25°enancib
Encontro Nacional de Pesquisa e Pós-graduação em Ciência da Informação
Informação, Decolonialidade e Direitos Difusos para o Desenvolvimento Sustentável
3 a 7 de novembro de 2025 - Rio de Janeiro

XXV ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO - XXV ENANCIB

GT 8 – Dados, Informação e Tecnologia

DETECÇÃO DE NOTÍCIAS FALSAS EM PORTUGUÊS: UMA REVISÃO SISTEMÁTICA DA LITERATURA COM ÊNFASE EM DEEP LEARNING E INTELIGÊNCIA ARTIFICIAL EXPLICÁVEL

FAKE NEWS DETECTION IN PORTUGUESE: *A SYSTEMATIC LITERATURE REVIEW WITH EMPHASIS ON DEEP LEARNING AND EXPLAINABLE ARTIFICIAL INTELLIGENCE*

Daniela Souza Moreira da Silva – Universidade Federal de Santa Catarina (UFSC)
Angel Freddy Godoy Vieira – Universidade Federal de Santa Catarina (UFSC)

Modalidade: Trabalho Completo

Resumo: as notícias falsas, tradução literal de fake news, correspondem a conteúdos que são criados como o objetivo de desinformar e/ou manipular a opinião pública. Independente do contexto em que ocorram: político, econômico, científico, na saúde, elas podem gerar impactos negativos na sociedade, sobretudo diante do crescente número de usuários conectados. Assim, o objetivo desta pesquisa foi identificar, por meio de uma revisão sistemática da literatura (RSL), como é realizado e quais técnicas, são empregadas nos trabalhos que realizam a análise de notícias textuais para classificá-las, e se há alguma ferramenta de IAE para auxiliar o usuário na compreensão do resultado. A RSL foi conduzida de acordo com o protocolo estabelecido o que resultou na seleção de 56 estudos, cuja maior parte envolve modelos de *deep learning* (DL) voltados para o idioma inglês. Nos modelos de DL há uma diversidade de abordagens híbridas que utilizam mais de uma técnica em um sistema de classificação de *fake news*, porém, entre os modelos mais avaliados os que apresentam melhores resultados são as abordagens que utilizam modelos transformadores (BERT ou alguma de suas variações com arquitetura CNN e LSTM). Ao utilizar modelos de DL cuja tomada de decisão deste tipo de modelo costuma ser uma caixa preta na visão do usuário surge a oportunidade de realizar novas pesquisas que utilizem outros recursos, como a inteligência artificial explicável.

Palavras-chave: notícias falsas; aprendizado profundo; inteligência artificial explicável.

Abstract: fake news, a literal translation of fake news, corresponds to content that is created with the aim of misinforming and/or manipulating public opinion. Regardless of the context in which it occurs: political, economic, scientific, or health, it can generate negative impacts on society, especially given the growing number of online users. Thus, the objective of this research was to identify, through a systematic literature review (SLR), how it is carried out and which techniques are used in studies that analyze textual news to classify them, and whether there is any IAE tool to help the user understand the result. The SLR was conducted according to the established protocol, which resulted in the selection of 56 studies, most of which involve deep learning (DL) models focused on the English language. In DL models, there is a diversity of hybrid approaches that use more than one technique in a fake news classification system. However, among the most evaluated models, those that present the best results are the approaches that use transformer models (BERT or some of its variations with CNN and LSTM architecture. When using DL models, whose decision-making of this type of model is usually a black box in the user's view, the opportunity arises to carry out new research that uses other resources, such as explainable artificial intelligence.

Keywords: fake news; deep learning; explainable artificial intelligence.

1 INTRODUÇÃO

Em 2025, observa-se maior acesso à informação, bem como diferentes formas de produzir e compartilhar conteúdo com as pessoas. Por outro lado, há um grande desafio em relação à informação: selecioná-la, analisá-la e transformá-la, de fato, em conhecimento (Mazzeto; Souza, 2022).

Para Ançanello, Casarin e Furnival (2023) o avanço das Tecnologias Digitais de Informação e Comunicação (TDIC), aliado à crescente utilização de mídias sociais e aplicativos de mensagens instantâneas, revela uma nova dinâmica de consumo informacional marcada pelo compartilhamento imediato e pela ampla disseminação de conteúdo. Essa realidade, no entanto, pode favorecer a circulação de informações enviesadas ou até mesmo falsas.

Fake News, ou notícias falsas (tradução literal do termo), são informações noticiosas que têm como objetivo alertar o público acerca de uma determinada situação ou apresentar um ponto de vista específico sobre um acontecimento, porém com conteúdo que não é verdadeiro (Paula; Silva; Blanco, 2018). Para Polonini (2023) *fake news* são conteúdos falsos produzidos com apresentação semelhante a notícias e amplamente divulgados com o intuito de causar danos, podendo ser consideradas um fenômeno de distúrbio informacional que, à medida que são consumidas, podem provocar conflitos nas relações humanas. Além disso, por serem resultantes do contexto social, político, econômico e tecnológico, podem ser consideradas como uma nova era jornalística ou, até mesmo, um novo paradigma jornalístico.

