de forma positiva ou negativa, constitui um processo inerente à comunicação entre indivíduos. No contexto contemporâneo, entretanto, esse fenômeno ganhou novas proporções devido à expansão da internet e das redes sociais digitais, que ampliaram significativamente a velocidade e o alcance da circulação de informações.

Cardoso (2021) cita que de acordo com uma pesquisa realizada pela DNPontocom, a geração z, com pessoas nascidas entre 1990 e 2000, não lê além dos títulos das informações. Esse comportamento amplia o impacto das informações falsas, podendo gerar consequências sociais relevantes, como polarização política, disseminação de teorias conspiratórias e perda de confiança em instituições científicas e jornalísticas.

Nesse cenário, destaca-se o fenômeno das chamadas *fake news*, o qual ganhou ampla visibilidade no Brasil especialmente durante o período eleitoral de 2018 e também no contexto da pandemia de COVID-19, momentos em que a circulação de informações falsas se intensificou nas plataformas digitais. De acordo com Cardoso (2021), a propagação dessas informações ocorre frequentemente com a intenção de distorcer fatos ou manipular a interpretação da realidade. Sena (2020) observa que a desinformação digital se tornou um problema global, envolvendo inclusive a monetização de conteúdos falsos na internet, além de discorrer sobre o financiamento de sites de notícias falsas por parte da Google e outras grandes empresas.

Determinados grupos populacionais podem apresentar maior vulnerabilidade à desinformação, como, por exemplo, Cardoso (2021) ressalta que jovens e usuários com menor nível de alfabetização digital podem apresentar maior probabilidade de acreditar ou compartilhar conteúdos falsos na internet.

As redes sociais desempenham papel central nesse processo. Vergna, Silva e Amorim (2022) destacam que essas plataformas ampliam significativamente o potencial de disseminação de conteúdos, incluindo informações falsas, devido à rapidez de compartilhamento e ao grande número de usuários conectados. Nesse contexto, cresce a necessidade de desenvolver mecanismos tecnológicos capazes de auxiliar na identificação e no combate à desinformação digital.

Entre as soluções tecnológicas investigadas atualmente, destacam-se técnicas baseadas em inteligência artificial e aprendizado de máquina. Esses métodos permitem analisar grandes volumes de dados textuais e identificar padrões associados à disseminação de conteúdos falsos. Entretanto, estudos recentes apontam que a utilização dessas tecnologias também apresenta desafios. Maraccini (2024) relata que determinados sistemas automatizados de verificação podem, paradoxalmente, reforçar a crença em manchetes falsas dependendo da forma como as informações são apresentadas aos usuários.

Diante desse contexto, pesquisadores da área de ciência da computação passaram a investigar o uso de técnicas de inteligência artificial como ferramentas capazes de auxiliar na detecção automatizada de desinformação. Carvalho, Santos e Conte (2025) afirmam que técnicas de aprendizado de máquina vêm se consolidando como ferramentas relevantes para apoiar processos automatizados de detecção de notícias falsas.

Assim, o presente artigo tem como objetivo geral analisar o uso de técnicas de aprendizado de máquina aplicadas à detecção de *fake news*. Como objetivos específicos, busca-se compreender os fundamentos teóricos das técnicas de classificação textual utilizadas nesse contexto e discutir a aplicação de modelos de inteligência artificial no reconhecimento automatizado de notícias falsas.

2. METODOLOGIA

O presente estudo caracteriza-se como uma revisão integrativa da literatura, com o objetivo de analisar a evolução das técnicas de classificação de texto no contexto do Processamento de Linguagem Natural, buscando identificar as abordagens que apresentam os melhores resultados na detecção de *fake news* por meio do aprendizado de máquina. Segundo Mendes, Silveira e Galvão (2008, p. 760), “A revisão integrativa da literatura consiste na construção de uma análise ampla da literatura, contribuindo para discussões sobre métodos e resultados de pesquisas, assim como reflexões sobre a realização de futuros estudos.”

A questão de pesquisa que orienta este trabalho é: qual a técnica de classificação de texto que apresenta melhores resultados no aprendizado de máquina para detecção de *fake news*?

Embora se trate de uma revisão integrativa, foram adotados procedimentos inspirados nas diretrizes do PRISMA (Page *et al.*, 2021), um guia criado com o objetivo de garantir maior organização e transparência na seleção dos estudos.

2.1 Estratégias de Busca

A busca foi realizada nas bases Google Scholar, SciELO e Semantic Scholar, considerando publicações no período de 2020 a 2025.

Foram utilizados descritores em português e inglês, com o uso de operadores booleanos (*AND*, *OR*) e estratégias de refinamento. Em algumas consultas, foram aplicados filtros para exclusão de dissertações, teses e trabalhos de conclusão de curso por meio dos operadores *-dissertação -tese -TCC*, enquanto em outras buscas esses documentos foram mantidos quando considerados relevantes, sendo essa uma limitação encontrada durante as buscas.

Também foram aplicados filtros como ordenação por relevância, restrição de idioma e, em alguns casos, seleção de estudos com acesso ao texto completo.

Dentre os principais termos utilizados, destacam-se: *comparative analysis of text classification techniques in fake news detection*; análise de técnicas de classificação de texto usadas para detecção de fake news; BERT *AND fake news detection* -dissertação -tese -dissertações -teses -tcc; *text classification for fake news detection*; evolução da classificação de texto em *machine learning*; "*fake news detection*" *using machine learning*; evolução da classificação de texto e *transformer*; classificação de texto *AND* processamento de linguagem natural classificação *OR transformer OR* processamento *OR* linguagem *OR* natural; *transformers* NLP bert *for text classification transformers OR* bert *OR naive OR bayes OR* tf-idf -dissertação -dissertações -tcc -tese -teses; melhor técnica de classificação de texto para detecção de *fake news*; detecção *fake news* portugues técnica de classificação; detecção *fake news* portugues técnica de classificação -dissertação -tcc -tese -"trabalho de conclusão de curso".

Os critérios de inclusão foram estudos publicados entre 2020 e 2025, que abordassem técnicas de classificação de texto aplicadas à detecção de *fake news*, incluindo artigos

científicos e, quando relevantes, dissertações e trabalhos acadêmicos, com disponibilidade de acesso ao texto completo.

Os critérios de exclusão foram estudos fora do período definido, trabalhos que não tratassem diretamente do tema, artigos duplicados e estudos sem acesso ao texto completo, além de documentos excluídos em buscas específicas por meio dos filtros aplicados.

2.2 Seleção dos estudos

A Figura 1, a seguir, ilustra as etapas de identificação do material usado na análise para a revisão integrativa.

