# Detecção de Fake News em Domínios Cruzados: Uma Revisão Sistemática

**Rafael R. Braz**[1], **Luciano A. Digiampietri**[1]

[1]Escola de Artes, Ciências e Humanidades – Universidade de São Paulo (USP)

{rafael.braz,digiampietri}@usp.br

***Abstract.*** *This paper presents a systematic review on the detection of fake news in cross-domain settings, where the challenge is to identify misinformation across various contexts, such as different themes, languages, or sources. The review reveals a preference for Domain Generalization (DG), which seeks to develop models capable of identifying fake news across a wide range of contexts without specific adjustments, over Domain Adaptation (DA), which aims to optimize the performance of a model trained in a source domain for specific target domains. The diversity of the datasets used underscores the need for standardized benchmarks for consistent evaluations. The study suggests exploring new techniques for domain generalization and adaptation to enhance the detection of fake news in different contexts.*

***Resumo.*** *Este artigo apresenta uma revisão sistemática sobre a detecção de fake news em domínios cruzados, em que o desafio é identificar desinformação em contextos variados, como diferentes temas, idiomas ou fontes. A revisão revela uma preferência pela Generalização de Domínio (DG), que busca desenvolver modelos capazes de identificar fake news em uma ampla gama de contextos sem ajustes específicos, em detrimento da Adaptação de Domínio (DA), que visa otimizar o desempenho de um modelo treinado em um domínio fonte para domínios-alvo específicos. A diversidade dos conjuntos de dados utilizados ressalta a necessidade de benchmarks padronizados para avaliações consistentes. O estudo sugere a exploração de novas técnicas de generalização e adaptação de domínio para aprimorar a detecção de fake news em diferentes contextos.*

## 1. Introdução

A propagação de notícias falsas, comumente referidas como "*fake news*", nas plataformas digitais, tem se revelado um problema crescente com profundas implicações para a democracia, saúde pública e liberdade de expressão [Allcott and Gentzkow 2017], desencadeando um debate global sobre a necessidade de mecanismos eficazes de detecção e contenção [Lazer et al. 2018, Ferreira et al. 2021]. Diante desse cenário, a detecção automatizada de *fake news* emergiu como um domínio crítico de pesquisa, visando atenuar os danos causados por essa propagação [Shu et al. 2017]. As abordagens para combater esse fenômeno são variadas, englobando desde a análise do conteúdo até a verificação da credibilidade das fontes, passando pelo exame dos padrões de disseminação e estilos de escrita [Zhou and Zafarani 2020]. Contudo, a capacidade de detectar *fake news* em contextos que atravessam fronteiras temáticas, culturais ou linguísticas — conhecida como detecção em domínios cruzados — apresenta desafios distintos, exigindo técnicas que possam identificar desinformação de forma ampla e adaptável.

Uma das maiores complexidades no estudo da disseminação de informações falsas na Internet está na definição e diferenciação entre os termos "*fake news*", "rumor" e "desinformação". Enquanto "*Fake news*" geralmente se referem a notícias fabricadas com a intenção de enganar, "rumores" podem ser entendidos como informações não verificadas que circulam entre o público, e "desinformação" abrange uma gama mais ampla de informações falsas ou enganosas, independente da intenção. Diferentes trabalhos acadêmicos adotam essas definições de maneiras variadas, o que pode levar a uma falta de uniformidade na compreensão e categorização dos estudos sobre o tema [Wardle and Derakhshan 2017]. Diante dessa variedade de interpretações, esta revisão sistemática opta por não estabelecer distinções estritas entre esses termos para a seleção de artigos, buscando incluir um espectro mais amplo de pesquisas sobre desinformação, *fake news* e rumores, de modo a fornecer uma análise inclusiva e compreensiva dos esforços para detectar e combater a disseminação de informações falsas em múltiplos contextos.

Embora a diferenciação entre informações verídicas e falsas constitua um desafio central na era digital, a abordagem tradicional de análise baseada exclusivamente em características superficiais, como a frequência de palavras ou o estilo de escrita, frequentemente não captura a complexidade do fenômeno em sua totalidade. Essas características tendem a se confinar a domínios ou idiomas específicos, reduzindo sua eficácia em uma gama mais ampla de cenários. Em contraste, a detecção de *fake news* em domínios cruzados busca identificar características latentes independentes de domínio, aproveitando a capacidade de generalização de modelos e algoritmos para reconhecer *fake news* em diversos contextos [Yu et al. 2022].

Esta revisão sistemática tem como objetivo principal compilar e analisar as metodologias, técnicas e desafios associados à detecção de *fake news* em domínios cruzados. Em um cenário marcado pela sofisticação crescente das *fake news* e pela expansão das plataformas digitais, torna-se crucial compreender como diferentes abordagens podem ser ajustadas ou reconfiguradas para enfrentar a desinformação em variados cenários.

Neste trabalho, são abordadas as principais representações de dados empregadas, as técnicas predominantes de extração de dados, os conjuntos de dados mais utilizados, os classificadores adotados e seus respectivos impactos nos resultados da detecção de *fake news*. Ademais, serão destacadas as adaptações dessas metodologias para a detecção em ambientes de domínios cruzados, considerando diferentes tópicos, línguas, fontes de notícias ou plataformas de redes sociais.