No âmbito da detecção e classificação de *fake News* observa-se uma predominância de estudos direcionados ao idioma inglês, cenário que também se reflete na disponibilidade de conjunto de dados (*datasets*) rotulados em português (Garcia, 2023). Para contornar essa situação, diferentes autores têm contribuído para a construção de *datasets* em português: como Monteiro *et al*, 2018; Garcia; Afonso; Papa, 2022; e Chavarro *et al.* (2023).

Ferreira *et al*. (2022) enfatizam a importância de abordar a limitação dos conjuntos de dados disponíveis na língua portuguesa, bem como a necessidade de implementação e compartilhamento de novos *corpora* voltados à detecção de *fake news*.

A inteligência artificial (IA) tem sido amplamente aplicada em diversos setores, nos quais algoritmos realizam processos decisórios de forma autônoma. Essas decisões podem

gerar impactos positivos ou negativos na vida das pessoas, a depender dos vieses presentes nos modelos utilizados.

A Inteligência Artificial Explicável (IAE), segundo Alves e Andrade (2022), busca esclarecer o processo de predição de modelos algorítmicos. Diante disso, definiu-se a seguinte questão de pesquisa: "como desenvolver um classificador de notícias falsas em português utilizando técnicas de *deep learning* que seja capaz de explicar as suas decisões de forma compreensível para os usuários?".

Sendo assim, o objetivo deste trabalho foi, por meio de uma revisão sistemática da literatura, identificar as técnicas empregadas na análise de notícias textuais para classificação e verificar a existência de ferramentas de IAE que auxiliem os usuários na compreensão dos resultados.

2 DESENVOLVIMENTO

Nesta seção será apresentado o referencial teórico do trabalho, seguido dos procedimentos metodológicos definidos para a pesquisa.

2.1 Fundamentação Teórica

Deep Learning (DL), ou aprendizagem profunda (AP) refere-se ao aprendizado profundo onde são utilizadas várias camadas de processamento onde os parâmetros das redes (os pesos) são modificados visando minimizar as perdas na etapa de treinamento do modelo (Krestel *et al.*, 2021).

Os modelos de DL são considerados "caixas-pretas" mediante a sua complexidade de implementação e dificuldade de interpretação. Em função destas características torna-se importante ter alguma forma de explicar o resultado(saída) deste tipo de modelo.

A definição de um modelo "caixa-preta", de acordo com Guidotti *et al* (2018), é de um sistema que foi desenvolvido ocultando a lógica interna do usuário, sendo construído por meio de uma diversidade de dados aplicadas em soluções de aprendizagem de máquina. Por este motivo o problema da barreira da explicabilidade é inerente às técnicas de DL, e que não estão presentes na IA de sistemas especialistas e dos modeles baseados em regras (Arrieta *et al*, 2020). E a Inteligência Artificial Explicável (IAE) tem como objetivo fornecer informações que ajudem a explicar o processo de predição de um modelo algorítmico (Alves; Andrade, 2022).

2.2 Procedimentos Metodológicos

Para realizar a RSL, foram seguidas as diretrizes propostas por Kitchenham e Charters (2007), abrangendo as principais etapas: planejamento, execução e relato da pesquisa. Por meio da RSL, é possível identificar estudos relevantes e/ou correlacionados com uma determinada questão de pesquisa, além de sua estrutura bem estabelecida contribuir para obtenção de resultados mais confiáveis, com possibilidade de reprodução e evolução em pesquisas futuras.

O protocolo da RSL contém a motivação e a pergunta de pesquisa, juntamente com as questões que se pretendem responder, seguido da estratégia de busca, na qual são especificadas as palavras-chaves e seus sinônimos para definir a *string* genérica de busca, as bases científicas a serem consultadas, bem como os critérios de inclusão e exclusão dos trabalhos, além da estratégia de extração dos dados dos trabalhos selecionados.

A pesquisa foi realizada por meio do Portal de Periódicos da Capes, consultando seis bases científicas, são elas: ACM DL, IEEE Xplore, ScienseDirect, Scopus, SpringerLink e Web Of Sciense. Para isso, utilizou-se a *string* genérica de busca definida a partir das palavras-chaves e seus sinônimos. A consulta às bases retornou 1.779 publicações, das quais foram selecionados os trabalhos publicados nos últimos cinco anos, entre 2019 e 2023, resultando em 1.754 trabalhos. Estes foram triados de acordo com os critérios de inclusão de exclusão estabelecidos, visando identificar pesquisas primárias que utilizassem *deep learning* nas tarefas de detecção e classificação de *fake News*, de modo a possibilitar o mapeamento das técnicas são utilizadas.

Na próxima seção será apresentada a metodologia utilizada para realizar a revisão sistemática, destacando cada uma de suas etapas.

3 METODOLOGIA

Foi realizada uma pesquisa exploratória com diferentes termos de busca até chegar a um conjunto de termos, quando combinados, apresentaram resultados considerados como satisfatórios para iniciar a execução do protocolo, retornando trabalhos relacionados ao tema de pesquisa.