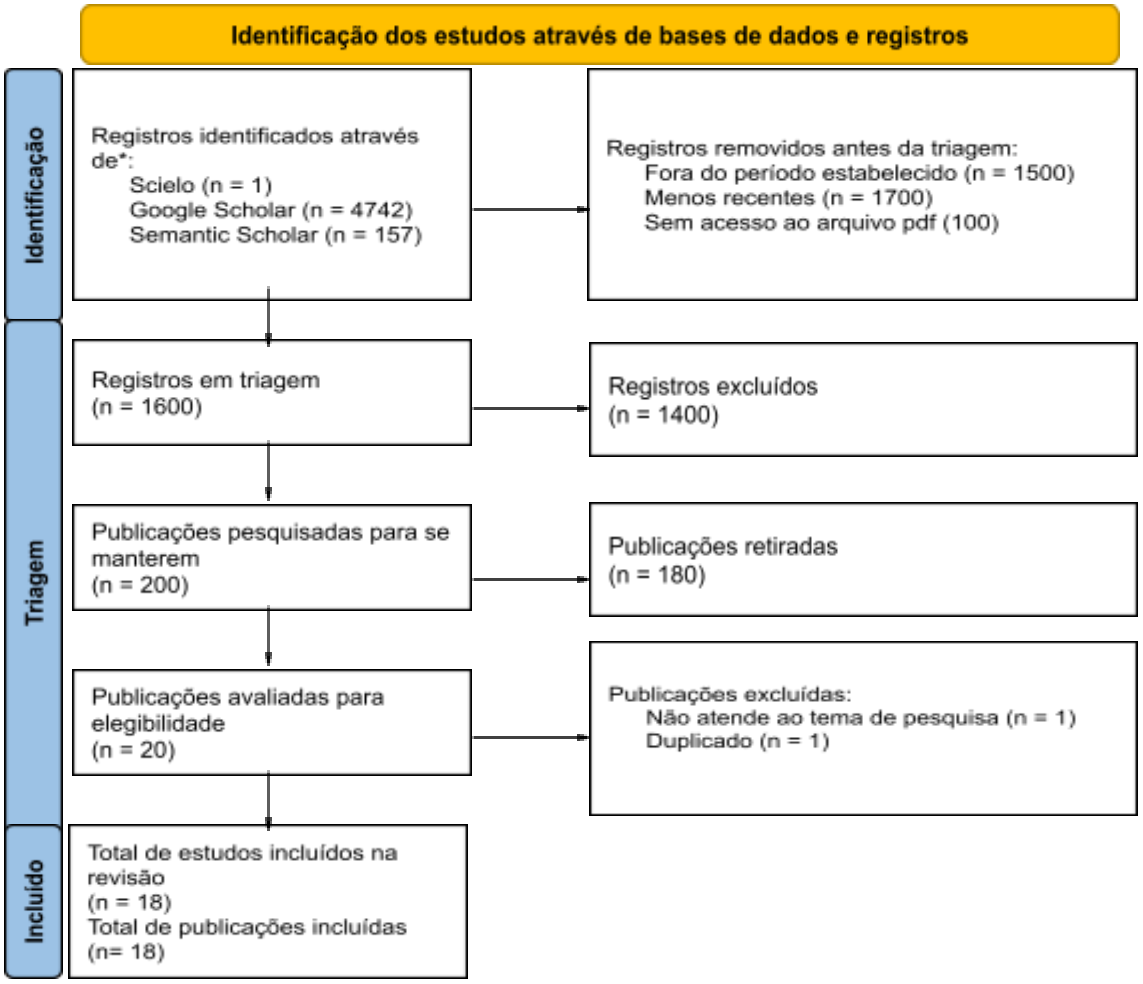

Figura 1 — Fluxograma do processo de seleção dos artigos

Inicialmente, foram identificados 4.900 registros nas bases de dados Scielo, Google Scholar e Semantic Scholar. Após a aplicação dos filtros de busca, incluindo período de

publicação, disponibilidade do texto completo em formato PDF e aderência aos descritores utilizados, 3.300 registros foram excluídos.

Na etapa de triagem, realizada por meio da leitura dos títulos e resumos, 1.400 publicações foram excluídas por não apresentarem relação direta com o tema da pesquisa. Dessa forma, 200 estudos permaneceram para análise.

Em seguida, os trabalhos selecionados foram submetidos à avaliação de elegibilidade, considerando sua relevância para o objetivo da pesquisa e a possibilidade de acesso ao texto completo. Nessa etapa, 180 publicações foram excluídas por não atenderem aos critérios estabelecidos.

Os 20 estudos restantes foram lidos integralmente. Após a leitura completa, dois trabalhos foram excluídos: um por não atender adequadamente ao tema proposto e outro por corresponder a um artigo derivado de um Trabalho de Conclusão de Curso já incluído na amostra. Ao final do processo, 18 estudos compuseram a revisão integrativa.

O processo de seleção ocorreu em etapas: leitura de títulos, análise de resumos e leitura completa dos estudos selecionados.

Como exemplo, a busca *"bert AND fake news detection -dissertação -tese -dissertações -teses -tcc"* no Google Scholar resultou em 39 trabalhos, dos quais apenas 1 foi considerado relevante após as etapas de triagem. Já no Semantic Scholar, a busca por *"melhor técnica de classificação de texto para detecção de fake news"*, com delimitação temporal entre 2020 e 2025, retornou até 88 resultados, posteriormente refinados conforme critérios de relevância.

2.3 Extração e Análise dos Dados

Os estudos selecionados foram organizados em uma planilha contendo informações como autores, ano e local de publicação, e principais resultados. A análise foi conduzida de forma qualitativa, com foco na comparação entre abordagens tradicionais, como *Bag of Words* e *Term Frequency–Inverse Document Frequency* (TF-IDF), e modelos mais recentes baseados em arquiteturas do tipo *Transformer*, como o *BERT*.

3. FUNDAMENTAÇÃO TEÓRICA

3.1 *Fake News* Textuais: Conceituação e Impacto

A disseminação de informações falsas não é um fenômeno recente, mas o surgimento das redes sociais digitais ampliou significativamente sua capacidade de propagação. O consumo de notícias em veículos tradicionais perdeu espaço para conteúdos que circulam em plataformas digitais, cuja facilidade de compartilhamento e atualização instantânea favorecem a viralização de informações - verdadeiras ou não (Shu *et al.*, 2017).

Wardle e Derakhshan (2017) classificam o ecossistema da desinformação em três categorias: *misinformation*, informações falsas disseminadas sem intenção de dano; *disinformation*, conteúdo falso distribuído deliberadamente para causar prejuízo - categoria na qual se inserem as fake news em sentido estrito; e *malinformation*, informações verdadeiras usadas com intenção de prejudicar uma pessoa ou instituição.

O impacto desse fenômeno tornou-se especialmente evidente em contextos eleitorais e de saúde pública, em que indivíduos ou grupos "criam e distribuem intencionalmente informação falsa para manipular a percepção pública, desmerecer oponentes ou promover agendas específicas" (Ilham; Mujiyono, 2025, p. 81). Diante desse cenário, cresce a necessidade de mecanismos tecnológicos capazes de auxiliar na identificação automatizada de desinformação.

3.2 Processamento de Linguagem Natural (PLN)

O Processamento de Linguagem Natural (PLN), é uma área da Inteligência Artificial (IA) que estuda o desenvolvimento de métodos computacionais para a compreensão e geração de linguagem humana, incluindo tarefas como tradução, busca de informações e classificação textual (Santos, 2020). Sistemas computacionais não compreendem linguagem humana, portanto, esse entendimento acontece por meio da transformação de palavras e frases em representações numéricas (processo nomeado como representação vetorial), viabilizando que algoritmos extraiam valor de textos em linguagem natural (Pereira, 2025).

Um dos maiores desafios do PLN é o tratamento de nuances como ambiguidades, ironias e regionalismos (Fernandes, 2025). Para que um modelo opere adequadamente sobre

textos, são necessárias etapas de pré-processamento. Pereira (2025) organiza esse fluxo nas seguintes etapas: normalização, que padroniza os *tokens* ou palavras ao converter para letras minúsculas ou maiúsculas, remover caracteres especiais e acentuação para evitar que variações desnecessárias sejam tratadas como termos distintos; a remoção de *stopwords*, que elimina palavras comuns e com pouco valor semântico que geram ruídos no processamento, como artigos e preposições; a tokenização, que divide o texto em unidades menores chamadas *tokens*; e a lematização e *stemming*, técnicas que reduzem palavras à sua forma base.

A lematização considera gramática e contexto, enquanto o *stemming* aplica regras simplificadas de remoção de sufixos para extrair o radical das palavras, sendo mais simples, porém menos preciso (Pereira, 2025). Por exemplo, na lematização, "observei" e "observaram" seria reduzido a "observar", enquanto no *stemming*, a forma base ficaria "observ".

3.3 Representação de Texto: Abordagens Clássicas

Após a limpeza do texto na etapa de pré-processamento, é preciso realizar a representação vetorial. Um dos modelos mais simples e antigos é o *Bag of Words*, que cria um vetor contando a frequência com que cada palavra aparece no texto. A grande desvantagem desse modelo é que ele ignora a ordem das palavras e o contexto em que foram usadas, tratando a frase apenas como um conjunto de termos soltos (Pereira; Manica, 2023).

