

INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DO CEARÁ
PRÓ-REITORIA DE PESQUISA, PÓS-GRADUAÇÃO E INOVAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Alice Feitosa Barbosa

**Revisão Sistemática e Análise das Etapas do Processo de Extração de
Conhecimento para otimizar a Classificação de Notícias Falsas em
Português**

Fortaleza – CE

Alice Feitosa Barbosa

Revisão Sistemática e Análise das Etapas do Processo de Extração de Conhecimento para otimizar a Classificação de Notícias Falsas em Português

Dissertação submetida ao Programa de Pós-Graduação em Ciência da Computação do Instituto Federal de Educação, Ciência e Tecnologia do Ceará, como requisito parcial para a Defesa da Dissertação do Mestrado em Ciência da Computação.

Linha de Pesquisa: Inteligência Artificial

Orientador: Prof. Dr. Reinaldo Bezerra Braga

Instituto Federal de Educação, Ciência e Tecnologia do Ceará (IFCE)

Pró-Reitoria de Pesquisa, Pós-Graduação e Inovação (PRPI)

Programa de Pós-Graduação em Ciência da Computação (PPGCC)

Fortaleza – CE



Instituto Federal de Educação, Ciência e Tecnologia do Ceará (IFCE)
Pró-Reitoria de Pesquisa, Pós-Graduação e Inovação (PRPI)
Programa de Pós-Graduação em Ciência da Computação (PPGCC)

Alice Feitosa Barbosa

Esta dissertação foi julgada adequada para obtenção do título de Mestre em Ciência da Computação, sendo aprovada pela Coordenação do Programa de Pós-Graduação em Ciência da Computação do Instituto Federal de Educação, Ciência e Tecnologia do Ceará e pela banca examinadora:

Orientador: Prof. Dr. Reinaldo Bezerra Braga
Instituto Federal de Educação, Ciência e Tecnologia do
Ceará (IFCE)

**Prof. Dr. Francisca Raquel de Vasconcelos
Silveira**
Instituto Federal de Educação, Ciência e Tecnologia do
Ceará (IFCE)

Prof. Dr. Tales Paiva Nogueira
Universidade da Integração Internacional da Lusofonia
Afro-Brasileira (UNILAB)

Fortaleza – CE

Este trabalho é dedicado a todas as pessoas que sempre estiveram ao meu lado, acreditaram no meu potencial e cuidaram de mim em minhas crises de ansiedade. Aqui se encontra o fruto dos nossos esforços, muito obrigada.

Agradecimentos

Agradeço, primeiramente, à Deus por estar ao meu lado, não só durante este trabalho, mas durante toda a minha vida, cuidando e dando força.

Agradeço a minha mãe Cláudia Feitosa por sempre me incentivar nos estudos, se doar para criar oportunidades para nossa educação e apoiar nos nossos sonhos. Obrigada por tudo! E as minhas irmãs Rogéria Feitosa, Tamires Maria Feitosa e Clarisse Maria Feitosa, por sempre me darem forças e me fazerem rir. O apoio de vocês fez e sempre fará a diferença na minha vida.

Agradeço, com muito amor, ao meu noivo Anderson Queiroz, por me ajudar ao longo dessa caminhada, não só no processo desse mestrado, mas na vida, aguentando minhas reclamações, angústias e crises de ansiedade. Obrigada por estar sempre ao meu lado, acreditando em mim quando eu mesma já tinha perdido as esperanças.

Agradeço aos meus avós Luiz Feitosa e Maria do Socorro, por sempre cuidarem de mim, sem pensar duas vezes em me proporcionar viver meus sonhos.

Em especial, agradeço ao meu orientador Prof. Reinaldo Bezerra Braga, por sua confiança e otimismo. Por sempre acreditar no meu potencial e me acompanhar nessa jornada. Como também a Prof. Carina Oliveira, por seu incentivo, conselhos e ajuda na minha carreira acadêmica.

Agradeço imensamente aos meus amigos, Mércia Ferreira, Luan Douglas e Felipe Harlen, por sempre me ouvirem, incentivarem e ajudarem ao longo desse processo, nas trocas de conhecimentos e conselhos. Tenho um carinho muito grande por vocês.

E agradeço a todas as pessoas que de alguma forma colaboraram para que este trabalho fosse realizado.

“Palavras na minha não tão humilde opinião são nossa inesgotável fonte de magia”
(J.K Rowling - Harry Potter e as Relíquias da Morte)

Resumo

Nos últimos anos, a adesão às redes sociais tem apresentado um aumento significativo, o que tem permitido o acesso à informação com apenas um clique, oferecendo um ambiente interativo e mecanismos de ampla divulgação. No entanto, o grande volume de informações compartilhadas e a dificuldade de validar sua veracidade proporcionam um ambiente favorável para a rápida proliferação de notícias falsas, popularmente conhecidas como *Fake News*. Devido ao aumento significativo de *Fake News* nas redes sociais, diversos estudos e estratégias têm sido desenvolvidos nos últimos anos com o intuito de identificá-las. Dentre elas, destaca-se a aplicação de algoritmos de Processamento de Linguagem Natural (PLN), juntamente com algoritmos de Aprendizagem de Máquina (AM) e Aprendizado Profundo (AP) em especial as Redes Neurais Artificiais (RNA). Ou seja, diversos estudos e estratégias têm sido desenvolvidos nos últimos anos com o intuito de identificar notícias falsas. Sendo assim, a presente dissertação apresenta uma avaliação de desempenho das etapas do processo de extração de conhecimento para detecção automática de *Fake News* em Português. Mais especificamente, é realizada uma análise do processo de construção dos algoritmos, investigando o impacto das tomadas de decisão no processo metodológico no resultado final. Para este fim, foi desenvolvido uma metodologia dividida em quatro etapas, sendo estas: I) Premissas de Análise; II) Base de dados; III) Construção dos cenários e IV) Testes e validações. As premissas de análise foi construída por meio de uma revisão sistemática, a primeira contribuição deste trabalho, que está relacionada à identificação da pesquisa científica em torno da detecção de notícias falsas em português. A base de dados selecionada foi a *Fake.Br Corpus*, que apresenta 7.200 artigos de notícias em português. Destaca-se que a pesquisa realizada concentrou-se em analisar tanto os textos, quanto os seus respectivos metadados. Desta forma, após uma análise das combinações, obteve-se uma média da acurácia de 97%.

Palavras-chave: Notícias Falsas. Processamento de Linguagem Natural. Língua Portuguesa.

Abstract

In recent years, adherence to social networks has shown a significant increase, which has allowed access to information with just one click, offering an interactive environment and mechanisms for wide dissemination. However, the large volume of shared information and the difficulty of validating its veracity provide a favorable environment for the rapid proliferation of false news. Due to the significant increase in *Fake News* on social networks, several studies and strategies have been developed in recent years in order to identify them. Among them, the application of Natural Language Processing (NLP) algorithms stands out, along with Machine Learning (ML) and Deep Learning (AP) algorithms, especially Artificial Neural Networks (ANN). That is, several studies and strategies have been developed in recent years in order to identify false news. Therefore, this dissertation presents a proposal for the automatic identification of fake news. More specifically, an analysis of the algorithm construction process is carried out, investigating the impact of decision-making in the methodological process on the final result. For this purpose, a methodology divided into four stages was developed, namely: I) Analysis premises; II) Database; III) Construction of scenarios and IV) Tests and validations. The analysis premises was built through a systematic review, the first contribution of this work, which is related to the identification of scientific research around the detection of fake news in Portuguese. The selected database was the *Fake.Br Corpus*, which presents 7,200 news articles in Portuguese. It is noteworthy that the research carried out focused on analyzing both the texts and their respective metadata. Thus, after analyzing the combinations, an average accuracy of 97% was obtained.

Keywords: Fake News. Natural Language Processing. Portuguese Language.

Lista de ilustrações

Figura 1 – Representação do Word2vec (Fonte: (AMAZON, 2018)).	30
Figura 2 – Representação do CBoW (Fonte: (RONG, 2016)).	31
Figura 3 – Processo de convolução.	33
Figura 4 – Max-Pooling.	34
Figura 5 – Arquitetura do LSTM.	35
Figura 6 – Diagrama PRISMA.	40
Figura 7 – Distribuição anual dos trabalhos.	42
Figura 8 – Fontes de publicação dos trabalhos.	42
Figura 9 – Conteúdos estudados.	44
Figura 10 – Redes sociais adotadas nos trabalhos.	44
Figura 11 – <i>Datasets</i> utilizados nos trabalhos.	45
Figura 12 – Algoritmos aplicados nos trabalhos.	46
Figura 13 – Resultado dos algoritmos aplicado nos trabalhos selecionados.	46
Figura 14 – Sequência de etapas metodológicas.	47
Figura 15 – Resultados da CNN.	54
Figura 16 – Matriz de confusão de teste do algoritmo CNN.	54
Figura 17 – Matrizes de confusão do processamento dos metadados.	56

Lista de tabelas

Tabela 1	– Tabela PICO.	37
Tabela 2	– Sintaxe de busca.	38
Tabela 3	– Trabalhos selecionados.	48
Tabela 4	– Categorias da base <i>Fake.Br Corpus</i>	49
Tabela 5	– Estatísticas descritivas referentes ao número de palavras das notícias. .	55
Tabela 6	– Resultados dos Metadados.	55
Tabela 7	– Resultados do primeiro experimento.	56
Tabela 8	– Resultados do segundo experimento.	57

Sumário

1	INTRODUÇÃO	21
1.1	Objetivo geral e objetivos específicos	23
1.2	Estrutura do trabalho	23
2	FUNDAMENTAÇÃO TEÓRICA	25
2.1	Fake News	25
2.2	Processamento de Linguagem Natural (PLN)	27
2.3	Modelos de Representação Textual	28
2.3.1	Term Frequency-Inverse Document Frequency (TF-IDF)	28
2.3.2	Word2Vec	29
2.3.3	Continuous Bag of Words (CBOW)	30
2.3.4	GloVe	31
2.4	Redes Neurais Convolucionais (CNN)	32
2.5	Redes de Memória de Longo e Curto Prazo	34
3	REVISÃO SISTEMÁTICA	37
3.1	Etapa 1: Planejamento	37
3.1.1	Questões de Pesquisa (QP)	37
3.1.2	Protocolos de busca	38
3.1.3	CrITÉrios de Qualidade	38
3.1.4	CrITÉrios de Inclusão e Exclusão	39
3.2	Etapa 2: Condução	40
3.3	Resultados	41
3.3.1	<i>Como estão distribuídos os trabalhos de detecção automática de fake news ao longo dos anos?</i>	41
3.3.2	<i>Quais os tipos de publicações científicas mais recorrentes no contexto das fake news?</i>	42
3.3.3	<i>Quais países têm mais pesquisadores publicando sobre detecção de fake news?</i>	42
3.3.4	<i>Quais tipos de conteúdo estão sendo considerados como solução na detecção de fake news?</i>	43
3.3.5	<i>Quais as redes sociais mais utilizadas no desenvolvimento dos trabalhos?</i>	43
3.3.6	<i>Quais os datasets mencionados nos trabalhos selecionados?</i>	44
3.3.7	<i>Quais algoritmos de inteligência artificial são aplicados nas pesquisas?</i>	45
3.3.8	<i>Quais estudos apresentam resultados mais qualificados quanto às métricas de avaliação?</i>	46

4	PROPOSTA	47
4.1	Premissas de Análise	47
4.2	Base de dados	49
4.3	Construção dos cenários	50
4.3.1	Pré-processamento	50
4.3.2	Transformação	50
4.3.3	Treinamento	50
4.3.4	Teste	51
5	RESULTADOS	53
5.1	Análise dos Textos	53
5.2	Análise dos metadados	55
5.3	Avaliação de Etapas do Algoritmo	55
6	CONSIDERAÇÕES FINAIS	59
6.1	Trabalhos Futuros	60
	REFERÊNCIAS	61

1 Introdução

O avanço da tecnologia tem trazido novas práticas na vida social, proporcionando meios modernos de comunicação e oferecendo facilidade e celeridade na realização de tarefas cotidianas. Esta realidade representa um fator importante para o aumento do número de usuários ativos que acessam a Internet regularmente no mundo. Segundo uma pesquisa publicada pela (DATAREPORTAL, 2023), atualmente, existem 5,16 bilhões de usuários de Internet no mundo, o que significa que 64,4% da população total do mundo está online.

Em paralelo a este cenário, percebe-se também a evolução das redes sociais. O estudo da (DATAREPORTAL, 2023) revela ainda que, existem 4,76 bilhões de usuários de mídia social em todo o mundo, um pouco menos de 60% da população global total. Destaca-se que, o crescimento de usuários de mídia social abrandou nos últimos tempos, com a adição líquida de 137 milhões de novos usuários neste ano, o que equivale a um crescimento anual de apenas 3%.

Ao mesmo tempo que o avanço das redes sociais tem sido importante para a sociedade, elas também têm impulsionado o consumo de notícias *online*. Em uma pesquisa realizada pelo (REUTERS; OXFORD, 2021) em 12 países, cerca de 66% dos entrevistados declararam fazer uso de duas ou mais redes sociais para consumir, compartilhar ou discutir notícias. Entre as mais citadas estão WhatsApp, Instagram, TikTok, Telegram e Facebook, respectivamente.

Entretanto, a facilidade de acesso, a velocidade com que as informações são compartilhadas, aliadas ao baixo custo, também proporcionam um espaço propício para a proliferação de notícias falsas. Segundo (FULLER; BIROS; WILSON, 2009), com o crescimento maciço da comunicação online, o potencial para enganar as pessoas por meio da Comunicação Mediada por Computador (CMC) também cresceu, e esse engano pode resultar em danos desastrosos de longo alcance em muitas áreas do cotidiano, incluindo os mercados financeiros (CARVALHO; KLAGGE; MOENCH, 2011), (KOGAN; MOSKOWITZ; NIESSNER, 2019) e eventos políticos (ALLCOTT; GENTZKOW, 2017), (ARO, 2016).

Os autores (ALLCOTT; GENTZKOW, 2017) definem *fake news* como artigos de notícias cujos conteúdos são intencional e comprovadamente falsos, sendo capazes de enganar os leitores, muitas vezes para a obtenção de vantagens financeiras, ou por motivações ideológicas, como o favorecimento de certos candidatos a cargos políticos em campanhas eleitorais. De acordo com os autores (KUMAR et al., 2021), a difusão e a acessibilidade da Internet facilitam o acesso às informações para as grandes massas em um

clique. Porém, essa acessibilidade à rede não oferece aos usuários as mesmas facilidades para a verificação da veracidade das informações.

Devido ao risco que a propagação de *fake news* representa à democracia, muitos esforços têm sido empenhados no sentido de desenvolver estratégias para a identificação de notícias falsas e, assim, minimizar os seus danos à sociedade (SHARMA et al., 2019). Dentre as estratégias adotadas, a mais convencional é a checagem de fatos, que consiste na análise humana para verificação de alegações contra evidências baseadas em fatos previamente falados ou escritos. No entanto, à medida que o número de usuários da Internet e a velocidade da disseminação de informações estão crescendo, a automatização do processo de detecção de notícias falsas tem ganhado cada vez mais interesse na comunidade de pesquisa acadêmica (OSHIKAWA; QIAN; WANG, 2018).