Especificamente, será investigada a prevalência e a eficácia das abordagens de Adaptação de Domínio (DA) e Generalização de Domínio (DG) dentro do corpus de estudos selecionados. Enquanto a Adaptação de Domínio busca otimizar o desempenho do modelo em domínios-alvo específicos, valendo-se de conhecimentos obtidos de domínios-fonte, a Generalização de Domínio procura desenvolver modelos com a capacidade de performar adequadamente em qualquer domínio-alvo, inclusive aqueles não contemplados durante o treinamento [Ben-David et al. 2010, Wang et al. 2023b]. Assim, esta análise contribuirá para elucidar a versatilidade e eficiência das técnicas correntes de detecção de *fake news*, apontando caminhos para inovações e melhorias futuras nesse campo vital de pesquisa.

## 2. Método

Diante do desafio de elucidar os mecanismos de detecção de notícias falsas em domínios variados, esta revisão sistemática [Kitchenham 2004] é guiada por um protocolo concebido para garantir uma abordagem ampla na coleta de dados, habilitando uma análise das metodologias e dos desafios presentes na identificação de notícias falsas em ambientes de domínio cruzado. As questões de pesquisa que orientam este inquérito acadêmico são:

1. **Análise das Estratégias e Tecnologias Aplicadas:** Quais são as estratégias e tecnologias empregadas na detecção de notícias falsas em contextos de domínios cruzados?
2. **Relevância de Conjuntos de Dados:** Quais os conjuntos de dados utilizados na detecção de notícias falsas em domínios cruzados?
3. **Utilização de Características, Modalidades e Embeddings de Texto:** Quais tipos de características, modalidades de dados e técnicas de embeddings textuais são utilizados, e como estes contribuem para a detecção de notícias falsas?
4. **Aplicação de Técnicas de Adaptação de Domínio (DA) e Generalização de Domínio (DG):** De que maneira as metodologias de Adaptação de Domínio e Generalização de Domínio são empregadas nos estudos sobre detecção de *fake news*, e qual o impacto dessas técnicas na capacidade dos modelos para operar eficazmente em múltiplos domínios?
5. **Identificação de Lacunas e Direções para Pesquisas Futuras:** Quais lacunas na literatura atual sobre detecção de notícias falsas em domínios cruzados podem ser identificadas, sugerindo áreas para pesquisas futuras?

Este protocolo estabelece a metodologia para a revisão sistemática, orientando a exploração do conhecimento atual sobre a detecção de notícias falsas em diferentes domínios. O objetivo é consolidar o entendimento existente e identificar áreas ainda não exploradas que possam contribuir para o avanço da pesquisa.

### 2.1. Estratégia de Busca e Fontes de Informação

Inicialmente, a estratégia de busca foi definida visando a cobrir uma ampla gama da literatura relevante. Para isso, duas bases de dados acadêmicas de destaque da área foram selecionadas: ACM Digital Library[1] e IEEE Xplore[2]. A escolha dessas plataformas foi motivada pela sua extensa cobertura de publicações nas áreas de ciência da computação, engenharia e tecnologias da informação, fundamentais para a pesquisa em detecção de *fake news*. Para a seleção dos estudos, a *string* de busca, enunciada abaixo, foi aplicada em cada mecanismo de busca selecionado:

```
("social" OR "facebook" OR "twitter") AND
("fake news" OR "desinformation" OR "misinformation"
    OR "hoax" OR "rumor" OR "rumour") AND
("detection" OR "identification" OR "classification")
AND ("domain" OR "unseen")
```

Esta *string* de busca foi aplicada aos artigos publicados entre 1 de janeiro de 2018 e 17 de janeiro de 2024, data esta em que a busca foi realizada nas bases.

[1]https://dl.acm.org/

[2]https://ieeexplore.ieee.org/

## 2.2. Critérios de Inclusão e Exclusão

A seleção dos estudos foi guiada por um conjunto de critérios de inclusão e exclusão definidos. Para serem incluídos, os trabalhos deveriam focar de forma explícita na detecção de *fake news* em contextos de domínios cruzados, relatar resultados empíricos obtidos por meio de métodos de detecção e serem publicados em inglês em periódicos ou conferências revisadas por pares. Por outro lado, foram excluídos da análise estudos que se limitavam a publicações breves, resumos de conferências, comentários editoriais e revisões de literatura ou trabalhos indisponíveis em formato de artigo completo. Também foram excluídos trabalhos que não se concentravam diretamente no fenômeno de domínios cruzados, ou seja, que não empregassem técnicas de adaptação ou generalização de domínio. Abaixo estão listados os critérios de exclusão utilizados:

a. Estudos que não estejam disponíveis na íntegra para acesso ou se referem a trabalho em andamento;
b. Estudos que não se concentram na detecção automatizada de informações falsas;
c. Estudos sobre detecção de notícias falsas fora do ambiente de redes sociais digitais;
d. Estudos que não são empíricos, como revisões, opiniões ou comentários;
e. Estudos que não fornecem informações suficientes sobre as abordagens de detecção automatizada utilizadas ou não apresentam métricas de desempenho;
f. Estudos que não envolvem ou não se concentram na classificação em contextos de domínio cruzado (*cross-domain*);
g. Estudos que apresentem duplicidade, independentemente de serem encontrados dentro da mesma base de dados ou por já terem sido selecionados de uma base distinta.