Para suprir essa limitação, surgiu o método TF-IDF *(Term Frequency - Inverse Document Frequency)*. De acordo com Pereira (2025), em vez de apenas contar palavras, ele atribui pesos a elas para demonstrar sua real relevância. Palavras que aparecem muitas vezes em um único documento ganham um peso maior (TF - *Term Frequency)*, enquanto termos que aparecem em quase todos os documentos da base de dados ganham um peso menor (IDF - *Inverse Document Frequency)*, diluindo a importância de palavras comuns e destacando os termos mais específicos. Ainda assim, esse é um método que não captura o contexto e as relações entre palavras, o que é uma de suas maiores limitações.

3.4 Algoritmos de Classificação

3.4.1 Algoritmos Clássicos

Santos (2020) define aprendizado de máquina supervisionado como a área que estuda algoritmos capazes de aprender a resolver problemas de maneira autônoma, reconhecendo e extraindo padrões de grandes volumes de dados rotulados. A classificação de texto possui duas etapas: fase de treinamento, em que o modelo recebe um conjunto de dados rotulados e aprende com ele, ajustando parâmetros de acordo com características extraídas desse conjunto; e a fase de inferência, em que o modelo recebe um documento não utilizado no conjunto de treinamento e faz a predição sobre a categoria e rótulo desse documento (Caseli; Nunes, 2026).

Dentre os métodos clássicos, o *Naive Bayes* se destaca por sua velocidade e simplicidade, utilizando o Teorema de *Bayes* para calcular a probabilidade de uma notícia pertencer a uma determinada classe (Fernandes, 2025). O algoritmo analisa a ocorrência de palavras e frases em cada categoria de documentos rotulados utilizados na fase de treinamento e, ao classificar um novo documento, "avalia a probabilidade de cada palavra do artigo pertencer à classe "verdadeiro" ou "falso" e combina essas probabilidades para determinar a categoria final do artigo" (Fernandes, 2025, p. 39). Sua variante, conforme explica Santos (2020), o *Multinomial Naive Bayes*, é amplamente aplicada por lidar de maneira eficiente com dados contáveis, como frequência de palavras.

Outras abordagens tradicionais adotam princípios de geometria, otimização, árvores e *ensembles*, segundo Caseli e Nunes (2026). O algoritmo *K-Nearest Neighbors* (KNN) é um método baseado na distância (similaridade) de um documento em relação a outros documentos do conjunto de treinamento, classificando-o de acordo com a classe que aparece com mais frequência entre seus "vizinhos". É um algoritmo preguiçoso (*lazy*), ideal para pequenos conjuntos de documentos, uma vez que a predição de um novo texto é feita ao compará-lo a todos os exemplos de treinamento armazenados em sua memória (Caseli; Nunes, 2026).

A Regressão Logística e o *Support Vector Machine* (SVM) estão entre os métodos de otimização aplicados à classificação textual. Segundo Caseli e Nunes (2026), esses métodos

buscam construir uma função de decisão capaz de separar classes no espaço de características. Na Regressão Logística, essa separação ocorre pela estimativa de um evento (no contexto desse trabalho, um texto) pertencer a determinada categoria com base em um conjunto de variáveis independentes, o que justifica sua aplicação em problemas que necessitem resultados binários, como 0 ou 1, sim ou não e a distinção entre notícias falsas e verdadeiras. (Sharma; Saran; Patil, 2021). O SVM, por sua vez, define um hiperplano de separação entre as classes e busca maximizar a margem entre elas, estabelecendo os vetores de suporte na fronteira de cada margem e comparando-os a novos textos para realizar a classificação. Tal característica favorece a utilização em tarefas de classificação textual com vetores de alta dimensionalidade (Pellegrini, 2025; Caseli; Nunes, 2026).

O *Random Forest* (RF) pertence ao grupo dos métodos baseados em árvores e *ensembles*. Nesse método, conforme explicam Caseli e Nunes (2026), várias árvores de decisão são combinadas para produzir uma classificação final robusta, reduzindo a dependência de uma única árvore, o risco de sobreajuste (*overfitting*) e aumentando a estabilidade do modelo. De acordo com Sharma, Saran e Patil (2021), o *Passive Aggressive Classifier* (PAC) mantém seus parâmetros quando a classificação está correta (*passive*) e os ajusta quando as predições estão incorretas (*aggressive*). O *Stochastic Gradient Descent* (SGD) não é um classificador, mas sim um algoritmo matemático que minimiza uma função de erro que atua como parte de uma estratégia de otimização para ajuste de parâmetros em modelos lineares (Pellegrini, 2025). Por fim, o *Multilayer Perceptron* (MLP) é uma rede neural de camadas conectadas capaz de identificar padrões complexos durante seu treinamento e aprender relações entre os vetores de entrada e as classes de saída, embora apresente uma estrutura mais simples do que arquiteturas profundas como Redes Neurais Convolucionais (CNNs), Redes Neurais Recorrentes (RNNs) e *Transformer* (Caseli; Nunes, 2026).

3.4.2 Aprendizado Profundo e Arquiteturas *Transformer*

Situadas no grupo de métodos conexionistas por Caseli e Nunes (2026), as redes neurais artificiais são modelos computacionais inspirados no funcionamento do cérebro humano. Esses modelos são compostos por camadas de neurônios artificiais capazes de aprender representações complexas a partir de grandes volumes de dados. O treinamento

dessas redes ocorre por meio de algoritmos de otimização baseados em *backpropagation*, que ajusta os pesos das conexões entre neurônios. As redes neurais artificiais se diferenciam dos métodos clássicos, que dependem com maior frequência de representações previamente definidas, como *Bag of Words* e TF-IDF.

As Redes Neurais Convolucionais, ou CNNs, foram inicialmente desenvolvidas para o processamento de imagens, mas também foram adaptadas para tarefas de classificação textual (Fernandes, 2025). Nesse contexto, utilizam filtros que percorrem sequências de palavras ou *tokens*, permitindo a identificação de padrões locais no texto. Por esse motivo, Caseli e Nunes (2026) descrevem as CNNs como "detectores de n-gramas" automáticos, pois elas reconhecem combinações de termos relevantes para diferenciar categorias. Na detecção de *fake news*, essa característica permite que o modelo contribua para a classificação de notícias verdadeiras e falsas a partir de padrões extraídos do conteúdo textual (Ilham; Mujiyono, 2025).

As Redes Neurais Recorrentes, ou RNNs, são arquiteturas neurais voltadas ao processamento de dados sequenciais, o que as torna relevantes para tarefas de PLN, já que a ordem das palavras interfere na interpretação do texto. Conforme explicam Caseli e Nunes (2026), esse tipo de modelo considera a sequência de entrada ao longo do processamento, permitindo que informações anteriores influenciem a análise dos próximos elementos. No entanto, RNNs tradicionais apresentam limitações para preservar relações entre termos distantes na frase ou no documento. Para lidar com esse problema, a LSTM (*Long Short-Term Memory*) utiliza mecanismos de memória que controlam quais informações devem ser mantidas, atualizadas ou descartadas ao longo da sequência, favorecendo o tratamento de contextos mais extensos em textos.

A arquitetura *Transformer*, proposta por Vaswani *et al.* (2017), representou um avanço significativo no campo do processamento de linguagem natural ao substituir a dependência de recorrência por mecanismos de autoatenção (*self-attention*). Segundo Pellegrini (2025), trata-se de um modelo de transdução de sequências, isto é, um modelo cuja entrada e saída são sequências, potencialmente de tamanhos distintos, baseado integralmente em mecanismos de autoatenção. Diferentemente das redes recorrentes, que processam os elementos de forma sequencial, os *Transformers* favorecem o paralelismo, entendido como a possibilidade de

processar diferentes partes da sequência de maneira simultânea durante o treinamento, uma das motivações apontadas para sua criação (Vaswani *et al.*, 2017; Pellegrini, 2025).