A elaboração deste protocolo foi realizada com base no trabalho de Takaki e Dutra (2023) e adaptada de acordo com os objetivos desta pesquisa.

A questão de pesquisa foi estabelecida com o objetivo de identificar como é conduzida a análise de notícias textuais, quais técnicas e metodologias são utilizadas para sua classificação e se há ferramentas de IAE capazes de auxiliar o usuário na compreensão dos resultados. A população corresponde ao grupo afetado pela intervenção, enquanto a intervenção representa o foco da investigação. A comparação refere-se aos elementos ou métodos utilizados como referência em relação à intervenção, e os resultados consistem nos dados observados a partir dessa intervenção, considerando o contexto em que a comparação ocorre. O Quadro 1 apresenta a questão de pesquisa definida, bem como a população, a intervenção, a comparação, os resultados e contexto estabelecidos para esta RSL.

Quadro 1 – Questão de pesquisa

Elemento	**Descrição**
Questão de Pesquisa	Quais técnicas de DL e IAE são utilizadas nos sistemas de classificações de notícias falsas textuais?
População	Notícias falsas textuais em português e inglês
Intervenção	Técnicas de aprendizado profundo aplicadas à classificação de notícias falsas e ferramentas de IAE.
Comparação	Comparação com modelos de aprendizagem de máquina sem DL
Resultados	Identificação de quais técnicas de DL apresentam melhor desempenho e se incorporam recursos de IAE.
Contexto	Classificação automática de notícias textuais

Fonte: elaborado pelos autores (2025).

Para viabilizar o mapeamento das técnicas identificadas nos estudos selecionados, foram definidas oito questões de pesquisa a serem respondidas em cada trabalho analisado, a saber: (Q1) quais *datasets* são utilizados; (Q2) quais modelos de DL são empregados; (Q3) quais etapas compõem o pré-processamento textual; (Q4) quais técnicas são aplicadas para a extração de informações; (Q5) quais métricas são adotadas para a avaliação dos modelos; (Q6) quais bibliotecas e linguagens de programação são utilizadas; (Q7) quais limitações são reportadas nos trabalhos; e (Q8) quais ferramentas de IAE são empregadas.

A estratégia de busca contempla estabelecer os termos de busca e o locais de consulta para que os trabalhos retornados, e posteriormente selecionados, sejam capazes de responder à pergunta de pesquisa. O Quadro 2 apresenta as fontes de pesquisa (repositórios das áreas da Ciência da Computação e Ciência da Informação) que foram consultadas juntamente com os termos de busca que irão compor as *strings* de cada base.

Quadro 2 – Fontes de Pesquisa (Bases) e Termos de Busca

Bases	**Termos em inglês**
ACM Digital Library ***IEEExplore*** ***Science Direct*** ***Scopus*** ***Springer Link*** ***Web of Science***	*Fake News* *Disinformation* *Fake News Classification* *Fake News Detection* *Fake News classification method* *Fake News detection method* *Deep Learning* *Natural Language Processing* *Explainable Artificial Intelligence*

Fonte: elaborado pelos autores (2025).

Os termos de busca foram definidos considerando os sinônimos identificados em publicações correlacionadas ao tema da pesquisa, juntamente com as palavras-chaves estabelecidas para RSL: *fake News, deep learning e Natural Language Processing*. Ressalta-se que os termos relacionados à IAE não foram incluídos na *string* genérica, pois restringiam os resultados. Contudo, nos trabalhos retornados foi verificado se havia a utilização de alguma técnica de IAE. Desta forma, optou-se por empregar com uma *string* mais ampla, de modo a obter um número maior de publicações, conforme apresentado no Quadro 3:

Quadro 3 – *String* genérica

("fake news" OR "disinformation") AND ("fake news classification" OR "fake news detection" OR "fake news classification method" OR "fake news detection method") AND "deep learning" AND "natural language processing"

Fonte: elaborado pelos autores (2025).

Quanto aos critérios de inclusão e exclusão, apresentados no Quadro 4, é importante justificar a escolha do período de 2019 a 2023. Essa decisão baseia-se no aumento de publicações relacionadas ao tema *fake news* após as eleições dos EUA em 2016, na infodemia decorrente da COVID-19 e na evolução significativa da capacidade computacional envolvendo IA neste período. Já os critérios de exclusão foram estabelecidos para desconsiderar as pesquisas teóricas (revisões da literatura) e estudos cujo acesso integral ao conteúdo não estivesse disponível por meio do Portal de Periódicos da CAPES.