Neste sentido, boa parte dos estudos que buscam apresentar soluções a este problema fazem uso das técnicas de Processamento de Linguagem Natural (PLN). Segundo (MEDEIROS; BRAGA, 2020) as técnicas de mineração de texto, combinadas a algoritmos de Aprendizagem de Máquina (AM), vêm apresentando bons resultados. O PLN, segundo (GONZALEZ; LIMA, 2003) trata-se computacionalmente dos diversos aspectos da comunicação humana, como sons, palavras, sentenças e discursos, considerando formatos e referências, estruturas e significados, contextos e usos. De acordo com (RUSSELL; NORVIG, 2010), para que um agente possa adquirir conhecimento, é necessário que este seja capaz de compreender, ainda que parcialmente, o que o autor chama de "confusa e ambígua" linguagem usada pelos humanos.

Tendo em vista os impactos nocivos que as *fake news* podem vir a ocasionar, diversos estudos desenvolveram diferentes abordagens na busca de detectar notícias falsas com alta precisão. Porém, os resultados obtidos por meio da análise das métricas de validação acabam sendo semelhantes. Sendo assim, o presente estudo apresenta uma proposta para identificação automática de *fake news*. Mais especificamente, é realizada uma avaliação de desempenho das etapas do processo de extração de conhecimento para detecção de *fake news* em português, ou seja, é analisado o processo de construção dos algoritmos, investigando o impacto das tomadas de decisão no processo metodológico no resultado final, proporcionando assim, encontrar a melhor combinação na literatura atual para a detecção de notícias falsas escritas em português.

Para alcançar este objetivo foi desenvolvida uma metodologia dividida em quatro etapas, sendo estas: I) Premissas de Análise; II) Base de dados; III) Construção dos cenários e IV) Testes e validações. Destaca-se que, para os testes, foi utilizada a mesma base de dados, intitulada Fake.Br Corpus, que apresenta 7.200 artigos de notícias escritas em português.

A presente dissertação traz como contribuição a análise não apenas do texto e metadados, mas também do processo de construção dos algoritmos, identificando os pontos

que causam maior impacto no resultado final. Frisa-se também que como um diferencial este estudo foca em textos escritos em português, haja vista que grande parte dos estudos nesta área são focados no Inglês.

1.1 Objetivo geral e objetivos específicos

Tendo em vista os impactos nocivos que as *fake news* podem vir a ocasionar, o presente estudo tem como objetivo apresentar uma revisão sistemática e análise das etapas do processo de extração de conhecimento para otimizar a classificação de notícias falsas em português. Ou seja, é analisado o processo de construção dos algoritmos, investigando o impacto das tomadas de decisão no processo metodológico no resultado final, proporcionando assim, encontrar a melhor combinação na literatura atual para a detecção de notícias falsas escritas em português.

Especificamente, o objetivo deste estudo consiste em:

- Apresentar uma revisão sistemática que proporcione identificar o estado da arte, bem como as oportunidades de contribuições que motivaram este trabalho;
- Apresentar uma análise das etapas do processo de extração de conhecimento para otimizar a classificação de notícias falsas em português, que proporcione encontrar os melhores algoritmos para uma determinada necessidade;
- Oferecer insumos capazes de otimizar a aprendizagem do comportamento das soluções de detecção automática de *fake news* a partir de um processo metodológico;

1.2 Estrutura do trabalho

Este trabalho está organizado da seguinte forma:

No Capítulo 2, trata-se da fundamentação teórica, no qual aborda as técnicas estudadas e analisadas. A princípio, é realizada uma contextualização quanto a proliferação de notícias falsas na Internet e suas particularidades. Em seguida, são abordados conceitos referentes às técnicas de processamento textual, modelos de representação textual, redes neurais e aprendizagem profunda.

No Capítulo 3, é descrita a revisão sistemática produzida. Neste capítulo é detalhado a metodologia aplicada para realizar o mapeamento de trabalhos relacionados, proporcionando assim identificar o estado da arte e as oportunidades de pesquisa encontradas nessa linha de pesquisa.

No Capítulo 4, é detalhada a proposta, assim como a metodologia desenvolvida para a sua concretização.

No Capítulo 5, são apresentados e discutidos os resultados obtidos com os experimentos realizados.

Por fim, no Capítulo 6, são elencadas as considerações finais obtidas ao longo da pesquisa e perspectivas de trabalhos futuros acerca do tema.

2 Fundamentação Teórica

Esta seção apresenta alguns conceitos necessários para melhor compreensão deste projeto. Tendo em vista que esta é uma linha de pesquisa relativamente nova, é indicado, inicialmente, contextualizar o problema da proliferação das *fake news*, expondo suas características e formas de detecção. Em seguida, são abordadas algumas estruturas de Redes Neurais Artificiais aplicadas na presente pesquisa.

2.1 Fake News

O compartilhamento de notícias falsas não é um fenômeno recente. Este problema existe há 500 anos, com a invenção da imprensa, quando notícias passaram a ser amplamente divulgadas. Desde então, inúmeros desafios vêm se somando no contexto das *fake news*, desde a sua definição até sua detecção e intervenção de propagação.

Atualmente, o termo "*Fake News*" tem sido largamente utilizado para referir-se a desinformação e notícias falsas compartilhadas, principalmente, por meio das mídias sociais. Apesar disto, ainda não existe um conceito universal para notícias falsas, mesmo no meio jornalístico (ZHOU; ZAFARANI, 2020). Entretanto, com a evolução das pesquisas acadêmicas, inúmeras definições têm surgido, como a declarada pelos autores de (ALLCOTT; GENTZKOW, 2017), que conceituam *fake news* como artigos de notícias cujos conteúdos são intencional e comprovadamente falsos e capazes de enganar os leitores.

Uma revisão sobre estudos anteriores encontrados na literatura revela que diversos tipos de conteúdo são categorizados como *fake news*, dentre eles estão:

- **Boatos:** conhecidos popularmente como “disse me disse”, os boatos consistem em notícias publicadas cuja fonte é desconhecida, em grande parte sem nenhum fundamento, e que acabam sendo amplamente compartilhados e divulgados.
- **Sátiras:** são paródias carregadas de ironia e humor, apresentadas em um formato típico do jornalismo e muito populares nas redes sociais e em sites sensacionalistas. Existem diferentes tipos de sátiras, como: a que usa comentários satíricos e esquetes cômicos para comentar eventos de notícias do mundo real e as que apresentam notícias totalmente fictícias.
- **Rumores:** notícias que normalmente apresentam ambiguidade e que, no momento do seu compartilhamento, não tem sua veracidade validada. Segundo os autores de (ZUBIAGA et al., 2018), de acordo com o dicionário de *Oxford English Dictionary*, um rumor consiste em “uma história ou relatório, atualmente em circulação, de verdade incerta ou duvidosa”.

- **Clickbait:** trata-se de uma estratégia de manipulação de conteúdo que tem como objetivo suscitar a curiosidade ou o interesse exagerado em leitores, com o propósito de incentivá-los a acessarem em determinados conteúdos mesmo que estes sejam menos impactantes ou precisos do que o título implica. Em outras palavras, o *clickbait* corresponde a uma estratégia intencional que visa atrair a atenção e incentivar os visitantes a consumir informações falsas e compartilhá-las (KUMAR et al., 2021) (CHEN; CONROY; RUBIN, 2015).
- **Propagandas:** consiste em um método de propagar notícias de forma concentrada, ou seja, com fins específicos. Muitas vezes são direcionadas para questões ideológicas, como promoção de ideias, causas e princípios políticos ou religiosos.
- **Notícias Enviesadas:** são notícias que apresentam um conteúdo agudamente tendencioso. Normalmente, estas notícias estão vinculadas às questões políticas, tais como partidos, figuras políticas e eventos.
- **Teorias da conspiração:** de acordo com a (UNESCO, 2021), as teorias da conspiração são uma tentativa de criar uma narrativa que promete atribuir sentido a eventos que de outro modo são aleatórios, tomando-os como parte de um enredo vasto e abrangente. Ou seja, são histórias que buscam de alguma forma fornecer explicações a eventos sem fundamento.

Ao considerar a afirmação de (GOLBECK et al., 2018), que definem *fake news* como notícia falsa, apresentada como verdadeira, e que tem intenção de enganar, pode-se concluir que sátiras e paródias de notícias não devem ser consideradas como meios de desinformação, pois não têm o intuito de enganar, desvelando sua falsidade pelo humor (RECUERO; GRUZD, 2019).

Nos últimos anos, o tópico *fake news* tem ganhado notoriedade diante do público e da mídia, principalmente após as eleições presidenciais dos Estados Unidos da América (EUA), em 2016. Desde então, a disseminação de notícias falsas tem se tornado uma grande preocupação para a sociedade, sendo tema recorrente de debates e pesquisas, devido à sua capacidade de manipular a opinião pública, bem como suas implicações político-sociais.

Os autores (ALLCOTT; GENTZKOW, 2017) descobriram, em suas pesquisas, que indivíduos são mais inclinados a acreditar em informações que estejam ideologicamente alinhadas com suas próprias crenças pessoais. Por conta disso, muitas publicações buscam manipular o leitor através de suas crenças, sobrepondo os fatos com apelos emocionais. Desta forma, esta vulnerabilidade individual pode ser explorada para a proliferação de *fake news* (SHARMA et al., 2019). Esta sobreposição dos fatos por apelos emocionais e crenças individuais, onde uma narrativa toma mais importância que os fatos, é conhecida como "pós-verdade".

2.2 Processamento de Linguagem Natural (PLN)

Processamento de linguagem natural (PLN) é a tradução do termo em inglês *Natural Language Processing (NLP)*, que consiste em uma área da computação conhecida por ser uma subárea da inteligência artificial. Segundo (VIEIRA; LOPES, 2010), PLN tem como objetivo trazer soluções de programas de computadores que analisam, reconhecem e geram textos em linguagens humanas ou naturais.

A técnica é um mecanismo criado não só para extrair as informações de textos, mas também um facilitador na entrada de dados e na sua estruturação. No PLN, é considerada a estrutura hierárquica da linguagem: várias palavras formam uma frase, várias frases fazem uma sentença e, em última análise, sentenças transmitem ideias (BARBOSA et al., 2017).

Destaca-se que, uma análise de texto usando PLN, por meio da compreensão e extração de seu significado, pode ser composta por cinco níveis:

- **Análise Morfológica:** Concentra-se na estrutura das palavras, explorando elementos como prefixos, sufixos e raízes.
- **Análise Sintática:** Trata das relações entre palavras em uma frase, revelando a organização gramatical e hierárquica.
- **Análise Semântica:** É construída com base nos resultados dos outros dois níveis anteriormente citados. Tendo como objetivo definir palavras e sentenças com base no conhecimento de sua estrutura, relacionamento e função.
- **Análise Pragmática:** Análise do significado em contexto, considerando a intenção do falante, a situação e as inferências.
- **Análise Discurso:** Focado na análise e compreensão de estruturas maiores de texto, como parágrafos, diálogos e conversações. Ele se concentra na coesão e coerência do discurso, ou seja, como as diferentes partes do texto se relacionam para formar uma mensagem coesa e lógica.

Destaca-se que boa parte das abordagens de mineração de texto se baseiam em um documento, que nada mais é do que um conjunto de palavras. Todavia, para a classificação desta estrutura, cada palavra passa a ser representada por um vetor com sua “importância numérica”, conhecido como processo de vetorização. Entretanto, até chegar nesta etapa, algumas técnicas de pré-processamento de texto são necessárias, sendo elas:

- **Tokenização:** Também conhecida como segmentação de palavras, a *tokenização* é um processo que envolve a divisão de uma sequência de caracteres em um texto, identificando os limites de cada palavra. Em outras palavras, a *tokenização* localiza os pontos onde uma palavra encerra e a próxima começa (BARBOSA et al., 2017). Esse processo é crucial no PLN, já que prepara o texto para análises mais aprofundadas, como análise morfológica e sintática, ao transformar

o texto em unidades significativas chamadas "tokens". Cada token pode ser uma palavra, pontuação, número ou símbolo, dependendo do contexto e das regras de *tokenização* aplicadas.

- **Lematização:** A lematização é um processo que estabelece uma relação entre as diferentes formas morfológicas de uma palavra e seus "*lemmas*", que representam as formas canônicas ou base dessas palavras conforme encontradas em um dicionário. Em vez de apenas remover a flexão das palavras, como ocorre na radicalização, a lematização envolve a redução das palavras às suas formas fundamentais, permitindo a análise e comparação mais precisa das palavras, independentemente de suas variações morfológicas.
- **Radicalização:** (ou "*Stemmização*", do inglês *Stemming*) que consiste em uma técnica computacional de pré processamento de textos, onde palavras mais complexas morfolologicamente são identificadas, decompondo em seu *Stem* invariante, ou melhor, na forma canônica do *lemma* e seus afixos, que ao fim do processo são deletados.
- **Stopwords:** O processo de remoção de "*stopwords*" envolve a exclusão de palavras e termos que são considerados irrelevantes para uma análise específica. Isso geralmente inclui palavras comuns, como artigos, preposições e numerais, que têm pouco impacto na compreensão do conteúdo semântico essencial do texto. A eliminação de "*stopwords*" é uma etapa importante no processamento de linguagem natural, especialmente em tarefas como análise de sentimentos e modelagem de tópicos, onde a presença dessas palavras não contribui significativamente para a extração de informações-chave e pode afetar negativamente os resultados da análise.
- **Filtragem:** A filtragem de texto inclui a remoção de *stopwords*, a detecção e remoção de caracteres especiais ou símbolos, a identificação de entidades nomeadas (como nomes de pessoas ou lugares) e a eliminação de ruídos ou informações irrelevantes.

2.3 Modelos de Representação Textual

É importante ressaltar que os algoritmos computacionais são capazes de compreender apenas dados estruturados. Assim, no contexto do processamento de texto, uma tarefa essencial é a construção de modelos de representação textual.

2.3.1 Term Frequency-Inverse Document Frequency (TF-IDF)

O *Term Frequency-Inverse Document Frequency* (TF-IDF) consiste em uma técnica amplamente usada no PLN e recuperação de informações para avaliar a importância relativa de uma palavra em um documento dentro de um conjunto de documentos.

Esta representação é amplamente aplicada para a ponderação de palavras com base em quão frequentemente aparecem em um documento específico em comparação com sua frequência em todo o corpus de documentos. O TF-IDF é composto por duas partes principais:

- *Term Frequency* (TF - Frequência do Termo): Essa parte mede a frequência com que um termo aparece em um documento específico. Geralmente, é calculada como a razão entre o número de vezes que um termo ocorre em um documento e o número total de termos no documento.
- *Inverse Document Frequency* (IDF - Frequência Inversa de Documento): Essa parte avalia a raridade de um termo em relação a todo o corpus de documentos. Ela é calculada como o logaritmo do número total de documentos dividido pelo número de documentos que contêm o termo.

O método clássico TF-IDF combina duas estatísticas essenciais: a Frequência de Termo (TF) e a Frequência Inversa de Documento (IDF), conforme proposto por (JOACHIMS, 1997). O TF captura a frequência com que um termo específico, representado por 't', ocorre em um documento 'd'. Por sua vez, o IDF avalia o valor informativo desse termo em relação a todo o corpus de documentos. Isso é calculado tomando o logaritmo do número total de documentos no corpus 'N' dividido pelo número de documentos que contêm o termo 't' ('dft'). Essa abordagem resulta em um peso 'w' para o termo 't' no documento 'd', representado como:

$$W_{t,d} = tf_{t,d} * \log\left(\frac{N}{df_t}\right) \quad (2.1)$$

Essa fórmula permite a atribuição de um valor que reflete a importância relativa de 't' dentro do contexto de 'd' em relação ao conjunto completo de documentos.