Em tarefas de classificação textual, os *Transformers* se destacam porque produzem representações contextuais, ou seja, representações que consideram o significado das palavras de acordo com o contexto em que aparecem. Pereira (2025) explica que modelos como *Word2Vec* geram uma representação fixa para cada palavra, o que dificulta o tratamento de casos como a polissemia. Já os mecanismos de autoatenção permitem gerar representações dinâmicas e contextualizadas, nas quais cada palavra é representada a partir das relações estabelecidas com os demais termos da sequência.

3.5 Modelos Pré-Treinados: BERT e Variantes

Proposto por Devlin *et al.* (2019), o modelo BERT (*Bidirectional Encoder Representations from Transformers*) é uma arquitetura baseada em *Transformers* que permite o processamento bidirecional do texto. Isso significa que o modelo considera simultaneamente o contexto à esquerda e à direita de cada palavra, permitindo melhor compreensão semântica. O BERT foi pré-treinado em grandes volumes de texto por meio de duas tarefas principais: o *masked language modeling* (MLM), em que palavras são ocultadas aleatoriamente e o modelo deve prevê-las com base no contexto, e o *next sentence prediction* (NSP), em que o modelo aprende a identificar se duas frases se seguem no texto original (Devlin *et al.*, 2019).

Uma das principais características do BERT é a utilização do *token* [CLS], que representa o vetor utilizado para tarefas de classificação textual. Caseli e Nunes (2026) explicam que esse *token* é inserido no início do documento e, após passar pelas camadas de atenção da rede, seu vetor resultante funciona como uma síntese matemática do texto analisado. Na classificação de textos, esse vetor pode ser conectado a uma camada de saída responsável por prever a classe do documento. Dessa forma, o modelo não atua apenas como extrator de características, mas também pode ser ajustado para tarefas de categorização textual (Devlin *et al.*, 2019; Caseli; Nunes, 2026).

Entre os modelos baseados em *Transformers*, o BERT e suas variações têm sido empregados em tarefas de classificação textual por sua capacidade de produzir representações

contextualizadas das palavras. Além do BERT original, há modelos adaptados para diferentes contextos linguísticos, como o BERTimbau, voltado ao português brasileiro, e o mBERT, desenvolvido para uso multilíngue. O BERTimbau é relevante nesse cenário por ter sido treinado com textos em português do Brasil, o que favorece a representação de estruturas, vocabulários e padrões próprios do idioma. O mBERT, por outro lado, permite o processamento de diferentes línguas em um único modelo, sendo útil em contextos multilíngues, embora modelos monolíngues possam apresentar maior adequação quando há disponibilidade de versões específicas para o idioma analisado (Vieira; Souza; Cavalcanti, 2025).

Outras variações associadas à família BERT/*Transformer* incluem RoBERTa e XLM-RoBERTa. De acordo com Vieira, Souza e Cavalcanti (2025), o RoBERTa pode ser compreendido como uma variação do BERT voltada ao aperfeiçoamento do processo de treinamento, enquanto o XLM-RoBERTa amplia essa lógica para um contexto multilíngue. Os autores ainda destacam modelos como BART e mBART, associados a tarefas de reconstrução textual, ampliando o conjunto de abordagens baseadas em *Transformers* aplicadas ao processamento de linguagem natural.

Para compreender as diferenças estruturais entre modelos baseados em *Transformers*, Pellegrini (2025) apresenta três organizações principais: *encoder-only*, *decoder-only* e *encoder-decoder*. Os modelos *encoder-only* são mais associados à compreensão textual e a tarefas de classificação; os *decoder-only* são mais direcionados à geração de texto; e os *encoder-decoder* combinam uma etapa de codificação, responsável pela representação da sequência de entrada, com uma etapa de decodificação, voltada à geração ou reconstrução de uma sequência de saída.

Além dessas arquiteturas, os modelos de linguagem de grande porte (LLMs), também passaram a ser considerados em tarefas de detecção de *fake news*. Emil e Brad (2024) abordam modelos como LLaMA 3.1, ChatGPT 4.0 beta e Gemini, sistemas voltados à interpretação e geração de linguagem natural em larga escala. Diferentemente de modelos ajustados especificamente para classificação textual, esses sistemas podem ser utilizados por meio de instruções em linguagem natural, isto é, *prompts*, para distinguir conteúdos verdadeiros e falsos. No entanto, por serem modelos generalistas, sua aplicação em tarefas

específicas de classificação pode depender da forma como a instrução é formulada e do nível de adaptação do modelo ao domínio analisado.

Também aparecem nesse campo modelos especializados ou derivados, como ALBERT, *FakeBERT*, *DeepFakE* e *EchoFakeD*. O ALBERT pode ser entendido como uma variação compacta da família BERT, voltada à redução de custo computacional e ao aumento de eficiência no treinamento. Já modelos como *FakeBERT*, *DeepFakE* e *EchoFakeD* são associados a estratégias específicas para identificação de notícias falsas, combinando representações profundas de texto com arquiteturas voltadas à classificação (Emil; Brad, 2024). Dessa forma, a detecção automática de *fake news* envolve tanto modelos pré-treinados de uso amplo quanto arquiteturas especializadas para a tarefa, variando conforme o idioma, o tipo de representação textual e o objetivo da classificação.

Assim, as variantes baseadas em BERT, *Transformers* e LLMs ampliam as possibilidades de classificação textual ao permitirem representações mais sensíveis ao contexto do que métodos tradicionais baseados apenas em frequência de termos. Enquanto abordagens como *Bag of Words* e TF-IDF representam o texto principalmente pela ocorrência de palavras, os modelos baseados em *Transformers* consideram as relações entre os termos ao longo da sequência, favorecendo a identificação de padrões semânticos mais complexos. No contexto da detecção de *fake news*, essa característica é relevante porque notícias falsas e verdadeiras podem apresentar diferenças sutis de construção textual, exigindo modelos capazes de captar relações contextuais, ambiguidades e padrões linguísticos menos evidentes.

3.6 Métricas de Avaliação

A avaliação dos modelos de classificação é essencial para verificar o desempenho das técnicas utilizadas na identificação de notícias falsas. Em tarefas de classificação textual, as métricas permitem analisar não apenas a quantidade de acertos do modelo, mas também a qualidade das decisões tomadas em relação às classes avaliadas. Entre as métricas mais utilizadas nesse contexto estão acurácia, precisão, *recall*, F1-*score*, matriz de confusão e curva ROC — *Receiver Operating Characteristic*.

A acurácia mede a proporção de classificações corretas realizadas pelo modelo em relação ao total de previsões (Fernandes, 2025). De acordo com Ilham e Mujiyono (2025), essa métrica indica o quanto o modelo é capaz de classificar corretamente as notícias analisadas. Em outras palavras, quanto maior a acurácia, maior é a quantidade geral de acertos. No entanto, conforme Fernandes (2025), embora uma alta acurácia possa indicar bom desempenho na identificação de notícias verdadeiras e falsas, essa métrica pode não ser suficiente quando há desequilíbrio entre as classes, pois o modelo pode apresentar muitos acertos gerais e, ainda assim, falhar na identificação adequada de uma das categorias.

A precisão avalia a proporção de previsões positivas que realmente pertencem à classe positiva. No contexto da detecção de *fake news*, essa métrica indica, por exemplo, entre as notícias classificadas como falsas, quantas de fato eram falsas. Ilham e Mujiyono (2025) explicam que uma precisão elevada demonstra que o modelo evita classificar incorretamente notícias de uma classe como pertencentes a outra. Assim, essa métrica é especialmente relevante quando o custo dos falsos positivos é alto, isto é, quando uma notícia verdadeira pode ser incorretamente classificada como falsa.