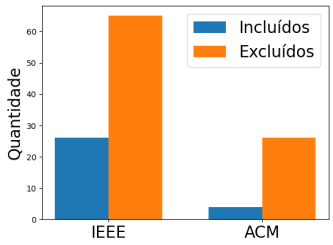
## 2.3. Processo de Seleção dos Estudos

O processo de seleção dos estudos foi realizado em duas etapas. Primeiro, uma triagem preliminar dos títulos e resumos aplicou os critérios de inclusão e exclusão na ordem enunciada, eliminando artigos não relevantes. Em seguida, uma análise minuciosa do texto completo dos trabalhos selecionados verificou sua pertinência ao tema de investigação.

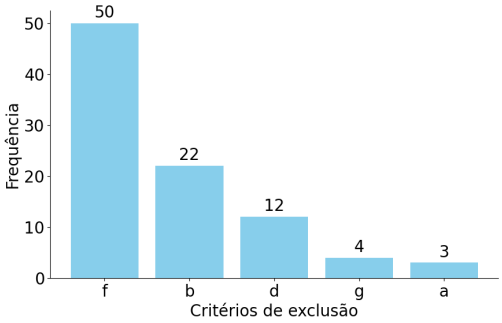
## 2.4. Extração e Síntese dos Dados

A partir da literatura selecionada, foi realizada a extração de informações essenciais, tais como autor(es), ano de publicação, objetivos, metodologias empregadas, conjuntos de dados utilizados, principais resultados e conclusões. Essas informações proporcionaram a base para uma síntese qualitativa, permitindo a identificação de padrões, temas comuns e divergências entre os estudos analisados. Especial atenção foi dedicada às metodologias de detecção empregadas, analisando os obstáculos enfrentados na generalização dos modelos em cenários de múltiplos domínios.

Essa abordagem visa a assegurar que esta revisão não somente sintetize o estado atual do conhecimento sobre a detecção de *fake news* em domínios cruzados, mas também identifique lacunas críticas no conhecimento existente, delineando caminhos promissores para pesquisas futuras.




**(a) Distribuição dos artigos por base de dados.**




**(b) Frequência de ocorrência dos critérios de exclusão.**

**Figura 1. Distribuição dos artigos por base de dados e frequência de ocorrência dos critérios de exclusão.**

## 3. Condução

A busca inicial nas duas bibliotecas digitais identificou um total de 121 artigos, 91 na IEEE e 30 na ACM. Contudo, após a aplicação dos critérios de inclusão e exclusão, o conjunto foi refinado para um total de 30 artigos escolhidos para análise detalhada. Destes, 26 foram selecionados na IEEE, representando 28,57% dos artigos iniciais, e 4 na ACM, correspondendo a 13,33%. A distribuição dos artigos incluídos e excluídos em cada base de dados é apresentada na Figura 1a.

O critério predominante para a exclusão de artigos foi a falta de foco direto em domínios cruzados. Uma representação dos critérios de exclusão aplicados e a frequência com que cada um impactou a seleção final dos estudos são apresentados na Figura 1b. Para uma visualização simplificada, nos casos em que um artigo satisfazia múltiplos critérios de exclusão, apenas o primeiro critério identificado foi considerado para esta representação. A Figura 2 apresenta a distribuição dos artigos selecionados por ano de publicação. Observa-se a concentração de artigos a partir de 2021.

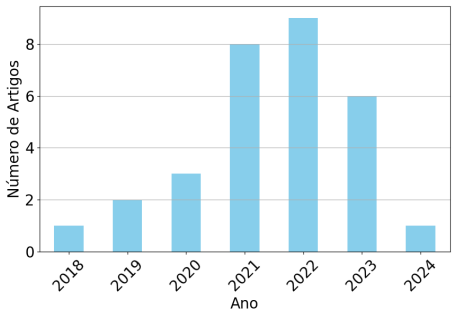



**Figura 2. Artigos selecionados por ano de publicação.**

## 4. Resultados

O propósito desta seção é abordar as questões formuladas no protocolo da revisão sistemática, utilizando como base os 30 artigos selecionados e resumidos na Tabela 1.