Quadro 4 – Critérios de inclusão e de exclusão

Tipo	Critério	Descrição
Inclusão	**CI0**	Pesquisas primárias publicadas entre 2019 e 2023 que utilizem *deep learning* para classificação de *fake news* e que citavam a utilização de IAE na pesquisa
Exclusão	**CE0**	Estudos teóricos (pesquisas secundárias e terciárias), trabalhos conceituais, editoriais e prefácios, capítulo de livro
Exclusão	**CE1**	Pesquisas em andamento (estudos incompletos e/ou com resultados parciais)
Exclusão	**CE2**	Trabalhos aplicados para *datasets* de línguas diferentes do inglês e português.
Exclusão	**CE3**	Trabalhos sem acesso ao texto completo da publicação;
Exclusão	**CE4**	Estudos sem metodologia clara de uso de PLN e DL
Exclusão	**CE5**	Trabalhos que não apliquem PLN sobre corpora textuais
Exclusão	**CE6**	Pesquisas que analisam imagens (multimodal)
Exclusão	**CE7**	Pesquisas voltadas exclusivamente para redes sociais
Exclusão	**CE8**	Pesquisas não voltadas para classificação/detecção de fake News e IAE

Fonte: elaborado pelos autores (2025).

Após a definição dos critérios de inclusão e exclusão, foi definido o procedimento de seleção dos estudos, triagem em duas etapas, que corresponde a tarefa de registrar em uma planilha a relação dos trabalhos que retornaram das consultas de cada base, com as seguintes informações: base, título do trabalho, autores, ano de publicação, resumo e DOI, sendo selecionados os trabalhos dentro do período estabelecido (entre 2019 e 2023), excluindo as publicações referentes a capítulos de livros, revisões da literatura, anais de eventos, bem como, os trabalhos duplicados (primeira etapa).

Na segunda etapa, após a consolidação de todos os trabalhos planilhados, foram aplicados os critérios de inclusão e exclusão analisando o título, o resumo e as palavras-chaves. Em caso de dúvidas para excluir/incluir o trabalho por informações consideradas insuficientes, ele era mantido para a próxima etapa.

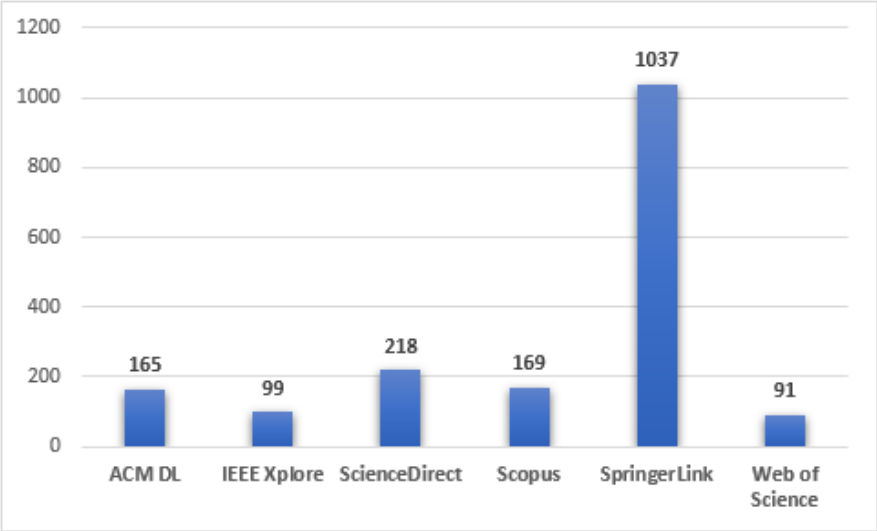
Após a análise da segunda etapa os trabalhos foram lidos na íntegra, para a aplicação da última triagem dos critérios de exclusão (caso fosse necessário). Em seguida, foi realizada a extração dos dados das publicações selecionadas de acordo com a estratégia definida.

A seguir, serão exibidos os resultados obtidos.

4 RESULTADOS

As consultas foram realizadas em setembro de 2023, de forma manual, nas seis bases de publicações definidas, utilizando a *string* genérica no campo de pesquisa de cada portal. Os resultados retornaram um total de 1.779 trabalhos científicos. A Figura 1 apresenta o resultado retornado em cada base acessada.

Figura 1 – Publicações científicas retornadas por base



Fonte: elaborada pelos autores (2025).

A partir dos resultados retornados apresentados na Figura 1 foi aplicado o filtro referente ao período temporal dos últimos 5 anos, entre 2019 e 2023. Houve uma redução de 1,41% resultando em 1.754 publicações. Em seguida, foi aplicado o filtro de tipo de publicação obtendo-se uma redução de 62,26% e resultando em 662 trabalhos. Sendo que filtro do tipo de publicação era para selecionar artigos de pesquisa. Ao remover as publicações duplicadas resultaram em 586 trabalhos. A Tabela 1 apresenta o detalhamento por base.

Tabela 1 – Filtros aplicados no processo inicial de triagem

Base	Publicações retornadas	Filtro Temporal (2019-2023)	Filtro por tipo de publicação	Exclusão duplicados
ACM	**165**	**161**	**103**	**102**
IEEE Xplore	**99**	**95**	**95**	**86**
ScienseDirect	**218**	**212**	**183**	**171**
Scopus	**169**	**166**	**62**	**15**
SpringerLink	**1.037**	**1.301**	**163**	**156**
Web of Science	**91**	**89**	**56**	**56**
Total	**1.779**	**1.754**	**662**	**586**

Fonte: elaborada pelos autores (2025).