Em resumo, o TF-IDF é obtido multiplicando-se o *Term Frequency* (TF) pelo *Inverse Document Frequency* (IDF) para cada termo em um documento específico. Essa pontuação resultante é uma representação numérica que reflete a importância de cada termo no contexto desse documento em relação a todo o conjunto de documentos. Quanto mais alto o valor do TF-IDF, mais importante a palavra é considerada para aquele documento.

2.3.2 Word2Vec

Assim como o TF-IDF, o *Word2Vec* consiste em uma técnica representação de palavras em linguagem natural. Todavia, o mesmo realiza o mapeamento das palavras individualmente para vetores contínuos em um espaço de alta dimensão. Essa representação vetorial é projetada de forma que palavras semanticamente semelhantes tenham vetores próximos no espaço, o que permite que o modelo capture significados e relações entre palavras.

Ele foi desenvolvido por Tomas Mikolov e seus colegas em uma série de artigos (MIKOLOV et al., 2013). No qual, destaca que o processo da representação começa com uma arquitetura de redes neurais e, à medida que avança, aprimora continuamente o vetor de palavras por meio de redes neurais, resultando em melhorias no desempenho. Em vez de depender da contagem de palavras e da criação de uma matriz de co-ocorrência, esse modelo refinado captura as características linguísticas de uma linguagem natural de maneira indireta.

Existem duas principais abordagens para o treinamento do *Word2Vec*:

- *Continuous Bag of Words* (CBOW): Nesta abordagem, o modelo tenta prever uma palavra-alvo com base em um contexto (uma janela de palavras vizinhas) em torno dela. O objetivo é maximizar a probabilidade de prever corretamente a palavra-alvo dadas as palavras do contexto. O CBOW é eficaz para representar o contexto das palavras.
- *Skip-gram*: Na abordagem *Skip-gram*, o objetivo é o oposto do CBOW. O modelo tenta prever as palavras do contexto com base em uma palavra de destino. O treinamento visa maximizar a probabilidade de prever as palavras do contexto a partir da palavra de destino. O Skip-gram é eficaz para capturar relações semânticas e analogias entre palavras.

O *Word2Vec* reúne palavras que compartilham significados semelhantes no espaço vetorial. Em outras palavras, esse algoritmo é capaz de criar representações distribuídas das palavras de maneira eficaz, preservando as relações semânticas que elas têm com as palavras vizinhas e resultando na formação de agrupamentos semânticos, como ilustrado na Figura 1.

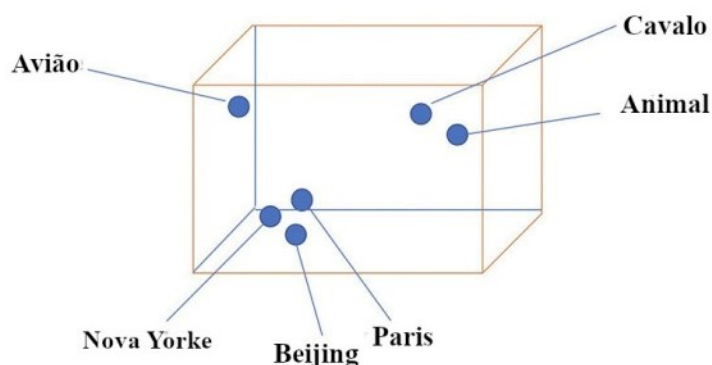


Figura 1 – Representação do Word2vec (Fonte: (AMAZON, 2018)).

2.3.3 Continuous Bag of Words (CBOW)

O *Continuous Bag-of-Words* (CBOW), como mencionado, é um dos dois modelos principais utilizados no *Word2Vec*. Este modelo é uma técnica de aprendizado de máquina

que visa aprender representações vetoriais de palavras com base em grandes volumes de texto.

Em essência, o CBoW é uma rede neural que faz previsões de palavras-alvo considerando o contexto das palavras circundantes contidas em uma janela definida. Nesta variante do *Word2Vec*, a camada de entrada consiste em um vetor *one-hot* que representa as palavras vizinhas, a camada oculta é onde ocorre o processamento do algoritmo e, por fim, a camada de saída produz a palavra prevista.

A Figura 2 mostra o modelo CBoW com uma configuração de contexto de várias palavras. Destaca-se que, ao calcular a saída da camada oculta, em vez de copiar diretamente o vetor de entrada da palavra de contexto de entrada, o modelo CBoW captura a média dos vetores das palavras de contexto de entrada e usa o produto da matriz de peso de entrada→oculta e o vetor médio como saída (RONG, 2016).

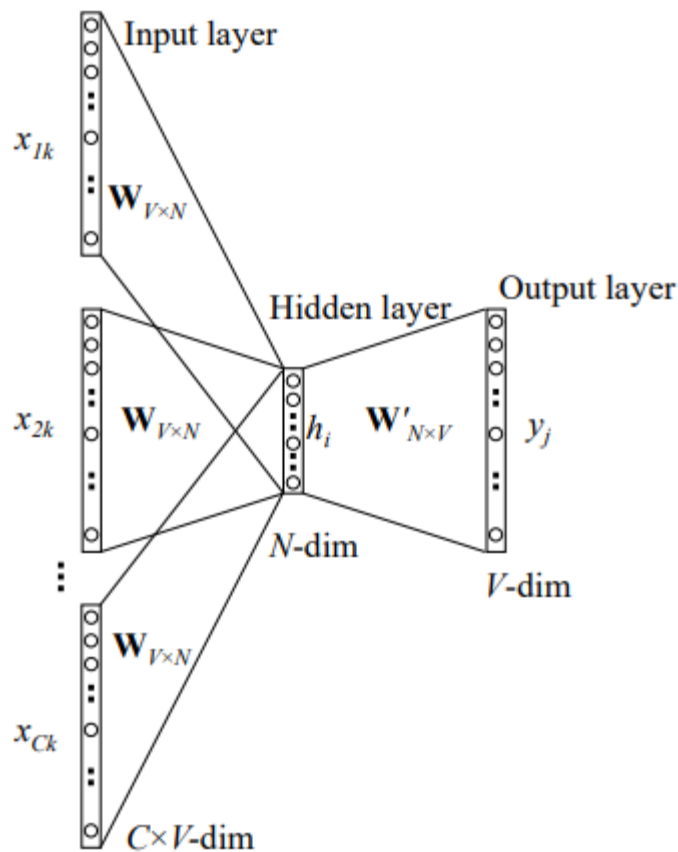


Figura 2 – Representação do CBoW (Fonte: (RONG, 2016)).

2.3.4 GloVe

O *Global Vectors for Word Representation* (GloVe), consiste em um modelo de aprendizado de máquina usado para aprender representações vetoriais de palavras em linguagem natural. Assim como o *Word2Vec*, o *GloVe* é uma técnica amplamente utilizada

para representação de palavras. Sua principal característica é que ele combina informações globais e locais para criar as representações vetoriais das palavras.

Destaca-se que o mesmo é um modelo global de regressão que combina as vantagens da Fatoração de Matrizes e da Janela de Contexto Local (PENNINGTON; SOCHER; MANNING, 2014). Os autores provam que a razão entre as probabilidades de co-ocorrência apresenta certos aspectos semânticos.

Desta forma, seja $P_{i,j} = P(j|i)$ a probabilidade de uma palavra j aparecer no contexto da palavra i . Sejam w_i e w_j palavras num contexto $\tilde{w}_k$ uma palavra em outro contexto. Tem-se que a função de associação é dada por:

$$F(w_i, w_j, \tilde{w}_k) = \frac{P_{ik}}{P_{jk}} \quad (2.2)$$

Dado que o GloVe compartilha princípios semelhantes com o Word2Vec, foi desenvolvida a seguinte função de custo para o GloVe:

$$J = \sum_{i,j=1}^V f(X_{ij})(w_i^T \tilde{w}_j + b_i + \tilde{b}_j - \log X_{ij})^2 \quad (2.3)$$

Na qual V é o tamanho do vocabulário, b são os vieses e $f(X_{ij})$ é uma função de ponderação na função de custo. A função de ponderação por sua vez é dada por:

$$f(x) = \begin{cases} (\frac{x}{x_{max}})^\alpha & x < x_{max} \\ 1 & x \geq x_{max} \end{cases} \quad (2.4)$$

2.4 Redes Neurais Convolucionais (CNN)

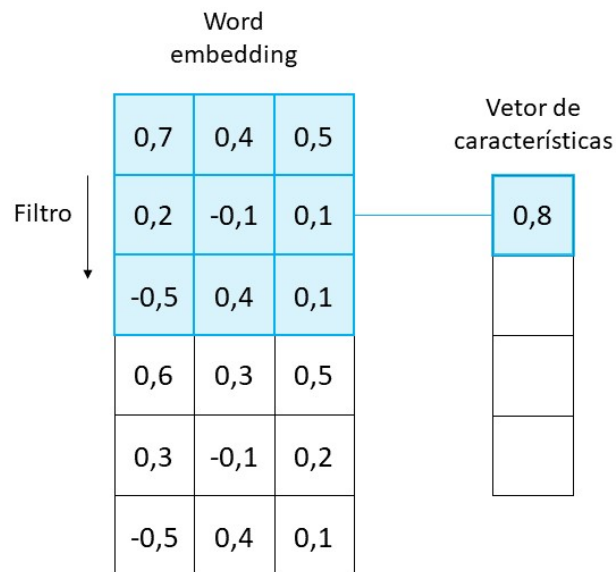
A Rede Neural Convolucional (do inglês *Convolutional Neural Networks* - *CNN*), é um algoritmo de *Deep Learning* (*DL*), inspirado em modelos biológicos (MATSUGU et al., 2003). As CNNs foram propostas no estudo de (LECUN; BENGIO, 1998), na qual consiste em um tipo de perceptron multicamadas que funciona de forma que os neurônios se conectam numa distribuição semelhante ao do córtex visual humano.

Tendo isso em vista, as CNNs passaram a ser fortemente aplicadas em soluções de visão computacional, entretanto a aplicação do modelo em tarefas de PLN também tem alcançado resultados satisfatórios, como, por exemplo, no *parsing* semântico e no processamento de texto (SHARMA et al., 2019). Nesta pesquisa, o CNN foi construído com a utilização de classes das bibliotecas *Keras* e *Tensorflow*.

No cenário de aplicação do CNN para processamento de texto se faz necessário que, previamente os dados sejam submetidos ao processo de transformação, que consiste

em converter dados de formato textual para representações numéricas. Desta forma, é aplicada uma técnica de PLN denominada *Word Embedding*. Esta técnica é utilizada para converter textos em vetores multidimensionais (matrizes) de baixo nível, de forma que termos cujo significado são semelhantes sejam representados por valores próximos entre si, preservando, assim, a relação semântica entre eles (LEVY; GOLDBERG, 2014). Dentre os métodos mais conhecidos de *Word Embedding*, pode-se citar o *Word2Vec* e o *GloVe*.

Figura 3 – Processo de convolução.



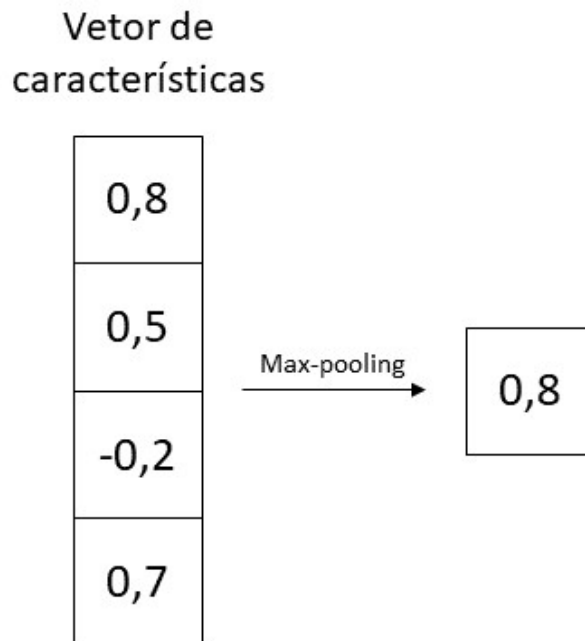
Fonte: Elaborada por (SOUSA et al., 2022)

Após a aplicação da técnica *Word Embedding*, os textos transformados são submetidos a uma camada convolucional. Esta camada tem como objetivo extrair recursos de alto nível dos dados de entrada e os repassa para a próxima camada na forma de mapas de recursos. No caso, são aplicados filtros sobre os dados de entrada, a fim de extrair as características dos textos. Este filtro, também conhecido como *kernel*, consiste em uma matriz com quantidade menor de linhas em relação à matriz de entrada, todavia, com a mesma quantidade de colunas. Desta forma, o filtro funciona como uma janela deslizante, deslocando-se de cima para baixo pela matriz, como representado na Figura 3. Destaca-se que o filtro contém inicialmente valores aleatórios, que passam a ser atualizados a cada iteração.

Dando continuidade, ao fim da camada convolucional é gerado um vetor resultante, denominado vetor de características. Este vetor passa a ser processado utilizando-se um método denominado *max pooling*, que constitui-se de uma função que é aplicada sobre os conjuntos de valores do vetor de características, a fim de manter apenas o maior valor, como pode ser visualizado na Figura 4. Este processo permite extrair as características mais relevantes do texto.

O processo de *max pooling* apresenta como saída vários valores, correspondentes

Figura 4 – Max-Pooling.



Fonte: Elaborada por (SOUSA et al., 2022)

ao número de filtros utilizados. Por este motivo, é necessário agrupá-los novamente na camada de *Flatten*, onde os valores são concatenados, formando um novo vetor. Este novo vetor é processado por uma nova camada totalmente conectada com a função de ativação *Rectified Linear Unit (ReLU)*, que, ao final, por tratar-se de um problema de classificação binária, utiliza-se a função de ativação sigmoidal. Na qual, a saída não é linear, ou seja, a função vai variar de 0 a 1 tendo um formato "S". No caso, a função essencialmente tenta alocar os valores para os extremos, na tentativa de classificá-los em uma classe específica.