**Tabela 1. Resumo dos artigos selecionados**

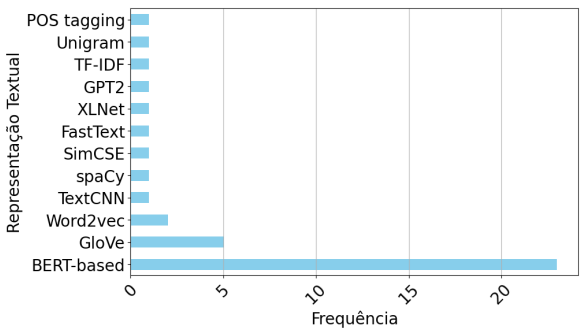
| Artigo | Características | Modalidades | Embeddings | Proc. de Imagem | Datasets | DA/DG |
|---|---|---|---|---|---|---|
| [Birunda and Devi 2021] | Conteúdo | Texto | TF-IDF | - | Kaggle | DG |
| [Blackledge and Atapour-Abarghouei 2021] | Conteúdo | Texto | BERT, DistilBERT, DeBERTa | - | ISOT, Combined Corpus | DG |
| [Gautam and Jerripothula 2020] | Conteúdo, gramática, sentimentos, handcrafted | Texto | GloVe | - | FakeNewsAMT, Celebrity | DG |
| [Goel et al. 2021] | Conteúdo | Texto | BERT, RoBERTa, XLNet, DeBERTa, GPT2 | - | FakeNewsAMT, Celebrity | DG |
| [Guo et al. 2022] | Conteúdo | Texto, imagem | BERT | Resnet-50 | Twitter, Weibo | DG |
| [Han et al. 2019] | Conteúdo | Texto | Word2vec, ELMo | - | PHEME, CrisisLexT26, CREDBANK | DG |
| [Joshi et al. 2023] | Conteúdo | Texto | GloVe | - | CoAID, MiSoVac | DG |
| [Kato et al. 2022] | Conteúdo | Texto | BERT | - | FakeNewsAMT | DG |
| [Kim et al. 2020] | Conteúdo, Propagação, Usuário (handcrafted) | Texto | Unigram, POS tagging | - | RumourEval2019, PHEME | DG |
| [Kong et al. 2023] | Conteúdo (handcrafted) | Texto | - | - | Fake.my-COVID19, Covid-19, BRACIS2019 | DG |
| [Li et al. 2021] | Conteúdo | Texto | RoBERTa | - | FakeNewsNet, Health DETERRENTS | DA |
| [Liu et al. 2024] | Conteúdo | Texto, imagem | TextCNN-roberta | ResNet50 | PHEME, Twitter | DG |
| [Lu et al. 2022] | Conteúdo, conhecimento externo, comentários | Texto | GloVe | - | Twitter, Weibo | DG |
| [Mosallanezhad et al. 2022] | Conteúdo, comentários, user interactions (vetor binário), | Texto | BERT | - | FakeNewsNet | DA |
| [Omrani et al. 2023] | Conteúdo | Texto | BERT | - | Covid-19 | DG |
| [Rastogi et al. 2021] | Conteúdo, Sentimento, Polaridade, handcrafted | Texto | BERT (não detalhado para CNN e LSTM) | - | FakeBan, PHEME, CoAID | DG |

| | | | | | | |
|---|---|---|---|---|---|---|
| [Shang et al. 2022] | Conteúdo, Artigos de notícias e fact-checking | Texto | BERT | - | Covid: Constraint, COVIDRumor, MMCoVaR, ANTiVax | DA |
| [Sicilia et al. 2021] | Conteúdo, Usuário, Rede (handcrafted) | Texto | Word2vec | - | Zika virus dataset, Vaccine dataset | DA |
| [Suprem et al. 2022] | Conteúdo | Texto | BERT, AlBERT, COVID-Twitter-BERT | - | Kaggle, CoAID, Covid19FN, Covid-19, COVID-CQ, CMU-MisCOV19, Covid-19-Rumor | DG |
| [Tang et al. 2023] | Conteúdo, Histórico e Interações | Texto | spaCy | - | Twitter | DA |
| [Wang et al. 2018] | Conteúdo | Texto, imagem | Embedding pré-treinado, não especificado | VGG-19 | Twitter, Weibo | DG |
| [Wang et al. 2021] | Conteúdo | Texto, imagem | FastText | VGG-19 | Twitter, Weibo | DA |
| [Wang et al. 2023a] | Conteúdo | Texto | BERT | - | Weibo21, Thu | DG |
| [Wang et al. 2022] | Conteúdo | Texto, imagem | AlBERT | VGG-19 | Twitter, Weibo | DG |
| [Yu et al. 2022] | Conteúdo, notícias | Texto | BERT, SimCSE para histórico | - | Weibo | DG |
| [Zeng et al. 2022] | Conteúdo | Texto | RoBERTa | - | GossipCop, LIAR, PHEME, Constraint, ANTiVax | DA |
| [Zhang et al. 2019] | Conteúdo, conhecimento | Texto, imagem | Glove | VGG-19 | Twitter, PHEME | DG |
| [Zhang et al. 2020] | Conteúdo, Site URL | Texto, imagem | BERT | VGG-19 | Twitter, Weibo | DG |
| [Zhang et al. 2021] | Conteúdo | Texto, imagem | GloVe | CNN | PHEME, PHEME$_{veracity}$ | DA |
| [Zhu et al. 2023] | Sentimentos, Estilo (handcrafted) | Texto | RoBERTa | - | FakeNews-Net, COVID | DG |

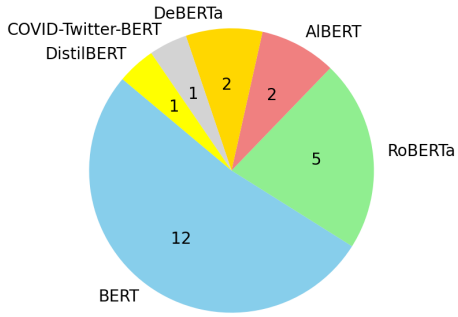
## 4.1. Respostas às Questões de Pesquisa