Em seguida, o processo de triagem foi realizado aplicando os critérios de inclusão (CI) e critérios de exclusão (CE), já apresentados no Quadro 4, e o resultado desta etapa está apresentado na Tabela 2.

Tabela 2 – Filtros aplicados com critérios de inclusão e exclusão

Base	Título	Resumo	Acesso	Leitura Trabalho Completo
ACM	**25**	**10**	**1**	**0**
IEEE Xplore	**40**	**22**	**22**	**18**
ScienseDirect	**37**	**22**	**22**	**16**
Scopus	**4**	**2**	**2**	**2**
SpringerLink	**32**	**14**	**14**	**9**
Web of Science	**27**	**14**	**11**	**11**
Total	**165**	**84**	**72**	**56**

Fonte: elaborado pelos autores (2025).

Os critérios foram aplicados no título das publicações, resumo (*abstract*), no acesso total para visualizar o conteúdo completo do trabalho e por fim, após a leitura do trabalho completo. A partir do conjunto de 586 trabalhos selecionados após a triagem inicial dos filtros foram aplicados os critérios nos títulos resultando em 165 publicações para serem analisadas na etapa seguinte. A Tabela 3 apresenta o detalhamento por critério de CE e CI0 indica a quantidade de trabalhos a ser avaliado na próxima etapa. Neste contexto, observou-se que o CE7 foi critério responsável pelo maior número de publicações removidas. Este critério está relacionado com pesquisas voltadas exclusivamente para análise de dados das redes sociais.

Tabela 3 – Critérios aplicados nos títulos das publicações

Critério	Nº Publicações	%
CE0	**73**	**12,46**
CE1	**12**	**2,05**
CE2	**56**	**9,56**
CE5	**16**	**2,73**
CE6	**52**	**8,87**
CE7	**144**	**24,57**
CE8	**68**	**11,60**
CI0	**165**	**28,16**
Total	**586**	**100%**

Fonte: elaborada pelos autores (2025).

Logo após, foram lidos os resumos dos 165 trabalhos selecionados e ao aplicar os critérios CE resultou em 84 trabalhos CI0) conforme detalhado na Tabela 4. Novamente o CE7 foi o critério responsável pelo maior número de exclusões. Após a leitura completa dos trabalhos, foram selecionados 56 trabalhos para a etapa de extração de dados.

Tabela 4 – Critérios aplicados no resumo das publicações

Critério	Nº Publicações	%
CE0	9	5,45
CE1	1	0,61
CE2	14	8,48
CE4	2	1,21
CE6	4	2,42
CE7	21	12,73
CE8	30	18,18
CI0	84	50,91
Total	165	100%

Fonte: elaborada pelos autores (2025).

A partir das 56 publicações selecionadas foi iniciado o processo para extrair os dados visando responder as perguntas estabelecidas para RSL, cujo foco estava em identificar quais técnicas de DL que são utilizadas na etapa de PLN dos detectores e classificadores de *fake news*, quais *datasets* e o idioma destes corpora textuais são utilizados, quais etapas são utilizadas no pré-processamento textual, bem como, as técnicas utilizadas para extração das informações. Além dessas informações, quais bibliotecas e linguagens de programação são utilizadas na etapa de implementação dos modelos, assim como, as limitações das pesquisas e as ferramentas de IAE que são utilizadas.

Com relação a pergunta “Q1: Quais *datasets* são utilizados?” verificou-se que a maioria das publicações, utilizava entre 2 ou 3 *datasets* para as etapas de treinamento e teste de seus classificadores, totalizando 41 trabalhos entre os 56, o que corresponde a 51,22%.

Há uma grande diversidade entre os datasets utilizados, tanto na quantidade de registros que cada conjunto possui quanto a origem de seus dados. Foram identificados 61 datasets diferentes citados nas publicações, sendo que, em dois trabalhos os autores citam

que foi realizada a coletada dos dados em sites de notícias, porém não disponibilizam o *dataset* para consulta. Destes 61 *datasets*, 60 são em inglês e apenas 1 em espanhol.

Para responder à questão "Q2 - Quais modelos de DL são utilizados?" foi verificado em cada trabalho quais técnicas de *deep learning* eram utilizadas, independente da etapa onde era implementada, seja no classificador ou na etapa de processamento de linguagem natural. Ao consolidar os dados observou-se que entre os 56 trabalhos há 46 soluções, aproximadamente 82%, utilizam arquiteturas mistas implementadas para realizar a tarefa de classificação. Observou-se que o BERT, e suas variações, é utilizado de alguma forma em 21 trabalhos e as arquiteturas LSTM e CNN também foram utilizadas em 39 trabalhos. As configurações de parâmetros destas redes variam em cada trabalho sendo necessário realizar diversos experimentos para definir a melhor arquitetura e sua configuração.