2.5 Redes de Memória de Longo e Curto Prazo

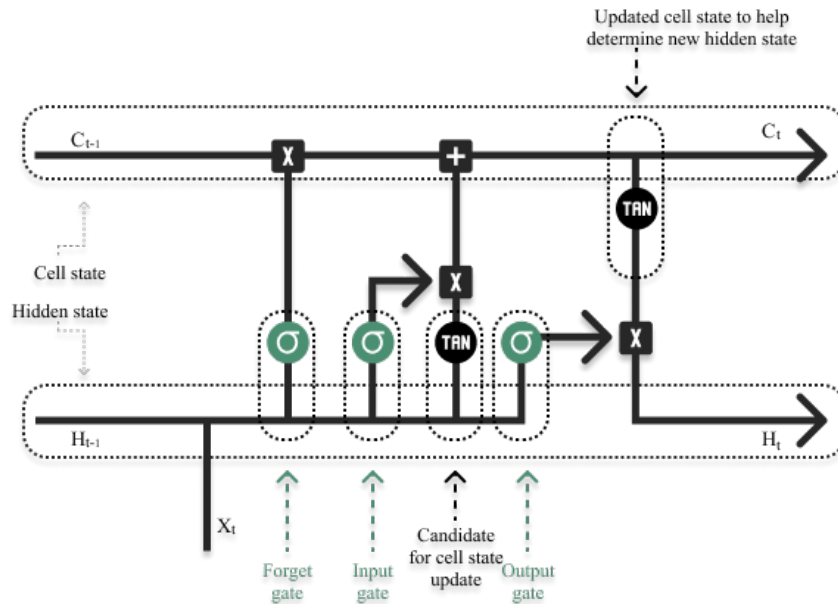
As Redes de Memória de Longo e Curto Prazo, mais conhecidas por seu termo em inglês *Long Short-Term Memory (LSTM)*, foi introduzido por Hochreiter e Schmidhuber (1997), que buscaram apresentar uma arquitetura de Rede Neural Recorrente (do inglês *Recurrent Neural Network (RNN)*) em conjunto com um algoritmo de aprendizado baseado em gradiente apropriado.

O LSTM, assim como o CNN, consiste em um algoritmo de aprendizado profundo, caracterizado por ser uma RNN capaz de armazenar informações por longos períodos. Destaca-se que a estrutura do modelo foi projetada com o objetivo de fazer com que a rede armazene uma maior quantidade de entradas passadas, isso sem perder os estados mais recentes. Sendo assim, o modelo se apropria dos estados dominantes *hidden state* e *cell state*.

O LSTM possui uma estrutura em cadeia, assim como as RNN. Todavia, em

vez de uma camada de rede o modelo apresenta quatro camadas. Em se tratando do funcionamento, a princípio, os dados de entrada passam pelo *Forget gate*, que realiza o controle de quais informações devem ser esquecidas, como pode ser observado na Figura 5. Destaca-se que este processo é realizado por meio da função *sigmóide*, que varia entre 0 e 1, sendo (0) os valores no estado da célula que devem ser descartados, (1) lembrados ou (algun valor entre 0 e 1) parcialmente lembrados. Em seguida parte-se para o *Input gate*, que tem como objetivo identificar elementos importantes que precisam ser adicionados ao *cell state*. Vale ressaltar que os resultados do *Input gate* são multiplicados pelo candidato ao *cell state*, com apenas as informações consideradas importantes pela *Input gate*, sendo adicionadas ao *cell state*.

Figura 5 – Arquitetura do LSTM.



Fonte: Elaborada pelos autores

Dando continuidade, após o processo mencionado anteriormente, é realizada uma atualização no *cell state*. Para isto, o estado anterior da célula é multiplicado pelos resultados do *Forget gate*, como está exibido na Figura 5, logo depois novas informações de [*Input gate* $\times$ *Cell state candidate*] são adicionadas, tendo em vista obter-se o estado da célula mais recente. A última parte é atualizar o estado oculto, no qual, o último estado da célula é passado pela função de ativação *tanh* e multiplicado pelos resultados da porta de saída.

3 Revisão Sistemática

Este capítulo apresenta a metodologia adotada para realização de uma Revisão Sistemática (RS) no contexto da detecção automática de notícias falsas. A metodologia está dividida em três etapas principais (KEELE et al., 2007), apresentadas a seguir. Todas as etapas foram realizadas com o auxílio do programa *Parsifal*¹.

3.1 Etapa 1: Planejamento

O planejamento compreende a realização da revisão bibliográfica, possibilitando uma familiarização com a temática a ser estudada e auxiliando no preenchimento dos protocolos de busca (CERRAO; CASTRO; JESUS, 2018). As atividades de planejamento são: (I) Definir as Questões de Pesquisa (QP) da revisão; (II) Produzir um protocolo de busca; (III) Estabelecer os critérios de qualidade e (IV) Determinar os critérios de inclusão e exclusão.

Tabela 1 – Tabela PICO.

Elemento	Descrição	Keywords
População	Meios sociais de propagação de fake news.	Social media e Social networks
Intervenção	Deteção de fake news, desinformação, boatos e rumores	fake news, Rumor e Misinformation
Comparação	Comparar desempenho dos algoritmos de IA	Machine learning e Deep learning
Outcomes	Validação dos resultados alcançados	-

3.1.1 Questões de Pesquisa (QP)

As questões que norteiam a presente RS são:

- Q1: Como estão distribuídos os trabalhos de detecção automática de *fake news* ao longo dos anos?
- Q2: Quais os tipos de publicações científicas mais recorrentes no contexto das *fake news*?
- Q3: Quais países têm mais pesquisadores publicando sobre detecção de *fake news*?
- Q4: Quais tipos de conteúdo estão sendo considerados como solução na detecção de *fake news*?
- Q5: Quais as redes sociais mais utilizadas no desenvolvimento dos trabalhos?
- Q6: Quais os *datasets* mencionados nos trabalhos selecionados?
- Q7: Quais algoritmos de inteligência artificial são aplicados nas pesquisas?
- Q8: Quais estudos apresentam resultados mais qualificados quanto às métricas de avaliação?

¹ <https://parsif.al/>

3.1.2 Protocolos de busca

Para os protocolos de busca foi aplicada uma tabela PICO (População, Intervenção, Comparação e Resultados). Desta forma, o protocolo de busca foi definido conectando os termos presentes na Tabela 1, usando o termo lógico “AND” e “OR” em todas as *keywords*.

A busca foi executada em cinco portais científicos, sendo eles:

- *ACM Digital Library* ²
- *IEEEExplore* ³
- *ScienceDirect* ⁴
- *SpringerLink* ⁵
- *SBC OpenLib* ⁶

A Tabela 2 apresenta os protocolos de busca utilizados em cada portal científico, já que cada um tem *layout* e sintaxe de busca diferente.

Tabela 2 – Sintaxe de busca.

Portal	Protocolo de busca
<i>ACM</i>	"query": {Fulltext:(social media OR social networks) AND Fulltext:(Fake News OR rumour OR misinformation) AND Fulltext:(machine learning OR deep learning)}
<i>IEEEExplore</i>	("Full Text & Metadata":social media) OR ("Full Text & Metadata":social networks) AND ("Full Text & Metadata":Fake News) OR ("Full Text & Metadata":rumour) OR ("Full Text & Metadata":misinformation) AND ("Full Text & Metadata":machine learning) OR ("All Metadata":deep learning)
<i>ScienceDirect</i>	all:{(social media OR social networks) AND (Fake News OR rumour OR misinformation) AND (machine learning OR deep learning)}
<i>SpringerLink</i>	detection AND learning AND social AND "Fake News" AND (machine OR deep OR media OR network OR news) filters: computer science And article
<i>SBC OpenLib</i>	Busca manual

3.1.3 Critérios de Qualidade

Os critérios de qualidade têm como propósito fornecer critérios de inclusão/exclusão mais detalhados, além de investigar se as variações de qualidade fornecem uma explicação para as diferenças nos resultados do estudo (KEELE et al., 2007). As avaliações de qualidade geralmente são baseadas em uma lista de verificação de fatores que precisam ser ponderados para cada estudo. Os itens estabelecidos no processo de avaliação recebem uma “nota”, que influencia na aprovação ou reprovação do estudo.

² <https://dl.acm.org/>

³ <https://ieeexplore.ieee.org/Xplore/home.jsp>

⁴ <https://www.sciencedirect.com/>

⁵ <https://link.springer.com/>

⁶ <https://sol.sbc.org.br/index.php/indice>

Para a seleção dos trabalhos, 10 critérios de qualidade foram estabelecidos. Para cada critério, há três possíveis respostas, sendo elas: **Presente (peso=10)**; **Parcialmente (peso=5)**; **Ausente (peso=0)**. Desta forma, o artigo que atingir nota maior ou igual 7 é classificado como aprovado. Os critérios são:

- C1: A motivação e a justificativa apresentadas no trabalho são evidentes e bem fundamentadas?
- C2: Os objetivos foram estabelecidos de forma coesa?
- C3: A pesquisa foi bem fundamentada teoricamente?
- C4: O trabalho expõe de forma transparente e objetiva a metodologia desenvolvida, sendo justificadas as decisões tomadas?
- C5: A solução proposta foi detalhada completamente?
- C6: A base de dados utilizada no trabalho foi apresentada de forma adequada? A sua escolha foi justificada?
- C7: Os resultados apresentados foram relatados?
- C8: As figuras e tabelas estão legíveis? As mesmas foram devidamente relatadas ao longo do trabalho?
- C9: Os resultados são consistentes com os objetivos apresentados?
- C10: De forma geral, o pesquisador conseguiu apresentar um trabalho claro e objetivo?

3.1.4 Critérios de Inclusão e Exclusão

Os critérios de inclusão foram:

- Trabalhos que apresentem um modelo com resultados;
- Apresentam métricas de avaliação como acurácia, F1, precisão e/ou revocação;

Os critérios de exclusão foram:

- Trabalhos não escritos em inglês ou português;
- Trabalhos publicados antes de 2016;
- Trabalhos em andamento;
- Trabalhos duplicados;
- Estudos muito específicos, como por exemplo: detecção com imagens ou apenas focado no perfil dos autores de *fake news*;
- Trabalhos de cunho teórico como, *Surveys* e revisões sistemáticas ou literárias;
- Trabalhos que não apresentem em seu texto referência à base de dados utilizada na pesquisa;
- Trabalhos que apresentem *dataset* com menos de 1.000 notícias verdadeiras e falsas;
- Trabalhos que não atingiram a pontuação estabelecida nos critérios de qualidade.

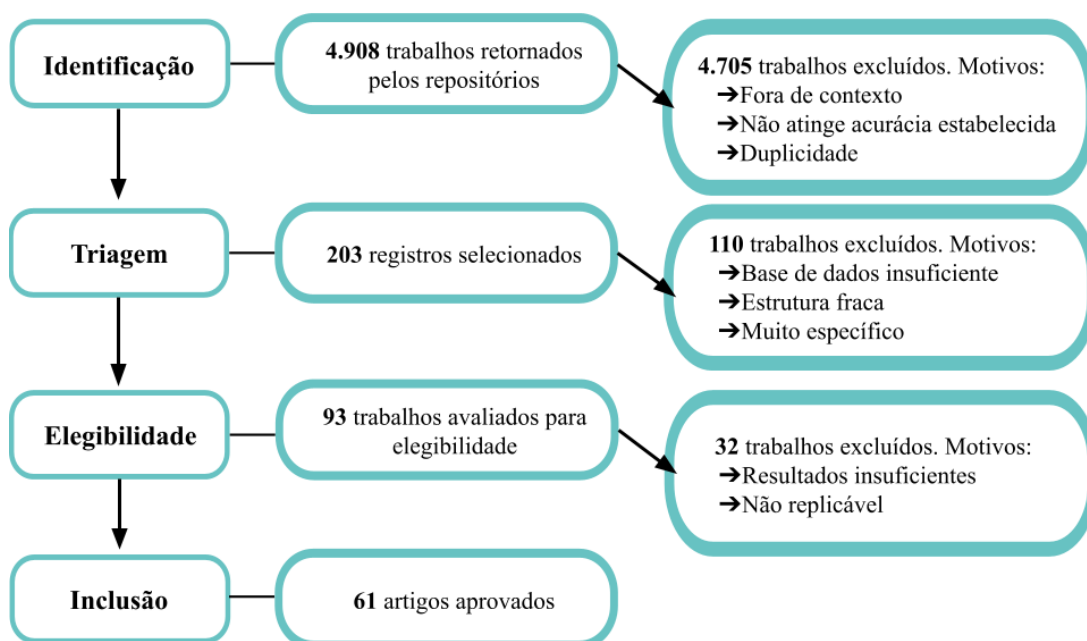
3.2 Etapa 2: Condução

Ao fim da etapa de planejamento, é iniciado o processo de condução. Ao todo foram reunidos 4.908 estudos. Vale ressaltar que os trabalhos coletados foram publicados dentro de um período de 7 anos, compreendidos entre 2016 e novembro de 2022 (ano do desenvolvimento do presente estudo). De posse dos trabalhos, o processo de avaliação para adesão ou eliminação é instaurado. Para isto, são utilizadas as recomendações PRISMA, Figura 6, que consistem em: Identificação, Seleção, Elegibilidade e Inclusão.

Na **Identificação**, pontos estratégicos foram avaliados com o objetivo de acelerar a aferição dos artigos. Foram analisados os títulos, *abstracts* e *keywords*, além de verificar artigos duplicados. Esse processo resultou na eliminação de 4.705 artigos. Neste ponto, com 203 artigos, é realizada a **Seleção** dos trabalhos. O processo consiste na aplicação dos 10 critérios de qualidade da Seção 3.1.3 e avaliação mediante às notas obtidas, o que resultou na eliminação de 110 trabalhos.

Com 93 estudos resultantes do processo de seleção, é iniciada a análise quanto à **Elegibilidade**. Nesse ponto são avaliadas a replicabilidade dos estudos e a qualidade dos resultados, não apenas acurácia, mas também precisão, matriz de confusão entre outras métricas de avaliação. Desta forma, 32 estudos foram eliminados. Por fim, 61 trabalhos foram **Incluídos** para a continuação da presente revisão.

Figura 6 – Diagrama PRISMA.



3.3 Resultados

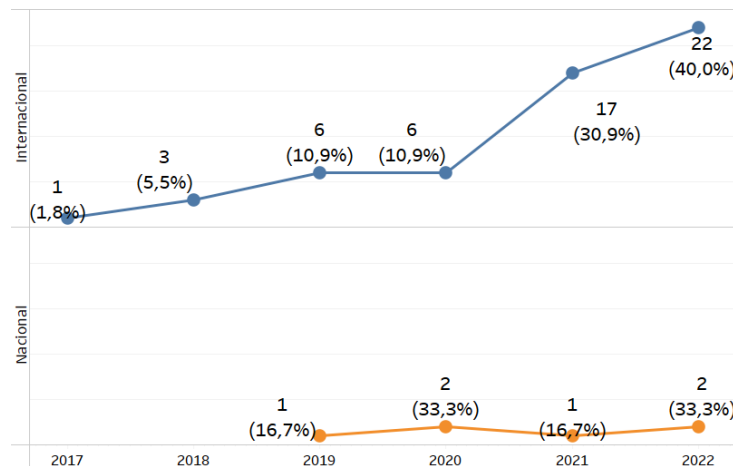
Nesta seção, são apresentados os resultados alcançados a partir da análise dos 61 artigos da RS, respondendo as QP da Seção 3.1.1. As análises são dispostas em dois cenários: Nacional (pesquisas com *dataset* em Língua Portuguesa) e Internacional (com *dataset* em idioma diferente do português). Os 61 artigos analisados foram:

(KHAN et al., 2021), (SOUZA et al., 2020b), (LI et al., 2021), (GADEK; GUÉLORGET, 2020), (QAWASMEH; TAWALBEH; ABDULLAH, 2019), (CHEN et al., 2021), (ABDELMINAAM et al., 2021), (AYOUB; YANG; ZHOU, 2021), (PAKA et al., 2021), (RUCHANSKY; SEO; LIU, 2017), (MORAES; SAMPAIO; CHARLES, 2019), (ZIANI et al., 2021), (DEEPAK; CHITTURI, 2020), (KALIYAR et al., 2020b), (SHU et al., 2019), (SASTRAWAN; BAYUPATI; ARSA, 2021), (WANG et al., 2018), (GEREME; ZHU, 2019), (SAWCZYN et al., 2021), (KALIYAR, 2018), (BAHAD; SAXENA; KAMAL, 2019), (NASIR; KHAN; VARLAMIS, 2021), (ZHOU et al., 2020), (AJAO; BHOWMIK; ZARGARI, 2018), (KALIYAR; GOSWAMI; NARANG, 2021), (KALIYAR et al., 2020a), (SEPÚLVEDA-TORRES et al., 2021), (NAYOGA et al., 2021), (ALMEIDA et al., 2021), (CHOUDRIE et al., 2021), (CHAUHAN; PALIVELA, 2021), (VEYSEH et al., 2019), (AHUJA; KUMAR, 2020), (MEESAD, 2021), (DONG; VICTOR; QIAN, 2020), (TESTONI et al., 2021), (BARBOSA et al., 2022), (ZHANG; HUANG; KHAN, 2022), (MAHMUD et al., 2022), (LAN et al., 2022), (HARRAG; DJAHLI, 2022), (MALHOTRA; MAHUR; ACHINT, 2022), (JAIN et al., 2022), (HAN et al., 2022), (SOUSA et al., 2022), (WU; LUO; DEMERTZIS, 2022), (RATHOD, 2022), (MARTÍN et al., 2022), (SHAILENDRA et al., 2022), (MATHEWS; PREETHI, 2022), (RANJAN; AGRAWAL, 2022), (KRISHNA; ADIMOOLAM, 2022), (HOTI et al., 2022), (SHISHAH, 2022), (FAYAZ et al., 2021), (PALANI; ELANGO; K, 2022), (ZERVOPOULOS et al., 2022), (SADEGHI; BIDGOLY; AMIRKHANI, 2022), (AGRAWAL; PANDEY; GOYAL, 2022), (KHAN; THAKUR, 2022) e (KHAN; THAKUR, 2022).