De acordo com Fernandes (2025), o *recall*, também chamado de revocação ou sensibilidade, mede a capacidade do modelo de identificar corretamente os casos positivos existentes na base de dados. Fernandes (2025) ainda explica que, em tarefas de classificação de *fake news*, um *recall* elevado indica que o modelo consegue reconhecer a maior parte das notícias falsas, reduzindo a ocorrência de falsos negativos. Dessa forma, enquanto a precisão observa a qualidade das classificações positivas feitas pelo modelo, o *recall* observa a capacidade de encontrar os casos positivos reais.

Segundo Fernandes (2025), O F1-*score*, ou medida-F1, combina precisão e *recall* em uma única métrica por meio da média harmônica entre elas. Essa medida é útil quando se deseja avaliar o equilíbrio entre a capacidade do modelo de evitar falsos positivos e sua capacidade de reduzir falsos negativos. Fernandes (2025) diz que um F1-*score* elevado indica bom equilíbrio entre precisão e *recall*. Para Ilham e Mujiyono (2025), o F1-*score* é importante especialmente em situações de desequilíbrio entre as classes, pois permite observar o desempenho do modelo de forma mais equilibrada do que a acurácia isolada.

A matriz de confusão é uma ferramenta utilizada para visualizar os acertos e erros do modelo em relação às classes previstas e às classes reais. De acordo com Fernandes (2025), ela apresenta a contagem de verdadeiros positivos, falsos positivos, verdadeiros negativos e falsos negativos, permitindo uma análise mais detalhada dos erros de classificação. Na detecção de *fake news*, essa matriz ajuda a identificar, por exemplo, se o modelo está classificando notícias verdadeiras como falsas ou deixando de reconhecer notícias falsas. Silva, Santos e Pardo (2024) relacionam a matriz de confusão à análise do desempenho do classificador.

A curva ROC (*Receiver Operating Characteristic*) é uma representação gráfica utilizada para avaliar a capacidade de um classificador distinguir entre classes. Fernandes (2025) explica que essa curva relaciona a taxa de verdadeiros positivos com a taxa de falsos positivos em diferentes limiares de decisão. A área sob essa curva, chamada de AUC-ROC, indica o desempenho geral do modelo: quanto maior essa área, melhor tende a ser a capacidade de separação entre as classes. Assim, essa métrica complementa as demais ao permitir uma análise visual e comparativa do comportamento do classificador.

Dessa forma, o uso conjunto dessas métricas permite uma avaliação mais completa dos modelos de detecção de *fake news*. Enquanto a acurácia apresenta uma visão geral dos acertos, a precisão e o *recall* analisam aspectos específicos das classificações positivas. O F1-*score* equilibra essas duas medidas, a matriz de confusão detalha os tipos de erro cometidos e a curva ROC contribui para observar a capacidade de separação entre as classes. Portanto, considerar diferentes métricas é importante para evitar conclusões baseadas apenas no percentual geral de acertos do modelo.

4. RESULTADOS E DISCUSSÃO

Nesta seção serão apresentados os resultados obtidos por meio da leitura dos 18 estudos selecionados e a discussão acerca do tema abordado.

Com o intuito de facilitar a interpretação dos resultados extraídos das fontes selecionadas, foi criado um quadro contendo o número do documento, seu título, objetivo, os

classificadores e algoritmos abordados pelos autores de cada um deles e seus resultados, conforme mostra o Quadro 1:

Quadro 1: Quadro comparativo

Nº	AUTOR / ANO / TÍTULO	OBJETIVO	CLASSIFICADORES/ ALGORITMOS	RESULTADOS
1	(Shushkevich; Alexandrov; Cardiff, 2022) Covid-19 Fake News Detection: A Survey	Analisar abordagens e conjuntos de dados utilizados na detecção de *fake news* sobre COVID-19.	Algoritmos clássicos, como regressão logística e SVM e os baseados em redes neurais e que utilizam arquitetura *Transformers*.	Modelos baseados em *Transformers*, como BERT e C-BERT, apresentaram melhor desempenho na classificação de *fake news*.
2	(Borges, 2024) Análise comparativa de algoritmos de classificação de texto	Comparar o desempenho de algoritmos de classificação de texto em português.	BERT, SVM e *Naive Bayes*.	BERT apresentou melhor desempenho em acurácia e F1-*score*, principalmente em bases maiores.
3	(Pires; Silva, 2024) Portuguese Fake News Classification with BERT models	Treinar e avaliar os modelos BERT, BERTimbau e mBERT.	BERTimbau, BERT e mBERT	BERTimbau apresentou os melhores resultados na classificação das notícias.
4	(Ilham; Mujiyono, 2025) Comparison of Text Classification Techniques in Fake News Detection in The Digital Information Age	Comparar a eficácia de técnicas baseadas em CNN e RNN na detecção de *fake news*.	CNN - *Convolutional Neural Network* e RNN - *Recurrent Neural Network*.	CNN apresentou melhor desempenho geral, enquanto RNN demonstrou boa sensibilidade na detecção.

Quadro 1 — Quadro comparativo (continuação)

Nº	AUTOR / ANO / TÍTULO	OBJETIVO	CLASSIFICADORES / ALGORITMOS	RESULTADOS
5	(Sharma; Saran; Patil, 2021) Fake News Detection using Machine Learning Algorithms	Avaliar algoritmos de *machine learning* na classificação de notícias online.	*Naive Bayes*, *Logistic Regression*, *Random Forest* e *Passive Aggressive Classifier* (PAC).	*Logistic Regression* foi o melhor modelo com acurácia de 65% e com otimização de parâmetro a acurácia foi de 75%.
6	(Pellegrini, 2025) Detecção de Fake News Usando Redes Neurais Baseadas em Transformers	Desenvolver redes neurais baseadas em *Transformers* para classificação de notícias em português.	*K-Nearest Neighbors* (KNN), *Random Forest* (RF), *Stochastic Gradient Descent* (SGD), *Support Vector Machine* (SVM), *Encoder-Only*, *Decoder-Only* e *Encoder-Decoder* (*Transformer*).	Modelos baseados em *Transformers* superaram os demais classificadores em todas as métricas.
7	(Pereira, 2025) Utilização de técnicas de Processamento de Linguagem Natural para construção de um sistema de recomendação baseado na similaridade de vetores de representação textual	Analisar técnicas de PLN e desenvolver sistemas baseados em similaridade textual.	TF-IDF e SBERT.	SBERT apresentou melhor capacidade de capturar relações contextuais do que TF-IDF.
8	(Oliveira *et al.*, 2020) Processamento de Linguagem Natural para Identificação de Notícias Falsas em Redes Sociais: Ferramentas, Tendências e Desafios	Apresentar algoritmos e técnicas utilizadas na detecção de *fake news* em redes sociais.	SVM, TF-IDF, *Random Forest*, KNN.	*One-Class* SVM apresentou desempenho mais homogêneo e melhores resultados de sensibilidade.

Quadro 1 — Quadro comparativo (continuação)

Nº	AUTOR / ANO / TÍTULO	OBJETIVO	CLASSIFICADORES / ALGORITMOS	RESULTADOS
9	(Santana; Oliveira; Nascimento, 2022) Text Classification of News Using Transformer-based Models for Portuguese	Propor o uso do BERTimbau para classificação de notícias em português.	BERTimbau.	BERTimbau apresentou desempenho superior aos métodos tradicionais baseados em BoW.
10	(Garrido-Merchan; Gozalo-Brizuela; Gonzales-Carvajal, 2023) Comparing BERT Against Traditional Machine Learning Models in Text Classification	Adicionar evidência com provas para defender o uso do BERT como padrão em tarefas de NLP.	BERT, *Voting Classifier, Logistic Regression*, *Linear SVC*, *Multinomial Naive Bayes*, *Ridge Classifier* e *Passive Aggressive Classifier*.	BERT apresentou melhor desempenho em comparação aos modelos tradicionais.
11	(Vieira; Souza; Cavalcanti, 2025) Detecção de Fake News em Português: Análise Comparativa entre Métodos de Representação em Português, Inglês e Multilíngues	Investigar métodos de representação textual para detecção de *fake news* em português.	BERTIMBAU, TEENYTINYLLAMA, TUCANO, BERT, ROBERTA, BART, MBERT, XLM-ROBERTA e MBART.	mBART + SVC apresentou melhor desempenho, enquanto BERTimbau + SVC obteve melhor equilíbrio entre precisão e eficiência.
12	(Emil; Brad, 2024) A Comparative Study in Large Language Models Usage for Fake News Detection	Avaliar a capacidade de LLMs na distinção entre notícias falsas e verdadeiras.	RoBERTa, LSTM-RRN, BiLSTM-RRN, ALBERT, FakeBERT, DeepFakE, EchoFakeD, BERT, LLaMA 3.1, LLaMA 3.1 8B *fine- tuned*, ChatGPT 4.0 beta, Gemini	LLMs apresentaram desempenho moderado e inferior aos modelos especializados em PLN.