### 4.1.1. Estratégias e Tecnologias Aplicadas

A análise dos 30 artigos selecionados revelou uma tendência crescente no uso de técnicas de aprendizado profundo para a detecção de *fake news* em domínios cruzados. Uma constatação é que todas as pesquisas incorporaram a análise de conteúdo textual. Notavelmente, a maioria dos estudos empregou variações de modelos BERT, como ilustrado nas Figuras 3a e 3b, evidenciando a robustez dessa arquitetura na compreensão e análise de conteúdo textual complexo em diversas línguas e domínios. Contudo, apenas 8 dentre os 30 estudos também exploraram a modalidade de imagem como parte integrante de suas abordagens de detecção. Dentre estes, a maioria (5 de 8) adotou o modelo pré-treinado VGG-19 para extração de características das imagens. O uso prevalente do VGG-19, um modelo de 16 camadas convolucionais que empregam filtros de tamanho 3x3 intercaladas com camadas de *max pooling*, destaca sua eficácia em capturar informações visuais complexas, e ressalta a importância de características visuais detalhadas na detecção de desinformação.




**(a) Representações textuais utilizadas.**




**(b) Embeddings baseados em BERT.**

**Figura 3. Distribuição das formas de representação textual utilizadas.**

Por se tratarem de modelos de aprendizagem profunda, em geral o classificador em si é composto por algumas camadas do tipo *feed-forward* seguidas por uma função *softmax*, com os esforços direcionados a produzir características latentes independentes de domínio. Entre as exceções temos [Kim et al. 2020] com o uso de *ensembles* de algoritmos clássicos de aprendizado de máquina e [Kong et al. 2023] que utiliza de programação genética para produzir uma equação matemática para detecção de *fake news*.

### 4.1.2. Relevância de Conjuntos de Dados

A diversidade dos conjuntos de dados utilizados nas pesquisas destaca-se como um obstáculo significativo para a construção de *benchmarks* padronizados e a comparação direta dos resultados entre os diferentes estudos. Essa variedade, embora enriqueça a pesquisa ao abranger múltiplas facetas da desinformação, introduz complexidade adicional na avaliação da eficácia comparativa das metodologias de detecção de *fake news*, limitando a capacidade de estabelecer conclusões amplamente aplicáveis.

Os conjuntos de dados mais utilizados foram extraídos da plataforma de mídia social Twitter (recentemente renomeada para X) seguidos pelos extraídos da rede social chinesa de *microblogging* Weibo. A recente utilização desses conjuntos os estabelece como os mais propícios a serem usados em *benchmarking* para a tarefa de classificação de *fake news* em domínios cruzados. Contudo, é notável que, apesar de sua popularidade, os estudos frequentemente adotam abordagens distintas na manipulação desses conjuntos de dados, incorporando diferentes tipos de informações adicionais, como em [Tang et al. 2023, Lu et al. 2022, Zhang et al. 2019].

É importante considerar a variação na definição de domínio de cada trabalho. Alguns estudos focam em conjuntos de dados multilíngues como forma de variação de domínio [Omrani et al. 2023], enquanto outros examinam como as distribuições se comportam em diferentes tópicos de notícias, como política, saúde e celebridades [Rastogi et al. 2021]. A variação dentro de eventos distintos, como Covid-19 e Monkeypox, também é considerada, expondo a dificuldade de encontrar um modelo abrangente o suficiente para funcionar com todos os tipos de fake news, inclusive as emergentes [Shang et al. 2022]. Além disso, vários estudos analisados concentraram-se na detecção de *fake news* relacionadas à pandemia da Covid-19 [Zhu et al. 2023, Omrani et al. 2023, Kong et al. 2023, Kim et al. 2020, Shang et al. 2022, Rastogi et al. 2021], evidenciando a necessidade de combater notícias falsas emergentes em tempos de crise de saúde pública.

Observa-se uma marcante predominância de conjuntos de dados em inglês e chinês nas pesquisas analisadas. Essa tendência evidencia um campo de oportunidade para enriquecer as investigações com uma maior diversidade linguística, questão abordada apenas em [Kong et al. 2023, Omrani et al. 2023]. Ao incorporar conjuntos de dados multilíngues, futuras pesquisas poderiam adotar uma perspectiva mais global, ampliando a eficácia das estratégias de combate às *fake news*.

### 4.1.3. Características, Modalidades e Embeddings de Texto

Conforme discutido na Seção 4.1.1 e ilustrado na Tabela 1, os estudos analisados abordaram as modalidades de texto e imagem. Para a modalidade textual, é visível a crescente adoção de *embeddings* baseados em BERT a partir de 2020, incluindo o RoBERTa, AlBERT, DeBERTa e DistilBERT. Apesar do surgimento e popularização do BERT, o GloVe se destaca entre as representações textuais por manter um desempenho competitivo em diversas aplicações de processamento de língua natural (PLN).