Quanto à questão: "Q3 - Quais etapas fazem parte da etapa de pré-processamento textual?" ao consolidar os dados extraídos verificou-se que em 23 trabalhos (41%) não havia uma descrição explícita da normalização textual realizada na etapa de pré-processamento. No entanto, para os demais 33 trabalhos (59%) os autores descreviam a etapa. A fase de pré-processamento pode ser formada por diferentes combinações entre as etapas mais comuns, dentre elas estão: uniformização (padronização maiúscula ou minúscula), tokenização, uso de expressão regulares, remoção de *stopwords*, remoção de pontuação, lematização e radicalização (*stemming*), e correção de grafia.

Já na questão: "Q4 - Quais técnicas são utilizadas para extrair informação?" observou-se que há uma diversidade na estratégia de extrair informação, bem como, na técnica de *word embeddings* utilizado. Contudo, há uma predominância no uso do BERT e GLOVE, sendo o BERT o modelo mais utilizado e com melhores resultados. Na etapa de *feature extraction* o TF-IDF e BoW foram os mais utilizados.

Com relação à pergunta: "Q5 - Quais métricas são utilizadas para avaliar os modelos?" observou-se que foram utilizados 20 conjuntos de métricas diferentes na avaliação dos 56 trabalhos. No entanto, foi observado que a métrica acurácia é utilizada em 16 conjuntos diferentes e que o conjunto das métricas Acurácia; Precisão; *Recall*; F1-*Score*, foi utilizado em 42,86% dos trabalhos.

Quanto a pergunta "Q6 - Quais as bibliotecas e linguagens de programação são utilizadas?" verificou-se que 38 trabalhos informam que foi utilizada a linguagem de programação Python, o que representa aproximadamente 68% do total de publicações

analisadas. O restante, 18 trabalhos (32%) não informam qual a linguagem foi utilizada na implementação dos modelos. Com relação às bibliotecas utilizadas em 22 trabalhos não há um detalhamento de quais são elas (39%) nos demais 34 trabalhos (61%) são informados pelo menos uma biblioteca utilizada em alguma etapa da pesquisa. A Tabela 14 apresenta a relação dessas bibliotecas mapeadas. Observa-se que para este conjunto de trabalhos a biblioteca NLTK é citada em 16 trabalhos e Spacy em 4 trabalhos como as ferramentas para a etapa de NLP. As bibliotecas Keras, Tensorflow e Pytorch utilizadas para implementas das redes neurais e DL, foram citadas em 20 trabalhos.

Nos trabalhos avaliados, em 32 deles (57,14%) não há informações sobre as configurações de hardware ou dos ambientes de desenvolvimento onde os experimentos foram realizados. Porém, em 8 trabalhos (14,29%) os autores informam que utilizaram o Google Colab, em 2 trabalhos (3,57%) foi utilizado Jupyter e em um trabalho (1,79%) foram utilizados o Google Colab e o Jupyter. Nos outros 13 trabalhos constam algumas informações relacionadas a configuração de hardware de processador e memória.

Com relação à "Q7 - Quais as limitações reportadas nos trabalhos?" os dados coletados são textuais e podem variar de acordo com o domínio de cada pesquisa, o que dificulta organizar em categorias para mensurar de forma quantitativa. No entanto, foi possível observar que em linhas gerais, na maioria dos trabalhos (em pelo menos 32 deles) há as seguintes limitações em suas pesquisas:

i) Ausência de uma análise detalhada dos recursos computacionais necessários para implementar o sistema, como uso de memória ou tempo de processamento.

ii) A generalização da abordagem proposta para diferentes tipos de notícias falsas ou diferentes idiomas não é abordada.

iii) Não fornece nenhuma visão sobre a interpretabilidade do modelo ou como ele identifica e pondera diferentes características nos artigos de notícias.

iv) Os modelos foram treinados e testados em artigos no idioma inglês, não sendo claro quão bem eles generalizariam para artigos de outros idiomas.

v) Ausência de uma análise detalhada das limitações da abordagem proposta ou potenciais desafios computacionais na sua implementação.

vi) Não apresenta uma discussão referente aos possíveis vieses ou limitações associados aos dados de treinamento usados para os experimentos.

Por fim, quanto à questão " Q8 – Quais ferramentas de IAE são utilizadas?", verificou-se que entre todos os trabalhos selecionados em apenas 2 deles foi proposto um modelo com foco na interpretação ou explicabilidade do que foi implementado. No trabalho de Gadek e Guélorget (2020) os autores utilizam arquitetura CNN com mapas de ativação de classes (CAMs) para propor uma previsão e interpretação precisa do resultado. Já Magistris *et al* (2022) mencionam que a classificação das notícias é explicada por meio do uso de técnicas de reconhecimento de entidades nomeadas (NER) e classificação de postura. No entanto, não fica claro quais bibliotecas de IAE foram utilizadas.

5 CONSIDERAÇÕES FINAIS

Esta pesquisa indica resultados relevantes com relação a detecção de notícias falsas, principalmente com relação ao português, tendo em vista que os trabalhos avaliados eram voltados para o idioma inglês o que indica a necessidade de mais pesquisas voltadas para o português.