Quadro 1 — Quadro comparativo (continuação)

Nº	AUTOR / ANO / TÍTULO	OBJETIVO	CLASSIFICADORES / ALGORITMOS	RESULTADOS
13	(Sultana; Nishino, 2023) Fake News Detection System: An implementation of BERT and Boosting Algorithm	Desenvolver um modelo baseado em BERT para identificação de *fake news*.	BERT e *Boosting Algorithm*.	O modelo apresentou alta precisão na classificação de notícias falsas.
14	(Roumeliotis; Tselikas; Nasiopoulos, 2025) Fake News Detection and Classification: A Comparative Study of Convolutional Neural Networks, Large Language Models, and Natural Language Processing Models	Comparar CNN, BERT e GPT na classificação de *fake news*.	CNN, BERT e GPT.	Modelos baseados em GPT apresentaram desempenho superior ao BERT e às CNNs.
15	(Pereira; Manica, 2023) Comparação de Técnicas de Classificação para Identificação de Fake News em Português	Avaliar técnicas de classificação para identificação de *fake news* em português.	SVM, Regressão Logística e BERTimbau.	BERTimbau apresentou melhor desempenho geral, enquanto SVM superou a Regressão Logística entre os modelos clássicos.
16	(Santos, 2020) Detecção de fake news: um comparativo entre os modelos de aprendizado supervisionado passive agressive e multinomial naive bayes	Construir modelos supervisionados para detecção de *fake news*.	*Multinomial Naive Bayes* e *Passive Aggressive*.	*Passive Aggressive* apresentou desempenho superior ao *Multinomial Naive Bayes*.

Quadro 1 — Quadro comparativo (continuação)

Nº	AUTOR / ANO / TÍTULO	OBJETIVO	CLASSIFICADORES / ALGORITMOS	RESULTADOS
17	(Fernandes, 2025) Aprendizado de máquina aplicado à detecção de Fake News em português do Brasil	Analisar algoritmos de *machine learning* na detecção de *fake news* em português.	SVM, CNN e *Naive Bayes*.	SVM apresentou melhor desempenho, seguido por *Naive Bayes* e CNN.
18	(Silva; Santos; Pardo, 2024) Detecção automática de notícias falsas	Discutir métodos e abordagens para identificação automática de *fake news*.	Regressão logística, *perceptron* multicamadas, *Naive-Bayes*, *optimum-path forest*; LSTM, GRU, CNN e BERT (BERTimbau como versão para o português); DistilBERT (MVAE-FakeNews).	Modelos baseados em BERT e DistilBERT apresentaram os melhores desempenhos.

Fonte: Elaborado pelos autores

Os 18 estudos analisados evidenciam a utilização de mais de 30 modelos e algoritmos distintos para a tarefa de detecção de *fake news*. Dentre esses, os modelos baseados na arquitetura *Transformer*, especialmente o BERT e suas variações, foram os mais recorrentes, sendo identificados em pelo menos 12 estudos, seguidos pelo SVM, presente em 6 estudos, e pelo *Naive Bayes* e sua variante *Multinomial NB*, identificados em 5 estudos.

A predominância dos modelos baseados em *Transformers* também se reflete no desempenho apresentado. A maior parte dos estudos que realizaram comparações diretas entre modelos aponta a superioridade dessas abordagens em métricas como acurácia e *F1-score*, especialmente em cenários com maior volume de dados. Tal desempenho está associado à capacidade desses modelos de capturar o contexto semântico das palavras e compreender relações mais complexas presentes no texto.

Segundo Pereira (2025) e Pereira e Manica (2023), abordagens tradicionais baseadas em frequência de termos, como *Bag of Words* e TF-IDF, não consideram adequadamente o contexto e as relações semânticas entre as palavras. Como notícias falsas frequentemente

utilizam estruturas linguísticas semelhantes às de notícias verdadeiras, a compreensão contextual do texto torna-se um fator relevante para a classificação.

Entretanto, nem todos os estudos apresentam consenso absoluto. O estudo de número 14 (Roumeliotis; Tselikas; Nasiopoulos, 2025) indica que modelos baseados em GPT podem superar o BERT em determinadas condições, possivelmente devido à sua maior capacidade de generalização e treinamento em grandes volumes de dados heterogêneos. Por outro lado, o estudo 12 (Emil; Brad, 2024) demonstra que modelos de larga escala, quando utilizados em modo *zero-shot*, apresentam desempenho inferior, com acurácia média de 55,29%, quando comparados a modelos especializados treinados diretamente para a tarefa. Esse resultado sugere que modelos generalistas nem sempre superam modelos especializados, especialmente quando utilizados sem treinamento adicional voltado para a tarefa de classificação de *fake news*.

No que se refere aos modelos clássicos, o SVM destacou-se como a abordagem mais consistente entre os algoritmos tradicionais. O estudo 15 (Pereira; Manica, 2023) demonstrou que o SVM superou a Regressão Logística em termos de desempenho, “atingindo 90% na Medida-F1 em comparação aos 89% da Regressão Logística” (Pereira; Manica, 2023, p. 17). Os autores ainda destacam que a Regressão Logística obteve 90% no *recall*, sendo esse o mesmo resultado do SVM, contudo obteve precisão de 87% (Pereira; Manica, 2023).

Apesar da diferença marginal, o resultado reforça a robustez do SVM em tarefas de classificação textual. Mesmo diante da evolução dos modelos baseados em *deep learning*, o SVM continua sendo uma alternativa competitiva para tarefas de classificação textual. Seu desempenho consistente observado nos estudos analisados pode estar relacionado à sua capacidade de lidar com conjuntos de dados de alta dimensionalidade, característica comum em problemas de processamento de linguagem natural (Pellegrini, 2025; Caseli; Nunes, 2026).

No contexto da língua portuguesa, observa-se um destaque significativo para o modelo BERTimbau. Pereira e Manica (2023) apontam que esse modelo atingiu valores próximos de 99% em métricas de avaliação, consolidando-se como uma das melhores alternativas para o processamento de textos em português. Esse resultado é corroborado pelo estudo 11 (Vieira;

Souza; Cavalcanti, 2025), no qual a combinação BERTimbau e SVC foi identificada como uma solução que apresenta bom equilíbrio entre desempenho e eficiência computacional.

O desempenho do BERTimbau pode ser atribuído ao fato de o modelo ter sido pré-treinado especificamente com textos em português brasileiro (Vieira; Souza; Cavalcanti, 2025). Essa característica permite uma melhor compreensão de aspectos linguísticos da língua, incluindo vocabulário, construções gramaticais e padrões frequentemente encontrados em conteúdos de desinformação.

De modo geral, os resultados evidenciam uma clara evolução das técnicas de classificação de texto. Abordagens tradicionais, como *Bag of Words* e TF-IDF associadas a algoritmos clássicos, embora ainda apresentem desempenho satisfatório em cenários com menor complexidade ou restrições computacionais, vêm sendo progressivamente superadas por modelos baseados em arquiteturas *Transformer*. Esses modelos se destacam especialmente quando há disponibilidade de dados suficientes para o processo de *fine-tuning*, consolidando-se como a abordagem mais eficiente para a detecção de *fake news* na atualidade.