3.3.1 Como estão distribuídos os trabalhos de detecção automática de fake news ao longo dos anos?

Na Figura 7 é apresentada a distribuição anual das pesquisas selecionadas no contexto de detecção automática de *fake news* em nível nacional e internacional, na qual o número de estudos era escasso em 2017. Destaca-se que, nacionalmente, não há registros de estudos sobre este tema neste ano. Entre 2018 a 2022, é apresentado um aumento no número de estudos, sendo 2022 o ano com maior número de pesquisas, tanto internacionalmente, quanto nacionalmente, com 22 (40%) e 2 (33,33%) estudos, respectivamente. É importante destacar que este aumento no volume de estudos durante o ano de 2021 e 2022 pode ser atribuído à preocupação em torno da disseminação de *fake news* relacionadas à pandemia causada pelo COVID-19.

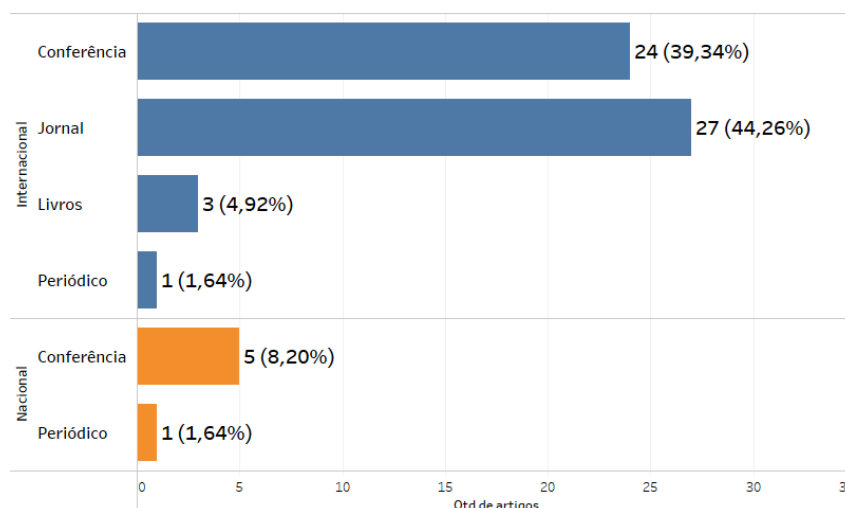
Figura 7 – Distribuição anual dos trabalhos.



3.3.2 Quais os tipos de publicações científicas mais recorrentes no contexto das fake news?

A Figura 8 exibe os tipos de publicações científicas mais recorrentes no contexto das notícias falsas. Como pode ser observado, no cenário internacional 27 (44,26%) dos artigos selecionados são publicados em jornais, já no cenário nacional o maior número de artigos é em conferências, no caso, 5 (8,20%) trabalhos. Todavia, no cenário internacional a diferença para o número de trabalhos publicados em conferências não é muito grande, no qual apresenta 24 (39,34%) artigos. Destaca-se que, o volume de trabalhos publicados em conferências pode ser atribuído ao processo de avaliação, como também ao escopo e as discussões técnicas.

Figura 8 – Fontes de publicação dos trabalhos.



3.3.3 Quais países têm mais pesquisadores publicando sobre detecção de fake news?

Mais de 50% dos autores dos estudos selecionados têm sua origem no continente asiático, 35 (57,4%), sendo a Índia o país com mais artigos, 20 (32,8%), seguido da China com 6

(9,8%). Destaca-se que, o continente americano é o segundo a apresentar um maior número de pesquisadores publicando no contexto de detecção de *fake news* com 13 (21,3%) estudos, sendo os Estados Unidos da América (EUA) o país com maior registro com 7 (11,5%) artigos, ficando atrás apenas da Índia no contexto mundial. Logo após os EUA temos o Brasil com 6 (9,8%) pesquisas. Vale ressaltar que, 80% dos trabalhos selecionados efetuam suas pesquisas apenas língua inglesa ou mesclam com outro idioma, nesta RS foi encontrado apenas um artigo, que foi a pesquisa publicada por Martín *et al.*,(2022), no qual apresenta um modelo multilíngue para detecção de *fake news* em inglês e espanhol. Com exceção destas duas situações, análise apenas em inglês e multilíngue, o outro idioma com maior registro de pesquisa é a língua portuguesa com apenas 5 (8,2%) estudos.

3.3.4 Quais tipos de conteúdo estão sendo considerados como solução na detecção de fake news?

Tendo em vista os diferentes conteúdos que vêm sendo utilizados no processo de detecção de *fake news*, o presente estudo buscou averiguar quais destes atualmente estão sendo amplamente aplicados. Na Figura 9, é possível ter uma visão das combinações de conteúdos considerados nos estudos.

No caso, o *Texto* é o principal fator para identificar uma notícia falsa, tanto no cenário nacional como internacional, com 4 (66,67%) e 36 (61,82%) estudos, respectivamente. Em contrapartida, a maioria dos estudos também mostram que considerando aspectos sociais é possível atingir bons resultados de detecção, sendo que, cerca de 37% dos estudos vêm realizando análise do conteúdo dos textos quanto ao *Contexto* e/ou *Sentimento*. Internacionalmente, essa pesquisa se faz mais presente, como a análise de *Texto + Contexto* em 10 (18,18%) estudos e *Texto + Sentimento* em 9 (9,09%) estudos. No cenário nacional, o estudo dos conteúdos apresenta poucos registros, com 1 (16,67%) estudo voltado para análise de *Texto + Sentimento* e 1 (16,67%) para *Texto + Sintaxe*.

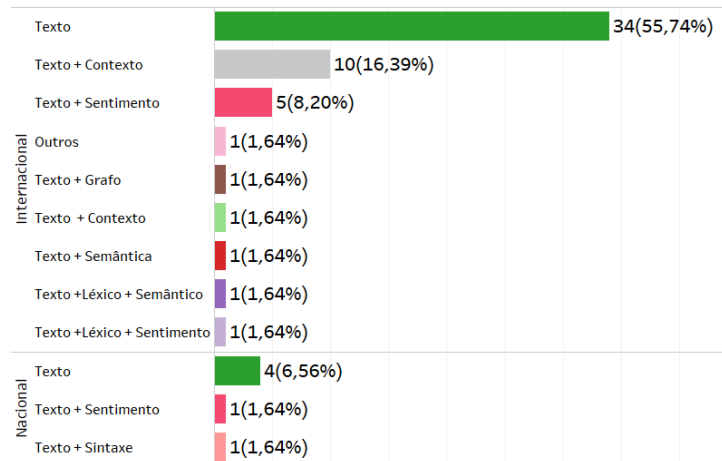
Vale ressaltar que, a análise de contexto é uma prática que envolve a compreensão e interpretação dos elementos circundantes a uma situação, texto, evento ou ambiente para uma melhor compreensão do assunto em questão. Já a análise de sentimentos, também conhecida como mineração de opinião ou mineração de emoções, é uma técnica utilizada para identificar, extrair e avaliar opiniões, sentimentos, emoções e atitudes expressas em texto. Destaca-se que, as análises de contexto e sentimento nos trabalhos desta RS foram parte do processo de análise.

3.3.5 Quais as redes sociais mais utilizadas no desenvolvimento dos trabalhos?

Como mostra a Figura 10, 27 (44,26%) direcionaram suas análises a detecção em redes sociais, sendo o *Twitter* a rede mais escolhida, com 25 (40,98%). Destaca-se que esse resultado já era esperado, tendo em vista que a plataforma é um ambiente que confortável para promover opiniões e notícias, sejam elas verdadeiras ou falsas.

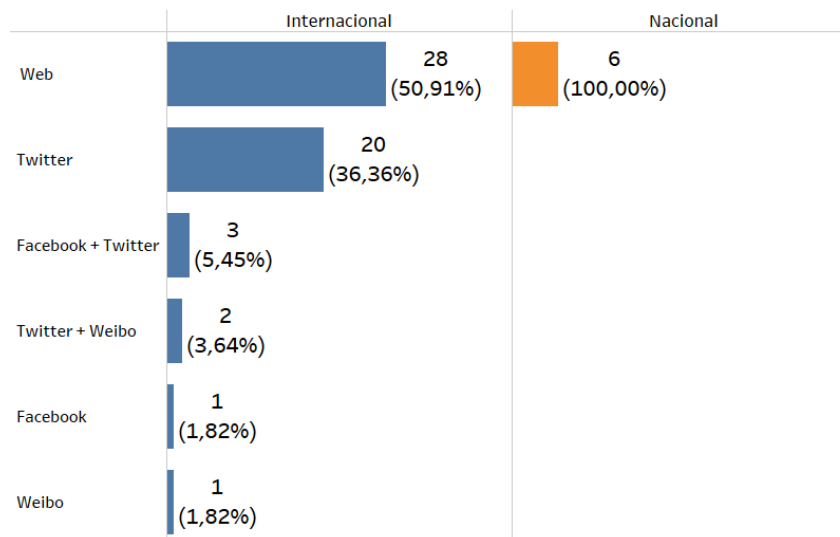
Dentre os 25 estudos, 20 (32,79%) trabalham somente com o *Twitter*, 3 (4,92%) combinam

Figura 9 – Conteúdos estudados.



o *Twitter* com o *Facebook* e 2 (3,28%) combinam o *Twitter* com o *Weibo*. Em contrapartida, 34 (55,74%) estudos vêm centralizando sua atenção para notícias publicadas em páginas jornalísticas. Nos estudos publicados nacionalmente, a predominância são em dados coletados da web, ou seja, de sites jornalísticos e revistas. Nos estudos publicados internacionalmente, 28 (50,91%) aderem também às notícias publicadas na web, como também de redes sociais como o *Twitter*, com 20 (36,36%).

Figura 10 – Redes sociais adotadas nos trabalhos.



3.3.6 Quais os datasets mencionados nos trabalhos selecionados?

Na Figura 11, é possível visualizar os *dataset* adotados pelos artigos selecionados, ao todo foram identificados 43 *dataset*. Vale destacar que mesmo tendo esses 43 *dataset*, 12 (12%) dos estudos selecionados criaram seus próprios *datasets*. Frisa-se que a construção do próprio *corpus* se dá tanto por meio de aplicações disponíveis pelas redes sociais, como também por meio de robôs de rastreamento, no caso, *web crawler*.

Mesmo a maior parte dos estudos criando seus próprios corpus, ainda há muitos estudos que optam por fazer uso de *datasets* já publicados, como as bases do *Kaggle* e *PolitiFact*. É importante frisar que, após a criação do próprio *dataset*, o *PolitiFact* é a base mais adotada com 10 (10%) estudos. Entretanto, em uma análise por cenários, é visível que nacionalmente há apenas duas bases de dados, sendo elas: a criação de uma própria com 3 (38%) estudo e a *Fake.br Corpus*, com 5 (63%) estudos. Destaca-se que muitos estudos fizeram uso de duas bases de dados, na qual uma é aplicada apenas para o treinamento e a outra para os testes, como no trabalho publicado por Almeida *et al.*, (2021).

Internacionalmente, a predominância é a base da *PolitiFact* com 10 (11%) registros e a criação de uma base própria com 9 (10%) estudos. Porém, diferente do cenário nacional, outras bases são adotadas, como a *Fake News detection - Kaggle* com 6 (7%) estudos.

Figura 11 – *Datasets* utilizados nos trabalhos.

		Internacional	Nacional
1	Corpus próprio	9 (14,29%)	3 (37,50%)
2	PolitiFacts	10 (15,87%)	
3	Fake News detec..	6 (9,52%)	
4	GossipCop	6 (9,52%)	
5	Fake.Br Corpus		5 (62,50%)
6	Fake or real new..	5 (7,94%)	
7	BuzzFeed	4 (6,35%)	
8	ISOT	4 (6,35%)	
9	Kaggle	4 (6,35%)	
10	LIAR	4 (6,35%)	
11	Twitter	4 (6,35%)	
12	Twitter15	2 (3,17%)	
13	Twitter16	2 (3,17%)	
14	Weibo	2 (3,17%)	
15	110-PT-BN-KP	1 (1,59%)	

3.3.7 Quais algoritmos de inteligência artificial são aplicados nas pesquisas?

Na Figura 12 é listado os algoritmos que foram adotados pelos artigos selecionados. Ao todo foram identificados 80 algoritmos, sendo que nestes estão incluídas combinações entre dois ou três algoritmos. É importante ressaltar que em diversos estudos os pesquisadores realizaram vários testes com 4 e/ou 5 algoritmos diferentes, no trabalho publicado por Khan *et al.*, (2022) por exemplo, 27 algoritmos foram testados.

Desta forma, o algoritmo de *Machine Learning* que apresentou maior número de registros foi o *Support Vector Machine* (SVM) com 39 (9,4%). Destaca-se que, o SVM predomina tanto nacionalmente como internacionalmente com respectivamente 7 (28%) e 32 (8,2%), em seguida tem-se *Naive Bayes* (NB) com 5 (30%) registros nacionalmente e 29 (7,4%) internacionalmente.

Algo que chama atenção é a adesão das redes neurais como, o *Long Short Term Memory* (LSTM) e *Convolutional Neural Network* (CNN) internacionalmente, que vem ganhando espaço na área, no qual o número de estudos são respectivamente 28 (7,2%) e 25 (6,4%). Todavia, nacionalmente esse registro é menor, no qual o CNN apresenta 1 (4%) registro como pode ser observado na Figura 12.

Figura 12 – Algoritmos aplicados nos trabalhos.