Na modalidade de imagem, o uso do VGG-19 prevaleceu, citado em diversos trabalhos [Wang et al. 2022, Wang et al. 2021, Zhang et al. 2020, Zhang et al. 2019, Wang et al. 2018]. Alternativamente, ResNet50 [Liu et al. 2024, Guo et al. 2022] e redes neurais convolucionais [Zhang et al. 2021] também foram empregadas.

Para incrementar a precisão na detecção de *fake news*, algumas estratégias incluíram a incorporação de informações adicionais, tais como a credibilidade das fontes de notícias [Birunda and Devi 2021], análises de conteúdo histórico e interações de usuários [Tang et al. 2023], bem como o aproveitamento de bases de conhecimento [Zhang et al. 2019, Lu et al. 2022]. Ainda, características como sentimentos, polaridade e atributos manualmente definidos continuam a ser relevantes [Zhu et al. 2023, Kong et al. 2023, Ras-

togi et al. 2021, Sicilia et al. 2021, Gautam and Jerripothula 2020, Kim et al. 2020].

### 4.1.4. Adaptação e Generalização de Domínio

A técnica de Generalização de Domínio (DG) foi a mais utilizada, aplicada em 22 dos 30 artigos. Em menor medida, a Adaptação de Domínio (DA) foi empregada em 8 estudos. Na literatura analisada, o uso de técnicas de DG é evidenciado pelo emprego de modelos como o EANN [Wang et al. 2018], baseado no conceito de Rede Neural Adversarial de Domínio, que utiliza uma função de perda adversarial com *Gradient Reversal Layer* para treinamento, buscando obter características latentes independentes de domínio. Alguns outros artigos se utilizam desse mecanismo de gradiente reverso em um classificador de domínio [Zhang et al. 2020, Wang et al. 2022, Joshi et al. 2023, Guo et al. 2022].

Uma segunda estratégia inovadora adotada é o uso de estruturas de memória para domínios. Um exemplo utilizado é uma rede de memória de eventos para armazenar e obter características invariantes por evento [Zhang et al. 2019]. Similarmente, [Zhu et al. 2023] utiliza um banco de memória de domínio que, ao ser utilizado em conjunto com rótulos de domínio múltiplos, facilita a adaptação de informações discriminatórias entre notícias de vários domínios.

O uso de conhecimento externo para tornar os modelos mais adaptáveis entre domínios é uma abordagem adotada diretamente [Lu et al. 2022], ou indiretamente baseada na capacidade de transferência de aprendizado de modelos como o BERT para obter características independentes de domínio [Goel et al. 2021, Rastogi et al. 2021].

Artigos mais simples não abordam diretamente a questão da generalização de domínio, mas se baseiam em características manualmente definidas [Birunda and Devi 2021, Gautam and Jerripothula 2020, Kim et al. 2020, Kong et al. 2023]. Outros simplesmente tentam se esquivar do problema, removendo artigos opinativos [Blackledge and Atapour-Abarghouei 2021], ou por meio de uma opção de rejeição baseada em *K-Means* [Suprem et al. 2022]. Uma outra questão interessante observada é a melhora na capacidade de generalização de modelos ao serem treinados utilizando notícias falsas e verdadeiras pareadas [Kato et al. 2022].

Semelhante a vários estudos no campo de Generalização de Domínio (DG), [Shang et al. 2022, Zhang et al. 2021] utilizam de uma função de perda adversarial para extrair características invariantes de domínio, mas para a função de adaptação entre domínios médicos, divergindo nas características adotadas. De maneira similar, [Li et al. 2021] também utiliza treinamento adversarial, mas incorporando conhecimento prévio do domínio alvo por meio de rótulos fracos. O estudo [Tang et al. 2023], por sua vez, utiliza de técnicas de transporte ótimo para alinhar as distribuições entre o domínio fonte e o domínio alvo, enquanto [Zeng et al. 2022] utiliza de aprendizado por contraste com objetivo similar. Analogamente, [Mosallanezhad et al. 2022] utiliza de aprendizado por reforço para transpor o espaço de características do domínio fonte para o alvo. Em outra abordagem, [Wang et al. 2021] utiliza meta-aprendizado e métodos de processo neural para lidar efetivamente com a detecção de notícias falsas em eventos emergentes com poucos dados rotulados

As abordagens distintas de adaptação e generalização de domínio oferecem van-

tagens e desvantagens. Enquanto a DG proporciona uma maior flexibilidade e potencial para generalização, a DA permite uma adaptação mais direcionada e eficaz a domínios específicos. A escolha entre DG e DA dependerá das necessidades específicas do sistema de detecção de *fake news* e do contexto em que será aplicado.

### 4.1.5. Lacunas e Direções Futuras

A análise dos estudos selecionados destacou algumas lacunas e direções promissoras para futuras pesquisas na área de detecção de *fake news* em domínios cruzados. Uma das principais dificuldades identificadas é a diversidade dos conjuntos de dados utilizados nos estudos, o que dificulta a comparação direta dos resultados e a avaliação da eficácia das abordagens de detecção. Há uma necessidade de estabelecer *benchmarks* padronizados que permitam uma avaliação consistente e comparável das técnicas de detecção de *fake news* em diferentes domínios.