5. CONCLUSÃO

A disseminação de desinformação digital representa um desafio crescente para sociedades contemporâneas altamente conectadas. Nesse contexto, técnicas de aprendizado de máquina e processamento de linguagem natural surgem como ferramentas promissoras para auxiliar na detecção automática de notícias falsas.

A revisão integrativa realizada neste estudo evidenciou a evolução das técnicas utilizadas para classificação textual, desde métodos estatísticos tradicionais até modelos avançados baseados em redes neurais profundas. Os resultados indicam que modelos baseados em *Transformers* e arquiteturas como BERT apresentam desempenho superior em tarefas de análise textual, tornando-se ferramentas relevantes para o enfrentamento da desinformação digital, respondendo à pergunta de pesquisa citada no início do artigo.

Apesar dos avanços tecnológicos, ainda existem desafios significativos relacionados à qualidade dos dados, à interpretabilidade dos modelos e às implicações éticas associadas ao uso de sistemas automatizados de análise de informação.

A inclusão de trabalhos de conclusão de curso e dissertações de mestrado foi necessária na construção deste artigo científico, sendo essa uma limitação apresentada durante as buscas, pois não houve grandes volumes de materiais para embasamento desta pesquisa. Apesar disso, nota-se que esse modelo de estudo explora e detalha minuciosamente os conceitos, sendo fundamental para esclarecimento de termos e conceitos ainda não amadurecidos ao longo dessa jornada.

Sugere-se, para trabalhos futuros, a análise de diferentes mídias como vídeos e imagens, considerando o cenário atual em que a qualidade das mídias produzidas por inteligência artificial tem sido elevada, estando cada vez mais próximas da realidade e representando um potencial vetor de disseminação de desinformação.

REFERÊNCIAS

BORGES, Beatriz Ribeiro. **Análise comparativa de algoritmos de classificação de texto**. 2024. Trabalho de Conclusão de Curso (Bacharelado em Sistemas de Informação) - Universidade Federal de Uberlândia, Uberlândia, Brasil, 2024. Disponível em: https://repositorio.ufu.br/handle/123456789/43701. Acesso em: 13 abr. 2026.

CARDOSO, Davi Valois. O impacto das "*fake news*" na educação dos jovens no Brasil. **Revista Ibero-Americana de Humanidades, Ciências e Educação**, *[S. l.]*, v. 7, n. 6, p. 614–625, 2021. DOI: 10.51891/rease.v7i6.1417. Disponível em: https://periodicorease.pro.br/rease/article/view/1417. Acesso em: 10 maio 2026.

CARVALHO , Arleson João Serrão; SANTOS , Wilker José Caminha dos; CONTE, Thiago Nicolau Magalhães de Souza. Avaliação do potencial da inteligência artificial como ferramenta de combate à desinformação: um estudo comparativo entre ChatGPT e Deepseek na verificação de *fake news*. ARACÊ , [S. l.], v. 7, n. 5, p. 21783–21794, 2025. DOI: 10.56238/arev7n5-047. Disponível em: https://periodicos.newsciencepubl.com/arace/article/view/4836. Acesso em: 10 maio 2026.

CASELI, H.M.; NUNES, M.G.V. (org.) Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português. 4 ed. BPLN, 2026. v. 3. ISBN 978-65-02-00167-7. Disponível em: https://brasileiraspln.ufscar.br/livro-pln-4ed-vol3. Acesso em: 13 abr. 2026.

DEVLIN, Jacob; CHANG, Ming-Wei; LEE, Kenton; TOUTANOVA, Kristina. BERT: pre-training of deep bidirectional transformers for language understanding. **arXiv**, Ithaca, 2019. DOI: 10.48550/arXiv.1810.04805. Disponível em: https://arxiv.org/abs/1810.04805. Acesso em: 11 mar. 2026.

EMIL, Repede Ştefan; BRAD, Remus. A comparative study in large language models usage for fake news detection. **Advances in Artificial Intelligence and Machine Learning**, Romênia, ano 2024, n. 4, p. 2810-2823, 13 nov. 2024. Disponível em: https://www.oajaiml.com/uploads/archivepdf/756844163.pdf. Acesso em: 5 abr. 2026.

FERNANDES, Miguel Ângelo Pinheiro. **Aprendizado de máquina aplicado à detecção de fake news em português do Brasil**. 2025. Trabalho de Conclusão de Curso (Bacharelado em Sistemas de Informação) – Instituto Federal de Educação, Ciência e Tecnologia de Goiás, Brasil, 2025. Disponível em: https://repositorio.ifg.edu.br/handle/prefix/2354. Acesso em: 27 mar. 2026.

GARRIDO-MERCHAN, Eduardo C.; GOZALO-BRIZUELA, Roberto; GONZALEZ-CARVAJAL, Santiago. Comparing BERT against traditional machine learning models in text classification. **Journal of Computational and Cognitive Engineering**, Espanha, ano 2023, v. 2, n. 4, p. 352-356, 21 abr. 2023. DOI: 10.47852/bonviewJCCE3202838. Disponível em: https://ojs.bonviewpress.com/index.php/JCCE/article/view/838. Acesso em: 5 abr. 2026.

ILHAM, Dimas Muhammad; MUJIYONO, Sri. Comparison of Text Classification Techniques in Fake News Detection in the Digital Information Age. **International Journal of Advances in Data and Information Systems**, v. 6, n. 1, p. 80-89, 2025. Disponível em: http://ijadis.net/index.php/ijadis/article/view/1365. DOI: 10.59395/ijadis.v6i1.1365. Acesso em: 23 mar. 2026.

MARACCINI, Gabriela. IA pode aumentar a crença em fake news? Estudo responde. **CNN Brasil**, 2024. Disponível em: https://www.cnnbrasil.com.br/tecnologia/ia-pode-aumentar-a-crenca-em-fake-news-estudo-responde/. Acesso em: 10 maio 2026.

MENDES, Karina Dal Sasso; SILVEIRA, Renata Cristina de Campos Pereira; GALVÃO, Cristina Maria. Revisão integrativa: método de pesquisa para a incorporação de evidências na saúde e na enfermagem. **Texto & Contexto - Enfermagem**, Florianópolis, v. 17, n. 4, p. 758-764, 2008. DOI: 10.1590/S0104-07072008000400018. Disponível em: https://www.scielo.br/j/tce/a/XzFkq6tjWs4wHNqNjKJLkXQ/. Acesso em: 10 maio 2026.

OLIVEIRA, Nicollas R. de; PISA, Pedro Silveira; COSTA, Bernardo; LOPEZ, Martin Andreoni; MORAES, Igor Monteiro; MATTOS, Diogo M. F. 2020. Processamento de Linguagem Natural para Identificação de Notícias Falsas em Redes Sociais: Ferramentas, Tendências e Desafios. In: **Minicursos do XX Simpósio Brasileiro de Segurança da Informação e de Sistemas Computacionais**, p.51-100. DOI: 10.5753/sbc.8285.6.2. Disponível em: https://www.researchgate.net/publication/360038529_Processamento_de_Linguagem_Natural_para_Identificacao_de_Noticias_Falsas_em_Redes_Sociais_Ferramentas_Tendencias_e_Desafios. Acesso em: 25 abr. 2026.

PAGE, Matthew J.; MCKENZIE, Joanne E.; BOSSUYT, Patrick M.; et al. **The PRISMA 2020 statement: an updated guideline for reporting systematic reviews**. BMJ, v. 372, n. 71, 2021. DOI: 10.1136/bmj.n71. Tradução de Verónica Abreu, Sónia Gonçalves-Lopes, José Luís Sousa e Verónica Oliveira. Disponível em: https://www.prisma-statement.org/translations. Acesso em: 26 mar. 2026.