		Internacional	Nacional
1	SVM	32 (12,0%)	7 (35,0%)
2	NB	29 (10,9%)	5 (25,0%)
3	RF	29 (10,9%)	
4	LSTM	28 (10,5%)	
5	DT	27 (10,2%)	
6	CNN	25 (9,4%)	1 (5,0%)
7	LR	22 (8,3%)	2 (10,0%)
8	KNN	21 (7,9%)	1 (5,0%)
9	BERT	15 (5,6%)	
10	BiLSTM	9 (3,4%)	
11	SGDClassifier	7 (2,6%)	2 (10,0%)
12	AdaBoost	5 (1,9%)	1 (5,0%)
13	Bert + NER	6 (2,3%)	
14	Gaussian NB	5 (1,9%)	1 (5,0%)
15	HAN	6 (2,3%)	

3.3.8 Quais estudos apresentam resultados mais qualificados quanto às métricas de avaliação?

A presente QP8, foi criada tendo em vista identificar os resultados mais promissores alcançados pelas pesquisas. Vale ressaltar que, como mencionado, muitos artigos testam diversos algoritmos como também diversos *datasets*. Desta forma, a análise da Figura 13 apresenta nas linhas os *datasets* e os respectivos algoritmos testados, e na coluna são dispostos os trabalhos selecionados, sendo assim é feita a relação com o resultados alcançados. Entretanto, por conta do tamanho para este artigo a análise está filtrada por cenário, constando assim apenas os artigos nacionais.

Sendo assim é notório que, o melhor resultado para o *dataset* Fake.br Corpus foi a do trabalho de ID A45 (referente ao artigo de (SOUSA et al., 2022)), com a rede neural CNN, com 97,3% de acurácia. Destaca-se que com a criação do próprio corpus o melhor resultado foi do artigo de ID A38 (referente ao artigo de (ZHANG; HUANG; KHAN, 2022)), com 96,39% de acurácia, com a combinação dos algoritmos *GRU + CBOW + CNN*. É notório que o CNN apresenta melhores resultados nos dois *datasets* disponíveis para a língua portuguesa.

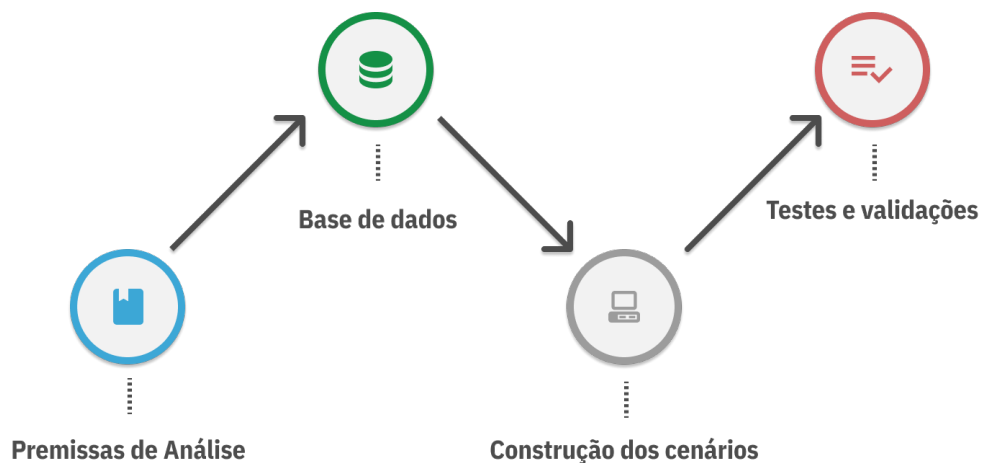
Figura 13 – Resultado dos algoritmos aplicado nos trabalhos selecionados.

		A2	A11	A29	A37	A38	A45
Fake.Br Corpus	AdaBoost		93,00%				
	CNN						97,30%
	NB	86,00%					
	SVM			80,00%	96,39%		
Corpus próprio	GRU + CBOW + CNN					96,39%	
	SGDClassifi..				92,90%		
	SVM			80,00%			

4 Proposta

Esta seção apresenta a metodologia aplicada para o desenvolvimento da proposta. Como mencionado na introdução especificamente, este trabalho propõe a aplicação de algoritmos de Aprendizado de Máquina e Redes Neurais Convolucionais para efetuar análises de padrões em textos de notícias em língua portuguesa, com o objetivo de realizar a detecção precisa de conteúdo enganoso. Tendo em vista os resultados alcançados é feita uma análise do processo de construção dos algoritmos, investigando o impacto das tomadas de decisão ao longo do processo metodológico no resultado final, proporcionando assim, encontrar a melhor combinação na literatura atual para a detecção de notícias falsas escritas em língua portuguesa. Para isto, foi desenvolvido uma metodologia dividida em etapas, ilustrado na Figura 14, sendo estas: I) Premissas de Análise; II) Base de dados; III) Construção dos cenários e IV) Testes e validações.

Figura 14 – Sequência de etapas metodológicas.



4.1 Premissas de Análise

Para a construção dos cenários foram levadas em consideração duas questões, sendo estas: I) Estrutura técnica e II) O que já há na literatura na área de automação de detecção de *fake news* em língua portuguesa. Desta forma, a partir da revisão sistemática foram selecionados quatro estudos e, destes, extraídas as decisões para a construção dos algoritmos, como pode ser visualizado na Tabela 3. Os estudos foram:

O estudo publicado por (SOUSA et al., 2022), que teve como objetivo apresentar uma análise de notícias em português e a detecção de *fake news*, utilizando Aprendizagem de Máquina (AM) e CNN. Destaca-se que o estudo concentrou-se em analisar tanto os textos, como também os metadados. Dentro dos processos para a construção do algoritmos foram utilizados como pré-processamento a remoção de *stopword* e a de remoção de caracteres estranhos, que são símbolos, letras ou caracteres especiais que não pertencem ao conjunto de caracteres normalmente esperado ou desejado, como *links* e pontuações. Em seguida, como transformador foi aplicado apenas o

Glove. É importante frisar que o treinamento foi realizado com quatro algoritmos, sendo estes: *Convolutional Neural Network* (CNN), *Naïve Bayes* (NB), *Support Vector Machine* (SVM) e *Multilayer Perceptron* (MLP). Desta forma, após uma análise sobre os algoritmos selecionados, obteve-se uma acurácia de 97%.

Tabela 3 – Trabalhos selecionados.

	Nível de análise	Pré-processamento	Transformação	Treinamento	ACC
Sousa et al., (2022)	Texto + Metadados	Remoção de stopword e caracteres estranhos	Glove	CNN, NB, SVM e MLP	97%
Rosa et al., (2019)	Texto	Remoção de stopword, caracteres estranhos e "Noise Corpus". Stemming.	Word Embedding não especificado	CNN e RNA	96%
Almeida et al., (2021)	Texto	Remoção de dados com conteúdo vazio, acentos e caracteres especiais, stopwords e números e pontuações	Bag of words TF - IDF	SVM e NB	80%
Souza et al., (2020)	Texto + Metadados	Classificação gramatical, polaridade e emoção. Extração de símbolos irrelevantes	Não especificado	NB, GB, AB, SVM e KNN	92,53%

O trabalho (ROSA et al., 2019), por sua vez, propõe uma solução para a detecção das *fake news*, utilizando 2 técnicas de AM, no caso, CNN e as Redes Neurais Artificiais de Múltiplas Camadas (RNA). No que diz respeito ao pré-processamento, o estudo fez uso da remoção de *stopword*, assim como remoção de caracteres estranhos, de "Noise Corpus", *stemming* e a frequência baixa. Para comprovar seus resultados o estudo exhibe uma análise quantitativa da acurácia, precisão, *recall*, *F1-Score*, bem como da matriz de confusão, na qual é possível observar que a CNN se sobressai a RNA, com a acurácia de 96% e 95%, respectivamente. Em contrapartida, o trabalho faz análise apenas de textos.

A pesquisa de (ALMEIDA et al., 2021) busca apresentar uma solução para identificar as notícias falsas brasileiras em um contexto político. Para isto, os autores propõem uma investigação quanto a qual algoritmo de aprendizado de máquina, entre SVM e NB, atinge o melhor resultado. Desta forma, foi adotada a base de dados Fake.br e uma própria, para o pré-processamento foi aplicado a remoção de dados com conteúdo vazio, assim como de acentos e caracteres especiais, tokenização, remoção de *stopwords* e a remoção de números e pontuações. É importante ressaltar que, no processo de transformação foram aplicados o *Bag of words*(BOW) e o *TF-IDF*. O melhor desempenho foi alcançado pela combinação de SVM (RBF) + BOW com 80% de acurácia.

O artigo publicado por (SOUZA et al., 2020a) propõe um método que visa realizar a classificação a nível de gramática e de sentimentos baseando-se na polaridade em textos escritos em português. Para este fim os autores utilizam também a base Fake.br Corpus, como pré-processamento é aplicado a classificação gramatical, extração de símbolos irrelevantes, classificação

de polaridade e a classificação de emoções. No que diz respeito ao treinamento foram testados os algoritmos NB, *Gradient Boost* (GB), *AdaBoost* (AB), SVM e *K-Nearest Neighbors* (KNN). Sendo que o algoritmo que apresentou o melhor desempenho foi o GB com 92,53% de acurácia.

É importante destacar que a escolha dos estudos descritos foi realizada no processo da RS desta dissertação. Os principais pontos para a seleção dos mesmos foi o idioma e a base de dados.

4.2 Base de dados

Para o desenvolvimento do presente trabalho foi adotado a base de dados *Fake.Br Corpus*. Essa base foi construída por pesquisadores da Universidade de São Paulo (USP) (MONTEIRO et al., 2018). Ela disponibiliza precisamente 7.200 artigos de notícias escritos em língua portuguesa, juntamente com seus respectivos metadados contendo 25 informações sobre os artigos, tais como:

1.Autor; 2.*Link*; 3.Categoria; 4.Data da publicação; 5.Número de *Tokens*; 6.Número de palavras sem pontuação; 7.Número de tipos; 8.Número de links dentro das notícias; 9.Número de palavras em maiúsculas; 10.Número de verbos; 11.Número de verbos no subjuntivo e no imperativo; 12.Número de substantivos; 13.Número de adjetivos; 14.Número de advérbios; 15.Número de verbos modais (principalmente verbos auxiliares); 16.Número de primeiro e segundo pronomes pessoais singulares; 17.Número de primeiros pronomes pessoais do plural; 18.Número de pronomes; 19.Pausa; 20.Número de caracteres; 21.Duração média da frase; 22.Comprimento médio da palavra; 23.Porcentagem de notícias com erros de ortografia; 24.Emotividade; 25.Diversidade.

Vale ressaltar que, dos 7.200 artigos, exatamente 3.600 notícias são consideradas verídicas e 3.600 são consideradas falsas. Destaca-se que, estes dados foram coletados dentro de um período de dois anos, de janeiro de 2016 a janeiro de 2018. O conteúdo das notícias da base *Fake.br* são divididos em 6 categorias como pode ser observado na Tabela 4, sendo estas: política, TV e celebridades, sociedade e notícias diárias, ciência e tecnologia, economia e religião. Destaca-se que, não existe nenhuma outra base de dados semelhante a esta disponível para a língua portuguesa, sendo esta a mais utilizada em pesquisas sobre *fake news* para este idioma.

Tabela 4 – Categorias da base *Fake.Br Corpus*.

CATEGORIA	AMOSTRA	%
Política	4.180	58,0%
TV e celebridades	1.544	21,4%
Sociedade e notícias diárias	1.276	17,7%
Ciência e tecnologia	112	1,5%
Economia	44	0,7%
Religião	44	0,7%

4.3 Construção dos cenários

Com base nos passos realizados pelos autores no processo de construção dos modelos de detecção de *fake news* e nos estudos dispostos na literatura, a presente pesquisa realiza a combinação entre estes processos focando em uma etapa por vez, buscando analisar o comportamento dos resultados. Ou seja, é tomado como foco a etapa de transformação e são analisadas as tomadas de decisão adotadas pelos autores acerca da construção dos algoritmos, o mesmo vale para o pré-processamento. A seguir é explicado as etapas estudadas e testadas no presente estudo.

4.3.1 Pré-processamento

O pré-processamento dos dados tem como objetivo padronizar os textos, no caso, remover *stopwords*, *tokenizar* os textos e eliminar caracteres irreconhecíveis pelos algoritmos. A etapa de pré-processamento dos textos segue os passos descrito a seguir:

- **Remoção de ruídos e caracteres estranhos:** Remoção de caracteres especiais ou em formato irreconhecível pelos algoritmos, assim como aqueles que não contribuem para o seu treinamento. Neste sentido, numerais, *urls*, caracteres de pontuação e outros caracteres especiais foram eliminados.
- **Remoção de *stopwords*:** Palavras consideradas comuns e, portanto, irrelevantes para a representação de um texto. São normalmente pertencentes às classes gramaticais de pronomes, preposições, artigos, advérbios e conjunções.
- **Tokenização:** também conhecida como segmentação de palavras, a *tokenização* consiste na quebra na sequência de caracteres do texto localizando o limite de cada palavra. As partes resultantes deste processo são denominadas de *tokens*.

4.3.2 Transformação

Com o objetivo de facilitar o processamento dos textos pelos algoritmos de AM, é necessário que os textos dos artigos sejam convertidos do formato textual para uma representação numérica. Para isso os textos são submetidos a um processo de vetorização, que consiste em converter cada termo contido no texto em um valor numérico. Desta forma, cada texto é representado por um vetor contendo todos os valores numéricos que representam cada um dos termos contidos nele. Com os textos vetorizados, o processo segue em utilizar a técnica denominada *word embedding*, que converte os textos em vetores multidimensionais, de forma a manter os seus valores tanto sintáticos quanto morfológicos.

4.3.3 Treinamento

Após a etapa de transformação, inicia-se a etapa de treinamento dos algoritmos. Etapa esta em que os algoritmos aprendem a partir dos dados de treinamento e ajustam seus parâmetros para realizar tarefas específicas, como classificação ou regressão. Desta forma, para o processamento

dos textos, foi implementada uma CNN e SVM para os textos e para os metadados foram aplicados SVM, NB e MLP. Haja vista que, ambos são comumente aplicados nos trabalhos relatados.

O CNN é um tipo especializado de rede neural amplamente utilizada em tarefas de visão computacional. No entanto, também têm sido aplicadas no processamento de texto. Sendo esta composta por camadas de convolução, que são capazes de extrair características relevantes de forma automática a partir dos dados. O SVM por sua vez, são algoritmos de aprendizado supervisionado que busca encontrar o hiperplano ótimo que melhor separa as classes no espaço de características. Essa abordagem é particularmente útil quando os dados são linearmente separáveis, mas também pode ser estendida para casos não lineares. O NB usa o teorema de Bayes para calcular a probabilidade de que uma determinada amostra pertença a uma classe específica, dadas as características dessa amostra. Já a MLP, é composta por camadas de neurônios conectados, incluindo camadas ocultas que aprendem representações complexas dos dados, sendo capaz de aprender relações não lineares.