Quanto aos modelos de DL há uma diversidade de abordagens mistas, isto é, abordagens híbridas que utilizam mais de uma técnica em um sistema de detecção e classificação de *fake news*, porém, entre os modelos mais avaliados os que apresentam melhores resultados são as abordagens que utilizam modelos transformadores (BERT ou alguma de suas variações com arquitetura CNN e LSTM. Reforçando ainda, que os modelos de DL apresentam resultados superiores as abordagens que utilizam apenas os algoritmos de machine learning, e nas etapas de processamento de linguagem natural os algoritmos de DL tem auxiliado na melhoria dos resultados.

Por outro lado, o fato de utilizar modelos de *deep learning* cuja tomada de decisão deste tipo de modelo costuma ser caixa preta surge a oportunidade de realizadas novas pesquisas que utilizem outros recursos, como a inteligência artificial explicável, para que a solução implementada faça uso de alguma técnica que permita explicar a tomada de decisão do modelo de tal forma que os resultados sejam mais confiáveis.

Além disso, foi possível observar que os trabalhos não exploram as questões relacionadas à implementação, desde os recursos computacionais necessários (configuração de hardware), bem como, as dificuldades técnicas relacionadas à implementação dos modelos de DL. Nas avaliações não foram realizados testes considerando às percepções dos usuários quanto ao resultado da classificação das notícias, de modo geral, os trabalhos realizaram

avaliações por meio de métricas. Já com relação às bibliotecas, a revisão apresenta uma diversidade de opções que podem ser utilizadas para diferentes etapas de um sistema de detecção e classificação de *fake news*. Foram identificadas ferramentas para etapas de PLN como a NLK, Spacy, e textblob e Keras, Tensorflow e Pytorch para implementação referente aos modelos de DL. A linguagem Python já era esperada ser a mais utilizada nos trabalhos da área, tendo em vista é uma linguagem cuja curva de aprendizado é suave, com uma grande diversidade de bibliotecas disponíveis e em desenvolvimento pela comunidade. É uma linguagem de alto nível utilizada com frequência em diferentes áreas, entre elas ciência de dados e IA.

Quanto as limitações relacionadas à RSL realizada é necessário destacar que o filtro aplicado nos últimos anos, compreendidos entre 2019 e 2023, podem ter deixado de incluir trabalhos relevantes para o contexto da pesquisa, bem como, a definição dos critérios de inclusão e exclusão também podem afetar a relação de trabalhos selecionados para análise.

REFERÊNCIAS

ALVES, Marco Antônio Sousa; ANDRADE, Otávio Morato de. Da “caixa-preta” à “caixa de vidro”: o uso da explainable artificial intelligence (xai) para reduzir a opacidade e enfrentar o enviesamento em modelos algorítmicos. **Direito Público**, [*s. l.*], v. 18, n. 100, p. 349-373, 27 jan. 2022. Disponível em: https://www.portaldeperiodicos.idp.edu.br/direitopublico/article/view/5973. Acesso em: 16 ago. 2025.

ANÇANELLO, Juliana Venancio; CASARIN, Helen de Castro Silva; FURNIVAL, Ariadne Chloe. Competência em informação, fake news e desinformação: análise das pesquisas no contexto brasileiro. **Em Questão**, v. 29, p. 125782, 2023. Disponível em: https://seer.ufrgs.br/index.php/EmQuestao/article/view/125782. Acesso em: Acesso em: 16 ago. 2025.

ARRIETA, Alejandro Barredo; DÍAZ-RODRÍGUEZ, Natalia; SER, Javier del; BENNETOT, Adrien; TABIK, Siham; BARBADO, Alberto; GARCIA, Salvador; GIL-LOPEZ, Sergio; MOLINA, Daniel; BENJAMINS, Richard; CHATILA, Raja; HERRERA, Francisco. Explainable Artificial Intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible ai. **Information Fusion**, [*s. l.*], v. 58, p. 82-115, jun. 2020. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S1566253519308103. Acesso em: 15 ago. 2025.

CHAVARRO, Juan Pablo; CARVALHO, Jonata Tyska; PORTELA, Tarlis Tortelli; SILVA, Jonathan Cardoso. FakeTrueBR: um corpus brasileiro de notícias falsas. *In*: ESCOLA REGIONAL DE BANCO DE DADOS (ERBD 2023), 18., 2023, Porto Alegre. **Anais** [...]. Porto Alegre: Sociedade

Brasileira de Computação – SBC, 2023. p. 108-117. Disponível em: https://sol.sbc.org.br/index.php/erbd/article/view/24352. Acesso em: 16 ago. 2025.

FERREIRA, Fernanda Vasques; VARÃO, Rafiza; BOSELLI, Marco Aurélio; SANTOS, Leandro Brito; MORET, Marcelo A. Uso de Python para detecção de fake news sobre a covid-19: desafios e possibilidades. **RECIIS**, Rio de Janeiro, v. 16, n. 2, p. 266-280, abr./jun. 2022. Disponível em: https://doi.org/10.29397/reciis.v16i2.3253. Acesso em: 16 ago. 2025.