PELLEGRINI, Lucas Guerreiro. **Detecção de fake news usando redes neurais baseadas em transformers**. 2025. Trabalho de Conclusão de Curso (Bacharelado em Ciência da Computação) - Universidade Federal de Uberlândia, Uberlândia, Brasil, 2025. Disponível em: https://repositorio.ufu.br/handle/123456789/47573. Acesso em: 12 abr. 2026.

PEREIRA, Matheus Ferreira; MANICA, Edimar. **Comparação de técnicas de classificação para identificação de fake news em português**. 2023. Trabalho de Conclusão de Curso (Bacharelado em Ciência da Computação) - Instituto Federal de Educação, Ciência e Tecnologia do Rio Grande do Sul, Ibirubá, Brasil, 2023. Disponível em: https://dspace.ifrs.edu.br/handle/123456789/1263. Acesso em:

12 abr. 2026.

PEREIRA, Pedro Henrique Mello. **Utilização de técnicas de processamento de linguagem natural para construção de um sistema de recomendação baseado na similaridade de vetores de representação textual**. 2025. Relatório de Projeto de Investigação (Mestrado em Ciência de Dados para Empresas) – Escola Superior Politécnico Setúbal, Setúbal, Portugal, 2025. Disponível em: https://comum.rcaap.pt/entities/publication/c8a1b33c-9ba3-4f01-84a7-0c5b06b8d121. Acesso em: 13 abr. 2026.

PIRES, Vinícius Baião; SILVA, Daniel Guerreiro e. **Portuguese Fake News Classification with BERT models**. In: ENCONTRO NACIONAL DE INTELIGÊNCIA ARTIFICIAL E COMPUTACIONAL (ENIAC), 21., 2024, Belém/PA. Anais [...]. Porto Alegre: Sociedade Brasileira de Computação, 2024. p. 834-845. ISSN 2763-9061. DOI: 10.5753/eniac.2024.245138. Disponível em: https://sol.sbc.org.br/index.php/eniac/article/view/33848. Acesso em: 30 mar. 2026.

ROUMELIOTIS, Konstantinos I.; TSELIKAS, Nikolaos D.; NASIOPOULOS, Dimitrios K. Fake news detection and classification: a comparative study of convolutional neural networks, large language models, and natural language processing models. **Future Internet**, Atenas, Grécia, ano 2025, v. 17, n. 28, p. 1-29, 9 jan. 2025. DOI: 10.3390/fi17010028. Disponível em: https://www.mdpi.com/1999-5903/17/1/28. Acesso em: 17 abr. 2026.

SANTANA, Isabel N.; OLIVEIRA, Raphael S.; NASCIMENTO, Erick G. S. Text classification of news using transformer-based models for Portuguese. **Journal of Systemics, Cybernetics and Informatics**, Reino Unido, ano 2022, v. 20, n. 5, p. 33-59. DOI: 10.54808/JSCI.20.05.33. Disponível em: https://www.iiisci.org/DOIJSCI/SA702PC22/#/. Acesso em: 17 abr. 2026.

SANTOS, Mirella Gadelha. **Detecção de fake news: um comparativo entre os modelos de aprendizado supervisionado passive agressive e multinomial naive bayes**. 2020. Trabalho de Conclusão de Curso (Bacharelado em Sistemas de Informação) - Centro Universitário Christus, Fortaleza, Brasil, 2020. Disponível em: https://repositorio.unichristus.edu.br/jspui/handle/123456789/1745. Acesso em: 14 abr. 2026.

SENA, Victor. Google e empresas de anúncios financiaram sites de Fake News com US$ 25 bi. **exame,** 2020. Disponível em: https://exame.com/tecnologia/google-e-empresas-de-anuncios-financiaram-sites-de-fake-news-com-us-25-bi/. Acesso em: 10 maio 2026.

SHARMA, Uma; SARAN, Sidarth; PATIL, Shankar M. Fake news detection using machine learning algorithms. **International Journal of Engineering Research & Technology**, Navi Mumbai, Índia, ano 2021, v. 09, n. 3, p. 509-518. Disponível em: https://www.academia.edu/46799010/IJERT_Fake_News_Detection_using_Machine_Learning_Algorithms. Acesso em: 13 abr. 2026.

SHU, Kai; SLIVA, Amy; WANG, Suhang; TANG, Juliang; LIU, Huan. Fake news detection on social media: A data mining perspective. **ACM SIGKDD explorations newsletter**, v. 19, n. 1, p. 22-36, 2017. Disponível em: https://arxiv.org/abs/1708.01967. Acesso em: 11 mar. 2026.

SHUSHKEVICH, Elena; ALEXANDROV, Mikhail; CARDIFF, John. Covid-19 Fake News Detection: A Survey. **Computación y Sistemas**, Cidade do México, México, ano 2022, v. 25, n. 4, p. 783-792, 28 fev. 2022. DOI: 10.13053/cys-25-4-4089. Disponível em: http://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S1405-55462021000400783&lng=es&nr

m=iso. Acesso em: 30 mar. 2026.

SILVA, Renato Moraes; SANTOS, Roney Lira de Sales; PARDO, Thiago Alexandre Salgueiro. Detecção automática de notícias falsas. In: CASELI, H. M.; NUNES, M. G. V. (org.). **Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português**. BPLN, 2024. 3. ed. [S.l.]: BPLN, 2024. cap. 27. ISBN 978-65-01-20581-6. Disponível em: https://brasileiraspln.com/livro-pln/3a-edicao/parte-aplicacoes/cap-fake-news/cap-fake-news.html. Acesso em 13 abr. 2026.

SULTANA, Raquiba; NISHINO, Tetsuro. Fake news detection system: an implementation of BERT and boosting algorithm. **EPiC Series in Computing**: Proceedings of 38th International Conference on Computers and Their Applications, Tóquio, Japão, ano 2023, v. 91, p. 124-137, 22 mar. 2023. DOI: 10.29007/d931. Disponível em: https://easychair.org/publications/paper/szdf. Acesso em: 10 abr. 2026.

VASWANI, Ashish; SHAZEER, Noam; PARMAR, Niki; USZKOREIT, Jakob; JONES, Llion; GOMEZ, Aidan N.; KAISER, Łukasz; POLOSUKHIN, Illia. Attention is all you need. In: ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS (NeurIPS), 31., 2017. Proceedings [...]. [S. l.]: NeurIPS, 2017. p. 5998–6008. DOI: 10.48550/arXiv.1706.03762. Disponível em: https://arxiv.org/abs/1706.03762. Acesso em: 11 mar. 2026.

VERGNA, Amanda Menezes; SILVA, Saulo Cardoso Malbar da Silva; AMORIM, Fábio Luiz Alves de. Fake news e a sociedade contemporânea: Os avanços das notícias falsas e os impactos na área da saúde. **destarte**, Brasil, ano 2022, v. 11, n. 2, p. 131-146, dez. 2022. Disponível em: https://estacio.periodicoscientificos.com.br/index.php/destarte/article/download/1507/1250. Acesso em: 10 maio 2026.

VIEIRA, Camila B.; SOUZA, José Vinicius de S.; CAVALCANTI, George D. C.. **Detecção de Fake News em Português: Análise Comparativa entre Métodos de Representação em Português, Inglês e Multilíngues**. In: BRAZILIAN WORKSHOP ON SOCIAL NETWORK ANALYSIS AND MINING (BRASNAM), 14., 2025, Maceió/AL. Anais [...]. Porto Alegre: Sociedade Brasileira de Computação, 2025. p. 187-199. ISSN 2595-6094. DOI: 10.5753/brasnam.2025.9062. Disponível em: https://sol.sbc.org.br/index.php/brasnam/article/view/36378. Acesso em: 5 abr. 2026.

WARDLE, Claire; DERAKHSHAN, Hossein. **Information disorder: Toward an interdisciplinary framework for research and policymaking.** Strasbourg: Council of Europe, 2017. Disponível em: https://edoc.coe.int/en/media/7495-information-disorder-toward-an-interdisciplinary-framework-for-research-and-policy-making.html. Acesso em: 14 fev. 2026.