4.3.4 Teste

Para avaliar o desempenho dos modelos treinados, foram selecionadas métricas específicas de avaliação, levando em consideração a natureza do conjunto de dados em análise. Entre as métricas escolhidas, destacam-se a acurácia e a matriz de confusão, haja vista que as mesmas são as mais recorrentes nos estudos selecionados nas premissas de análise e considerando o tamanho da base de dados que está sendo trabalhada. Essa seleção criteriosa foi influenciada pelo tamanho da base de dados em questão, bem como pela necessidade de uma avaliação robusta e abrangente. A acurácia oferece uma medida geral da precisão do modelo, enquanto a matriz de confusão fornece informações detalhadas sobre como o modelo classifica as diferentes classes, incluindo falsos positivos, falsos negativos, verdadeiros positivos e verdadeiros negativos. Essas métricas são fundamentais para avaliar o desempenho do modelo em um conjunto de dados de tamanho considerável.

5 Resultados

Na presente seção são discutidos os resultados alcançados nos testes realizados. Tendo em vista avaliar o desempenho dos modelos treinados, foram selecionadas as métricas de avaliação da acurácia e matriz de confusão. Estes critérios foram adotados considerando o tamanho da base de dados que está sendo trabalhada e a recorrência da aplicação destas métricas de avaliação entre os estudos selecionados nas premissas de análise.

Como mencionado, os algoritmos utilizados foram o CNN e SVM para textos, haja vista que o objetivo consiste em analisar as decisões na criação do algoritmo. É importante relatar que, antes de inicializar o desenvolvimento da proposta, foi realizado um estudo na base de dados. Como supramencionado, este artigo faz uso tanto de textos como de metadados.

No que diz respeito especificamente aos metadados, foi necessário selecionar as informações com maior relevância para o estudo em questão. Para isso, utilizou-se a técnica de *Backward Feature Elimination*, que consiste em iterativamente remover uma coluna por vez do conjunto completo de metadados, reavaliando o modelo a cada etapa. As colunas removidas são aquelas que não causam uma piora significativa no desempenho do modelo ou não contribuem de forma relevante para a análise, resultando em um subconjunto mais conciso e informativo de metadados para a pesquisa.

Inicialmente, com base na análise realizada na revisão sistemática, destacou-se que a adoção de modelos de aprendizado de máquina, como SVM, LSTM e CNN, foi mais pronunciada. Portanto, na primeira fase do desenvolvimento do modelo de detecção de notícias falsas, optou-se por utilizar a abordagem CNN para o processamento dos textos. Quanto ao processamento dos metadados, foram realizados estudos comparativos com três modelos: MLP, SVM e NB, com o objetivo de selecionar o método mais adequado aos dados.

5.1 Análise dos Textos

A princípio, é importante destacar que, para conduzir os testes e avaliações, a base de dados foi dividida em duas partes: 80% para treinamento e validação do modelo e 20% para os testes do desempenho. Como pode ser observado na Figura 15(a), o CNN alcançou uma acurácia de 97,3% na base de treinamento, 96,3% nos dados de validação e 96,5% nos testes. Além disso, a Figura 15(b) ilustra que a perda na base de treinamento foi de 7,4%, enquanto nos dados de validação foi de 11,2% e nos testes foi de 10%. Esses resultados indicam a eficácia do modelo. É importante destacar que, durante as fases de treinamento e teste, observou-se que o processo de aprendizado alcançou uma convergência rápida, com o algoritmo atingindo resultados satisfatórios em apenas três épocas de treinamento. Vale destacar que o critério de parada foi estabelecido empiricamente, observando o ponto em que ocorria o *overfitting*.

Desta forma, os resultados permitem inferir que o modelo CNN é promissor para o

processamento de textos e detecção de notícias falsas, devido à sua acurácia e rápido processo de aprendizado, mas vale a pena notar que a perda do modelo é maior nos dados de validação em comparação com a base de treinamento e os testes. Isso pode indicar uma ligeira tendência de *overfitting*, onde o modelo se adapta demais aos dados de treinamento.

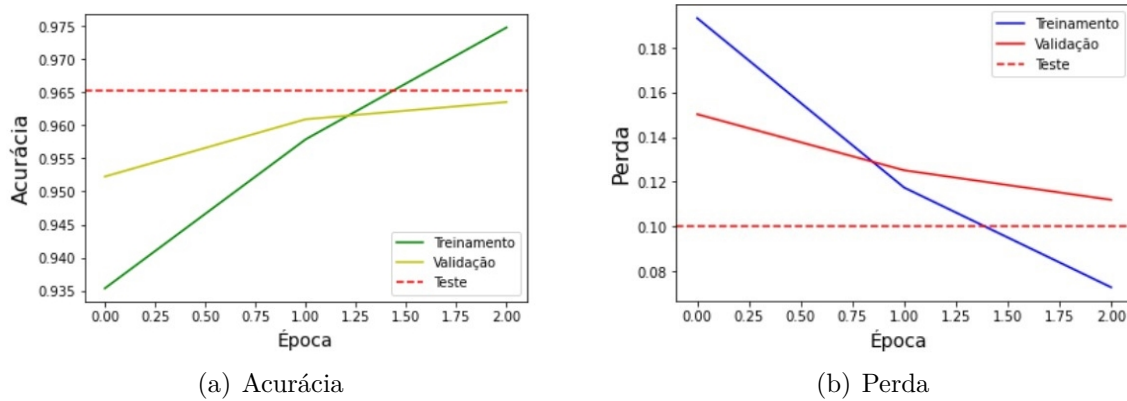


Figura 15 – Resultados da CNN.

Observando os resultados na matriz de confusão, conforme ilustrado na Figura 16, é perceptível que o algoritmo exibe uma tendência a cometer mais erros de falsos negativos, ou seja, classificar notícias verdadeiras como falsas. Buscando identificar a origem desse erro de classificação, foi observado que o comprimento dos textos desempenha um papel significativo, haja vista que o algoritmo demonstrou uma maior dificuldade em aprender com textos mais curtos, que contêm menos termos em sua composição. Essa dificuldade pode ser atribuída ao fato de que o conjunto de dados *Fake.br* frequentemente rotula como *fake news* os artigos com estruturas mais compactas, enquanto artigos com estruturas mais extensas tendem a ser classificados como verdadeiros. Salienta-se que, em uma análise quanto ao número de palavras de notícias falsas e verdadeiras do Fake.br Corpus a média, mediana e desvio padrão de palavras em uma notícias falsas é respectivamente 185.5, 157 e 127.8 palavras, já em uma verdadeira é de 1106.5, 922 e 686, como pode ser observado na Tabela 5.

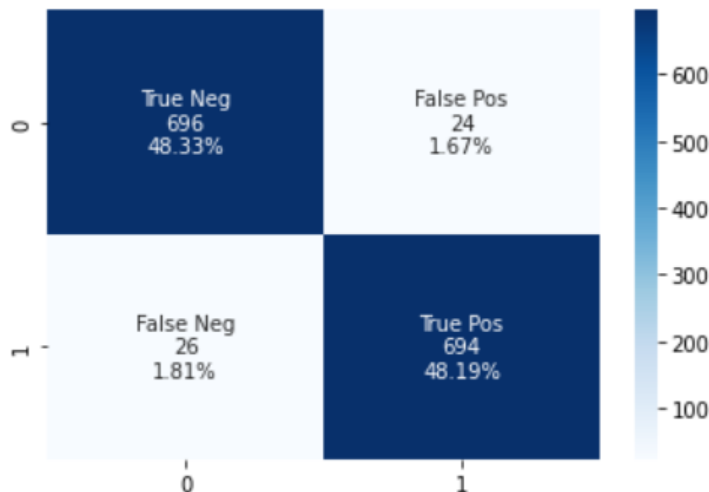


Figura 16 – Matriz de confusão de teste do algoritmo CNN.

Tabela 5 – Estatísticas descritivas referentes ao número de palavras das notícias.

	Média	Mediana	Desvio Padrão
Fake News	185.8	157.0	127.8
Verdadeiras	1106.5	922.0	686.0

5.2 Análise dos metadados

Nos testes no processamento dos metadados, o SVM apresentou melhor desempenho com uma acurácia de 97,5% e precisão de 98,3%. A performance deste e dos demais algoritmos, assim como as métricas de avaliação, podem ser conferidas na Tabela 6.

Tabela 6 – Resultados dos Metadados.

Algoritmo	Acurácia	Recall	Precisão	F1-Score
SVM	97.5%	96.6%	98.3%	97.4%
MLP	95.8%	94.3%	97.2%	95.7%
NB	96.4%	97.2%	95.7%	96.4%

Ao analisar as matrizes de confusão, apresentadas nas Figuras 17(a), 17(b) e 17(c), podemos observar que, assim como ocorreu com o CNN, os erros se concentram principalmente nos falsos negativos. Notavelmente, o algoritmo NB demonstrou o menor erro de classificação nesse aspecto entre os três algoritmos avaliados, enquanto o MLP exibiu uma taxa de erro de classificação mais elevada. É importante ressaltar que, de forma geral, o SVM se destacou como o algoritmo com o melhor desempenho, enquanto o MLP mostrou um desempenho inferior em comparação aos outros algoritmos testados, tanto nas matrizes de confusão quanto na tabela de métricas de avaliação.

5.3 Avaliação de Etapas do Algoritmo

Os primeiros testes tiveram como foco o processo de transformação, sendo assim, tanto a base de dados, como o nível de análise e o pré-processamento se mantiveram os mesmos para todas as combinações. Neste ponto é importante ressaltar que não foi possível replicar as combinações dos autores (ROSA et al., 2019) e (SOUZA et al., 2020a) uma vez que, o primeiro utilizou um *embedding* o próprio o qual não está disponível, enquanto o segundo não especificou qual foi aplicado.

Portanto, diante das combinações realizadas, como pode ser visualizada na Tabela 7, foi notório que a diferença entre os resultados é mínima, já que todos alcançaram uma média de 97% de acurácia, porém dentre as combinações o que mais se sobressaiu foi o transformador *Word2Vec* com 97,7%. De posse dos resultados constata-se que o nível de influência do transformador no algoritmo não é tão forte quanto outros fatores.

Os demais testes tiveram como foco o pré-processamento, porém é consensual que os processos de remoção de caracteres estranhos e tokenização são etapas fundamentais para a preparação dos textos para a análise. Sendo assim, a fim de analisar o impacto do pré-

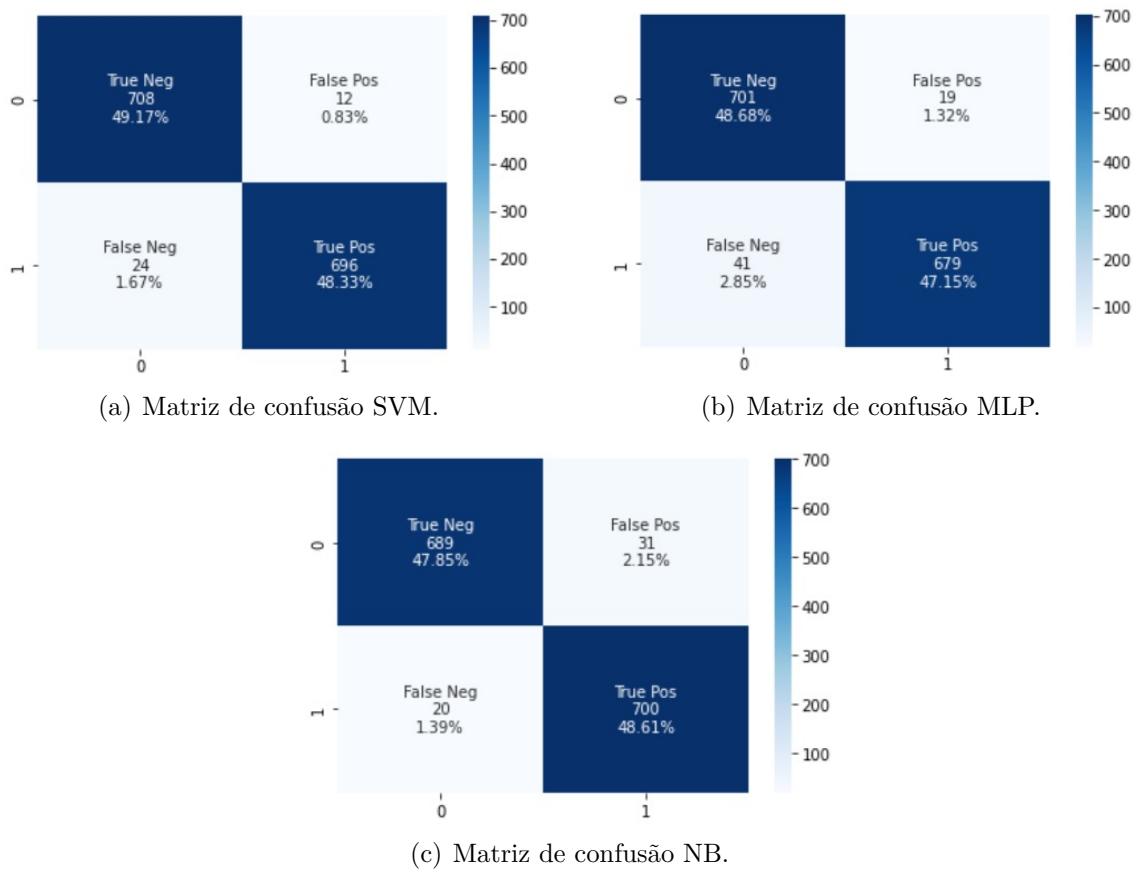


Figura 17 – Matrizes de confusão do processamento dos metadados.

Tabela 7 – Resultados do primeiro experimento.

	Transformação	Treinamento	Acurácia
Teste 1	Glove	CNN	97,0%
Teste 3	Wang2Vec	SVM	96,4%
Teste 4	Word2Vec	SVM	97,7%
Teste 5	TF-IDF	SVM	96,8%
Teste 6	FASTEXT	SVM	96,6%

processamento foram realizados testes comparativos a respeito da remoção ou manutenção de *stopwords*.

Como pode ser visualizado na Tabela 8 sem a remoção de *stopwords* foi o cenário que apresentou o melhor resultado com 97% de acurácia, aplicando o TF-IDF. Frisa-se que os demais resultados foram semelhantes, sendo que a combinação do Glove sem a remoção de *stopwords* apresentou menor acurácia. É importante frisar que o escolha do Glove para o CNN se deu pelo mesmo apresentar melhor adequação aos dados. Tendo em vista os resultados obtidos constatou-se que, especificamente para o cenário sem a remoção de *stopwords* para o TF-IDF, a remoção de *stopwords* pode levar a perda de informações contextuais, já que elas fornecem estrutura gramatical e conectividade entre as demais palavras. Isso pode resultar em uma representação menos rica dos documentos, levando a uma perda de informações relevantes.

No que diz respeito aos metadados os resultados foram bem próximos, porém o SVM

Tabela 8 – Resultados do segundo experimento.

	Pré - Processamento	Transformação	Treinamento	Acurácia
Teste 1	Com Stopwords	Glove	CNN	96,0%
Teste 2	Sem Stopword	Glove	CNN	96,3%
Teste 3	Com Stopword	TF-IDF	SVM	97,0%
Teste 4	Sem Stopword	TF-IDF	SVM	96,8%
Teste 5	Com Stopword	BOW	SVM	96,9%
Teste 6	Sem Stopword	BOW	SVM	96,8%

apresentou uma melhor adequação aos dados com 97,6% de acurácia, seguido do MLP com 97,4% e por fim o NB com 95,7%.