GADEK, Guillaume; GUÉLORGET, Paul. An interpretable model to measure fakeness and emotion in news. **Procedia Computer Science**, [*s. l.*], v. 176, p. 78-87, 2020. Disponível em: https://www-sciencedirect.ez46.periodicos.capes.gov.br/science/article/pii/S1877050920318330. Acesso em: 16 ago. 2025.

GARCIA, Gabriel Lino; AFONSO, Luis C. S.; PAPA, João P. FakeRecogna: a new brazilian corpus for fake news detection. **Lecture Notes In Computer Science**, Fortaleza, p. 57-67, mar. 2022. Disponível em: https://link.springer.com/chapter/10.1007/978-3-030-98305-5_6. Acesso em: 16 ago. 2025.

GARCIA, Gabriel Lino. **Detecção de Fake News utilizando aprendizado de máquina**. 2023. 79f. Dissertação (Mestrado em Ciência da Computação) - Programa de Pós-Graduação em Ciência da Computação, Universidade Estadual Paulista Júlio de Mesquita Filho, Bauru, 2023. Disponível em: https://repositorio.unesp.br/handle/11449/242304. Acesso em: 16 ago. 2025

GUIDOTTI, Riccardo; MONREALE, Anna; RUGGIERI, Salvatore; TURINI, Franco; GIANNOTTI, Fosca; PEDRESCHI, Dino. A Survey of Methods for Explaining Black Box Models. **Acm Computing Surveys**, [*s. l.*], v. 51, n. 5, p. 1-42, 22 ago. 2018. Disponível em: https://dl.acm.org/doi/10.1145/3236009. Acesso em: 16 ago. 2025.

KITCHENHAM, Barbara; CHARTERS, S. Guidelines for performing Systematic Literature Reviews in Software Engineering. **Technical Report (EBSE 2007).** Durham, UK, jul. 2007. 44 p. Disponível em: https://cdn.elsevier.com/promis_misc/525444systematicreviewsguide.pdf. Acesso em: 16 ago. 2025.

KRESTEL, Ralf; CHIKKAMATH, Renukswamy; HEWEL, Christoph; RISCH, Julian. A survey on deep learning for patent analysis. **World Patent Information**, [*s. l.*], v. 65, p. 1-13, jun. 2021. Disponível em: https://www.sciencedirect.com/science/article/pii/S017221902100017X. Acesso em: 16 ago. 2025.

MAGISTRIS, Giorgio de; RUSSO, Samuele; ROMA, Paolo; STARCZEWSKI, Janusz T.; NAPOLI, Christian. An Explainable Fake News Detector Based on Named Entity Recognition and Stance Classification Applied to COVID-19. **Information**, [*s. l.*], v. 13, n. 3, p. 137, mar. 2022. Disponível em: https://www-webofscience.ez46.periodicos.capes.gov.br/wos/woscc/full-record/WOS:000775045300001. Acesso em: 16 ago. 2025.

MAZZETO, Ana Carla Epitácio; SOUZA, Elisabete. Infodemia e Desinformação no Contexto da Pandemia da Covid-19: Reflexões À Luz da Noção de Competência em

Informação. **PontodeAcesso**, *[s. l.]*, v. 16, n. 2, p. 2–23, 2022. Disponível em: https://periodicos.ufba.br/index.php/revistaici/article/view/49151. Acesso em: 16 ago. 2025.

MONTEIRO, Rafael A.; SANTOS, Roney L. S.; PARDO, Thiago A. S.; ALMEIDA, Tiago A. de; RUIZ, Evandro E. S.; VALE, Oto A. Contributions to the Study of Fake News in Portuguese: new corpus and automatic detection results. **Lecture Notes In Computer Science**, Canela, p. 324-334, set. 2018. Disponível em: https://dl.acm.org/doi/10.1007/978-3-319-99722-3_33. Acesso em: 16 ago. 2025.

PAULA, Lorena Tavares de; SILVA, Thiago dos Reis Soares da; BLANCO, Yuri Augusto. Pós-verdade e fontes de informação: um estudo sobre fake news. **Revista Conhecimento em Ação**, v. 3, n. 1, p. 93-110, 2018. Disponível em: http://hdl.handle.net/20.500.11959/brapci/71135. Acesso em: 16 ago. 2025.

POLONINI, Janaína Fernandes Guimarães. **A construção social da informação**: análise do fact-checking no brasil. 2023. 210 f. Tese (Doutorado em Ciência da Informação) - Universidade Estadual Paulista Júlio de Mesquita Filho, Marília, 2023. Disponível em: https://repositorio.unesp.br/handle/11449/243455. Acesso em: 16 ago. 2025.

TAKAKI, Patrícia; DUTRA, Moisés Lima. Text mining applied to distance higher education: a systematic literature review. **Education And Information Technologies**, [*s. l.*], p. 1-28, out. 2023. Disponível em: https://link.springer.com/article/10.1007/s10639-023-12235-0. Acesso em: 16 ago. 2025.