Além disso, observou-se que os modelos baseados no BERT dominaram a paisagem atual das pesquisas, mas ainda existe espaço para explorar novas arquiteturas de aprendizado profundo. O uso de modelos generativos avançados, como o GPT-4, Gemini e Llama, pode oferecer representações textuais mais profundas e gerar dados sintéticos para enriquecer o treinamento e a avaliação dos sistemas de detecção de *fake news*. Essas inovações têm o potencial de aprimorar significativamente a capacidade de identificar e combater a disseminação de notícias falsas em diferentes contextos. Isso se alinha com a abordagem de generalização de domínio por manipulação de dados [Wang et al. 2023b], ao manipular a entrada para facilitar o aprendizado de representações genéricas.

Diversos estudos já exploraram o aprendizado de representações invariantes de domínio, mas métodos menos investigados incluem técnicas de Kernel e Minimização do Risco Invariante, além de estratégias como meta-aprendizado e operações no gradiente que ainda podem ser mais desenvolvidas[Wang et al. 2023b].

## 5. Conclusão

Neste estudo, foi realizada uma revisão sistemática para explorar o estado atual da detecção de *fake news* em domínios cruzados, um desafio crescente na era da desinformação digital. Por meio da análise de 30 artigos selecionados, foram identificadas tendências predominantes, desafios recorrentes e oportunidades futuras na pesquisa nesta área vital.

Observa-se uma clara preferência pelo uso de técnicas de aprendizado profundo, particularmente modelos baseados em BERT, para a análise de conteúdo textual na detecção de *fake news*. No entanto, a diversidade dos conjuntos de dados utilizados nos estudos ressaltou a necessidade de estabelecer *benchmarks* padronizados para facilitar comparações diretas e avaliações mais consistentes das abordagens de detecção.

Embora a modalidade textual tenha sido amplamente explorada, a inclusão de informações visuais por meio de imagens ainda é subutilizada, apontando para uma área de potencial desenvolvimento futuro. Além disso, a integração de conhecimento externo e o uso de modelos generativos foram identificados como estratégias promissoras para melhorar a adaptabilidade e a generalização dos modelos de detecção em diferentes contex-

tos. Ao abordar as lacunas identificadas e explorar as direções futuras propostas, espera-se contribuir para o avanço da pesquisa na detecção de *fake news* em domínios cruzados.

## Referências


Allcott, H. and Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, 31:211–236.

Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., and Vaughan, J. (2010). A theory of learning from different domains. *Machine Learning*, 79:151–175.

Birunda, S. S. and Devi, R. K. (2021). A novel score-based multi-source fake news detection using gradient boosting algorithm. In *2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS)*, pages 406–414.

Blackledge, C. and Atapour-Abarghouei, A. (2021). Transforming fake news: Robust generalisable news classification using transformers. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 3960–3968.

Ferreira, G., Santos, B., do Ó, M., Braz, R., and Digiampietri, L. (2021). Social bots detection in brazilian presidential elections using natural language processing. *Anais do Simpósio Brasileiro de Sistemas de Informação (SBSI)*.

Gautam, A. and Jerripothula, K. R. (2020). Sgg: Spinbot, grammarly and glove based fake news detection. In *2020 IEEE Sixth International Conference on Multimedia Big Data (BigMM)*, pages 174–182.

Goel, P., Singhal, S., Aggarwal, S., and Jain, M. (2021). Multi domain fake news analysis using transfer learning. In *2021 5th International Conference on Computing Methodologies and Communication (ICCMC)*, pages 1230–1237.

Guo, Y., Ge, H., and Li, J. (2022). Fake news detection based on two-branch network and domain adversarial. In *2022 IEEE 5th International Conference on Computer and Communication Engineering Technology (CCET)*, pages 172–176.

Han, S., Gao, J., and Ciravegna, F. (2019). Neural language model based training data augmentation for weakly supervised early rumor detection. In *2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 105–112.

Joshi, G., Srivastava, A., Yagnik, B., Hasan, M., Saiyed, Z., Gabralla, L. A., Abraham, A., Walambe, R., and Kotecha, K. (2023). Explainable misinformation detection across multiple social media platforms. *IEEE Access*, 11:23634–23646.

Kato, S., Yang, L., and Ikeda, D. (2022). Domain bias in fake news datasets consisting of fake and real news pairs. In *2022 12th International Congress on Advanced Applied Informatics (IIAI-AAI)*, pages 101–106.

Kim, Y., Kim, H. K., Kim, H., and Hong, J. B. (2020). Do many models make light work? evaluating ensemble solutions for improved rumor detection. *IEEE Access*, 8:150709–150724.

Kitchenham, B. (2004). Procedures for performing systematic reviews. *Keele, UK, Keele University*, 33(2004):1–26.

Kong, J. T. H., Wong, W. K., Juwono, F. H., and Apriono, C. (2023). Generating fake news detection model using a two-stage evolutionary approach. *IEEE Access*, 11:85067–85085.

Lazer, D., Baum, M., Benkler, Y., Berinsky, A., Greenhill, K., Menczer, F., Metzger, M., Nyhan, B., Pennycook, G., Rothschild, D., Schudson, M., Sloman, S., Sunstein, C., Thorson, E., Watts, D., and Zittrain, J. (2018). The science of fake news. *Science*, 359:1094–1096.

Li, Y., Lee, K., Kordzadeh, N., Faber, B., Fiddes, C., Chen, E., and Shu, K. (2021). Multi-source domain adaptation with weak supervision for early fake news detection. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 668–676.

Liu, H., Wang, W., Sun, H., Rocha, A., and Li, H. (2024). Robust domain misinformation detection via multi-modal feature alignment. *IEEE Transactions on Information Forensics and Security*, 19:793–806.

Lu, M., Huang, Z., Li, B., Zhao, Y., Qin, Z., and Li, D. (2022). Sifter: A framework for robust rumor detection. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:429–442.

Mosallanezhad, A., Karami, M., Shu, K., Mancenido, M. V., and Liu, H. (2022). Domain adaptive fake news detection via reinforcement learning. In *Proceedings of the ACM Web Conference 2022*, WWW '22, page 3632–3640, New York, NY, USA. Association for Computing Machinery.

Omrani, P., Ebrahimian, Z., Toosi, R., and Akhaee, M. A. (2023). Bilingual covid-19 fake news detection based on lda topic modeling and bert transformer. In *2023 6th International Conference on Pattern Recognition and Image Analysis (IPRIA)*, pages 01–06.

Rastogi, S., Gill, S. S., and Bansal, D. (2021). An adaptive approach for fake news detection in social media: Single vs cross domain. In *2021 International Conference on Computational Science and Computational Intelligence (CSCI)*, pages 1401–1405.

Shang, L., Zhang, Y., Yue, Z., Choi, Y., Zeng, H., and Wang, D. (2022). A knowledge-driven domain adaptive approach to early misinformation detection in an emergent health domain on social media. In *2022 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 34–41.

Shu, K., Sliva, A., Wang, S., Tang, J., and Liu, H. (2017). Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter*, 19.

Sicilia, R., Francini, L., and Soda, P. (2021). Representation and knowledge transfer for health-related rumour detection. In *2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS)*, pages 591–596.

Suprem, A., Vaidya, S., and Pu, C. (2022). Exploring generalizability of fine-tuned models for fake news detection. In *2022 IEEE 8th International Conference on Collaboration and Internet Computing (CIC)*, pages 82–88.

Tang, W., Ma, Z., Sun, H., and Wang, J. (2023). Learning sparse alignments via optimal transport for cross-domain fake news detection. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Wang, D., Zhang, W., Wu, W., and Guo, X. (2023a). Soft-label for multi-domain fake news detection. *IEEE Access*, 11:98596–98606.

Wang, J., Lan, C., Liu, C., Ouyang, Y., Qin, T., Lu, W., Chen, Y., Zeng, W., and Yu, P. S. (2023b). Generalizing to unseen domains: A survey on domain generalization. *IEEE Transactions on Knowledge & Data Engineering*, 35(08):8052–8072.

Wang, X., Li, X., Liu, X., and Cheng, H. (2022). Using albert and multi-modal circulant fusion for fake news detection. In *2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 2936–2942.

Wang, Y., Ma, F., Jin, Z., Yuan, Y., Xun, G., Jha, K., Su, L., and Gao, J. (2018). Eann: Event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, page 849–857, New York, NY, USA. Association for Computing Machinery.

Wang, Y., Ma, F., Wang, H., Jha, K., and Gao, J. (2021). Multimodal emergent fake news detection via meta neural process networks. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, KDD '21, page 3708–3716, New York, NY, USA. Association for Computing Machinery.

Wardle, C. and Derakhshan, H. (2017). *Information disorder: Toward an interdisciplinary framework for research and policymaking*, volume 27. Council of Europe Strasbourg.

Yu, W., Ge, J., Yang, Z., Dong, Y., Zheng, Y., and Dai, H. (2022). Multi-domain fake news detection for history news environment perception. In *2022 IEEE 17th Conference on Industrial Electronics and Applications (ICIEA)*, pages 428–433.

Zeng, H., Yue, Z., Kou, Z., Shang, L., Zhang, Y., and Wang, D. (2022). Unsupervised domain adaptation for covid-19 information service with contrastive adversarial domain mixup. In *2022 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 159–162.

Zhang, H., Fang, Q., Qian, S., and Xu, C. (2019). Multi-modal knowledge-aware event memory network for social media rumor detection. In *Proceedings of the 27th ACM International Conference on Multimedia*, MM '19, page 1942–1951, New York, NY, USA. Association for Computing Machinery.

Zhang, H., Qian, S., Fang, Q., and Xu, C. (2021). Multimodal disentangled domain adaption for social media event rumor detection. *IEEE Transactions on Multimedia*, 23:4441–4454.

Zhang, T., Wang, D., Chen, H., Zeng, Z., Guo, W., Miao, C., and Cui, L. (2020). Bdann: Bert-based domain adaptation neural network for multi-modal fake news detection. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

Zhou, X. and Zafarani, R. (2020). A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys*, 53.

Zhu, Y., Sheng, Q., Cao, J., Nan, Q., Shu, K., Wu, M., Wang, J., and Zhuang, F. (2023). Memory-guided multi-view multi-domain fake news detection. *IEEE Transactions on Knowledge and Data Engineering*, 35(7):7178–7191.