6 Considerações Finais

Conforme apresentado e discutido, esta dissertação propõe uma avaliação de desempenho das etapas do processo de extração de conhecimento para detecção de *fake news* em português. Realizando uma análise do processo de construção dos algoritmos, investigando o impacto das tomadas de decisão no processo metodológico no resultado final, proporcionando assim, encontrar a melhor combinação na literatura atual para a detecção de notícias falsas escritas em português. Para chegar a este objetivo uma série de estudos foi realizada, como discutido. No qual, foi possível identificar características de notícias falsas e verdadeiras, assim como a carência de estudos na área de detecção de *fake news* em português.

Analisando as combinações, foi perceptível que os maiores impactos nos resultados foram verificados quando o foco se encontrou no pré-processamento, mais especificamente com relação às *stopwords*. No caso sem a remoção de *stopwords* foi o cenário que apresentou o melhor resultado com 97% de acurácia, aplicando o *TF-IDF*. Em contrapartida, a combinação do *Glove* sem a remoção de *stopwords* apresentou menor acurácia.

No que diz respeito a revisão sistemática realizada, a presente dissertação buscou apresentar os trabalhos mais relevantes no contexto de detecção de *fake news* e suas respectivas soluções. Foi possível concluir que:

- Nos últimos três anos, o número de artigos científicos tem passado por um crescimento no número de publicações, especialmente em 2022.
- As publicações científicas mais recorrentes no cenário internacional são em jornais, já no cenário nacional são em conferências.
- A RS identificou a Índia como o país com o maior número de publicações e o Inglês como o idioma mais adotado nos trabalhos.
- Em relação aos tipos de conteúdo, o *texto* é considerado o principal, porém novas combinações vêm sendo apresentadas como solução, tais como análise de *texto + contexto* e *texto + sentimento*.
- Mostrou-se também que mais de 55% dos estudos centralizaram a atenção para notícias publicadas em páginas jornalísticas. Em seguida, tem-se o *Twitter*, a rede social mais adotada para estudos na área de detecção de *fake news*.
- Em paralelo as redes sociais, os pesquisadores têm buscado adotar base de dados próprias, porém, ainda há um número considerável que faz uso de bases consolidadas, sendo o *Kaggle* e o *PolitiFact* as mais aderidas internacionalmente e a *Fake.br Corpus* nacionalmente.
- O algoritmo de ML que apresentou maior número de registros foi o SVM com 39 (9,4%). Internacionalmente, chamou atenção a adesão das redes neurais como o LSTM e o CNN. Nacionalmente, essa adesão é menor.

6.1 Trabalhos Futuros

Como trabalhos futuros, pretende-se expandir as combinações, no caso, aplicar técnicas utilizadas em estudos voltados para a língua inglesa e ver o comportamento no português. Como também aplicar algoritmos de *Generative Pre-trained Transformer* (GPT). Destaca-se que pretende-se refazer os testes em outras bases de dados em português já consolidadas.

Referências

- ABDELMINAAM, D. S. et al. Coaid-deep: an optimized intelligent framework for automated detecting covid-19 misleading information on twitter. *Ieee Access*, IEEE, v. 9, p. 27840–27867, 2021. Citado na página 41.
- AGRAWAL, C.; PANDEY, A.; GOYAL, S. Fake news detection system based on modified bi-directional long short term memory. *Multimedia Tools and Applications*, v. 81, 2022. Citado na página 41.
- AHUJA, N.; KUMAR, S. *S-HAN: Hierarchical Attention Networks with Stacked Gated Recurrent Unit for Fake News Detection*. 2020. 873–877 p. Citado na página 41.
- AJAO, O.; BHOWMIK, D.; ZARGARI, S. *Fake news identification on twitter with hybrid cnn and rnn models*. 2018. 226–230 p. Citado na página 41.
- ALLCOTT, H.; GENTZKOW, M. Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, v. 31, p. 211–236, 2017. Citado 3 vezes nas páginas 21, 25 e 26.
- ALMEIDA, L. et al. Identificação de “fake news” no contexto político brasileiro: uma abordagem computacional. In: *Anais do II Workshop sobre as Implicações da Computação na Sociedade*. Porto Alegre, RS, Brasil: SBC, 2021. p. 78–89. ISSN 2763-8707. Disponível em: <<https://sol.sbc.org.br/index.php/wics/article/view/15966>>. Citado 2 vezes nas páginas 41 e 48.
- AMAZON. *Introduction to Amazon SageMaker Object2Vec*. 2018. <https://aws.amazon.com/pt/blogs/machine-learning/introduction-to-amazon-sagemaker-object2vec/>. Acesso em: 03 Maio 2023. Citado 2 vezes nas páginas 15 e 30.
- ARO, J. The cyberspace war: Propaganda and trolling as warfare tools. *European View*, v. 15, n. 1, p. 121–132, 2016. Disponível em: <<https://doi.org/10.1007/s12290-016-0395-5>>. Citado na página 21.
- AYOUB, J.; YANG, X. J.; ZHOU, F. Combat covid-19 infodemic using explainable natural language processing models. *Information Processing & Management*, Elsevier, v. 58, n. 4, p. 102569, 2021. Citado na página 41.
- BAHAD, P.; SAXENA, P.; KAMAL, R. Fake news detection using bi-directional lstm-recurrent neural network. *Procedia Computer Science*, Elsevier, v. 165, p. 74–82, 2019. Citado na página 41.
- BARBOSA, J. et al. Introdução ao processamento de linguagem natural usando python. *III Escola Regional de Informatica do Piauí*, v. 1, p. 336–360, 2017. Citado na página 27.
- BARBOSA, V. N. et al. A comparative study of machine learning algorithms for the detection of fake news on the internet. *XVIII Brazilian Symposium on Information Systems*, Association for Computing Machinery, New York, NY, USA, 2022. Citado na página 41.
- CARVALHO, C.; KLAGGE, N.; MOENCH, E. The persistent effects of a false news shock. *Journal of Empirical Finance*, v. 18, n. 4, p. 597–615, 2011. ISSN 0927-5398. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S092753981100020X>>. Citado na página 21.

- CERRAO, N. G.; CASTRO, F. F. d.; JESUS, A. F. d. O método de revisão sistemática da literatura (rs) na área da ciência da informação no brasil: análise de dados de pesquisa. *Informação & Tecnologia*, v. 5, n. 1, p. 105–116, 2018. Citado na página 37.
- CHAUHAN, T.; PALIVELA, H. Optimization and improvement of fake news detection using deep learning approaches for societal benefit. *International Journal of Information Management Data Insights*, Elsevier, v. 1, n. 2, p. 100051, 2021. Citado na página 41.
- CHEN, X. et al. Catch me if you can: A participant-level rumor detection framework via fine-grained user representation learning. *Information Processing & Management*, Elsevier, v. 58, n. 5, p. 102678, 2021. Citado na página 41.
- CHEN, Y.; CONROY, N.; RUBIN, V. *Misleading Online Content: Recognizing Clickbait as "False News"*. 2015. Citado na página 26.
- CHOUDRIE, J. et al. Machine learning techniques and older adults processing of online information and misinformation: A covid 19 study. *Computers in Human Behavior*, Elsevier, v. 119, p. 106716, 2021. Citado na página 41.
- DATAREPORTAL. *DIGITAL 2023: GLOBAL OVERVIEW REPORT*. 2023. Disponível em: <<https://datareportal.com/reports/digital-2023-global-overview-report>>. Acesso em: 02 Maio 2023. Citado na página 21.
- DEEPAK, S.; CHITTURI, B. Deep neural approach to fake-news identification. *Procedia Computer Science*, Elsevier, v. 167, p. 2236–2243, 2020. Citado na página 41.
- DONG, X.; VICTOR, U.; QIAN, L. Two-path deep semisupervised learning for timely fake news detection. *IEEE Transactions on Computational Social Systems*, IEEE, v. 7, n. 6, p. 1386–1398, 2020. Citado na página 41.
- FAYAZ, M. et al. Machine learning fake news classification with optimal feature selection. *Soft Comput*, 2021. Citado na página 41.
- FULLER, C. M.; BIROS, D. P.; WILSON, R. L. Decision support for determining veracity via linguistic-based cues. *Decision Support Systems*, v. 46, n. 3, p. 695–703, 2009. ISSN 0167-9236. Wireless in the Healthcare. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S0167923608001991>>. Citado na página 21.
- GADEK, G.; GUÉLORGET, P. An interpretable model to measure fakeness and emotion in news. *Procedia Computer Science*, Elsevier, v. 176, p. 78–87, 2020. Citado na página 41.
- GEREME, F. B.; ZHU, W. *Early detection of fake news "before it flies high"*. 2019. 142–148 p. Citado na página 41.
- GOLBECK, J. et al. *Fake News vs Satire: A Dataset and Analysis*. 2018. 17-21 p. Citado na página 26.
- GONZALEZ, M.; LIMA, V. L. S. d. Recuperação de informação e processamento da linguagem natural. *XXIII Congresso da Sociedade Brasileira de Computação*, 2003. Citado na página 22.
- HAN, S. et al. D-fend: A diffusion-based fake news detection framework for news articles related to covid-19. *ACM/SIGAPP Symposium on Applied Computing*, Association for Computing Machinery, New York, NY, USA, 2022. Citado na página 41.
- HARRAG, F.; DJAHILI, M. K. Arabic fake news detection: A fact checking based deep learning approach. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, Association for Computing Machinery, New York, NY, USA, v. 21, 2022. ISSN 2375-4699. Citado na página 41.

- HOTI, A. H. et al. Identifying fake news written on albanian language in social media using naïve bayes, svm, logistic regression, decision tree and random forest algorithms. *Mediterranean Conference on Embedded Computing (MECO)*, 2022. Citado na página 41.
- JAIN, P. et al. Classifying fake news detection using svm, naive bayes and lstm. *International Conference on Cloud Computing, Data Science & Engineering*, 2022. Citado na página 41.
- JOACHIMS, T. A probabilistic analysis of the rocchio algorithm with tfidf for text categorization. In: *Proceedings of the Fourteenth International Conference on Machine Learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1997. (ICML '97), p. 143–151. ISBN 1558604863. Citado na página 29.
- KALIYAR, R. K. *Fake news detection using a deep neural network*. 2018. 7 p. Citado na página 41.
- KALIYAR, R. K.; GOSWAMI, A.; NARANG, P. Fakebert: Fake news detection in social media with a bert-based deep learning approach. *Multimedia tools and applications*, Springer, v. 80, n. 8, p. 11765–11788, 2021. Citado na página 41.
- KALIYAR, R. K. et al. Fndnet—a deep convolutional neural network for fake news detection. *Cognitive Systems Research*, Elsevier, v. 61, p. 32–44, 2020. Citado na página 41.
- KALIYAR, R. K. et al. *DeepNet: an efficient neural network for fake news detection using news-user engagements*. 2020. 6 p. Citado na página 41.
- KEELE, S. et al. *Guidelines for performing systematic literature reviews in software engineering*. [S.l.], 2007. Citado 2 vezes nas páginas 37 e 38.
- KHAN, J. Y. et al. A benchmark study of machine learning models for online fake news detection. *Machine Learning with Applications*, Elsevier BV, 2021. Citado na página 41.
- KHAN, Y.; THAKUR, S. Fake news detection of south african covid-19 related tweets using machine learning. *International Conference on Artificial Intelligence, Big Data, Computing and Data Communication Systems (icABCD)*, 2022. Citado na página 41.
- KOGAN, S.; MOSKOWITZ, T. J.; NIESSNER, M. Fake news: Evidence from financial markets. *Available at SSRN*, v. 3237763, 2019. Citado na página 21.
- KRISHNA, N. L. S. R.; ADIMOOLAM, M. Fake news detection system using decision tree algorithm and compare textual property with support vector machine algorithm. *International Conference on Business Analytics for Technology and Security (ICBATS)*, 2022. Citado na página 41.
- KUMAR, S. et al. *A Survey on Analysis of Fake News Detection Techniques*. 2021. 894-899 p. Citado 2 vezes nas páginas 21 e 26.
- LAN, T. et al. An effective approach for rumor detection based on ernie-bigru-attention. *International Conference on Consumer Electronics and Computer Engineering (ICCECE)*, 2022. Citado na página 41.
- LECUN, Y.; BENGIO, Y. Convolutional networks for images, speech, and time series. In: _____. *The Handbook of Brain Theory and Neural Networks*. Cambridge, MA, USA: MIT Press, 1998. p. 255–258. ISBN 0262511029. Citado na página 32.
- LEVY, O.; GOLDBERG, Y. Dependency-based word embeddings. In: *52nd Annual Meeting of the Association for Computational Linguistics*. [S.l.: s.n.], 2014. p. 302–308. Citado na página 33.

- LI, X. et al. *A novel self-learning semi-supervised deep learning network to detect fake news on social media*. [S.l.]: Springer, 2021. 9 p. Citado na página 41.
- MAHMUD, F. B. et al. A comparative analysis of graph neural networks and commonly used machine learning algorithms on fake news detection. *2022 7th International Conference on Data Science and Machine Learning Applications (CDMA)*, 2022. Citado na página 41.
- MALHOTRA, R.; MAHUR, A.; ACHINT. Covid-19 fake news detection system. *2022 12th International Conference on Cloud Computing, Data Science & Engineering*, 2022. Citado na página 41.
- MARTÍN, A. et al. Facter-check: Semi-automated fact-checking through semantic similarity and natural language inference. *Knowledge-Based Systems*, v. 251, 2022. ISSN 0950-7051. Citado na página 41.
- MATHEWS, E. Z.; PREETHI, N. Fake news detection: An effective content-based approach using machine learning techniques. *2022 International Conference on Computer Communication and Informatics (ICCCI)*, 2022. Citado na página 41.
- MATSUGU, M. et al. Subject independent facial expression recognition with robust face detection using a convolutional neural network. *Neural Networks*, v. 16, p. 555–559, 2003. Citado na página 32.
- MEDEIROS, F. D.; BRAGA, R. B. Fake news detection in social media: A systematic review. In: *XVI Brazilian Symposium on Information Systems*. New York, NY, USA: Association for Computing Machinery, 2020. (SBSI'20). ISBN 9781450388733. Disponível em: <<https://doi.org/10.1145/3411564.3411648>>. Citado na página 22.
- MEESAD, P. Thai fake news detection based on information retrieval, natural language processing and machine learning. *SN Computer Science*, Springer, v. 2, n. 6, p. 1–17, 2021. Citado na página 41.
- MIKOLOV, T. et al. Distributed representations of words and phrases and their compositionality. In: *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*. Red Hook, NY, USA: Curran Associates Inc., 2013. (NIPS'13), p. 3111–3119. Citado na página 30.
- MONTEIRO, R. A. et al. Contributions to the study of fake news in portuguese: New corpus and automatic detection results. In: *Computational Processing of the Portuguese Language*. [S.l.]: Springer International Publishing, 2018. p. 324–334. ISBN 978-3-319-99722-3. Citado na página 49.
- MORAES, M. P.; SAMPAIO, J. de O.; CHARLES, A. C. *Data mining applied in fake news classification through textual patterns*. 2019. 321–324 p. Citado na página 41.
- NASIR, J. A.; KHAN, O. S.; VARLAMIS, I. Fake news detection: A hybrid cnn-rnn based deep learning approach. *International Journal of Information Management Data Insights*, Elsevier, v. 1, n. 1, p. 100007, 2021. Citado na página 41.
- NAYOGA, B. P. et al. Hoax analyzer for indonesian news using deep learning models. *Procedia Computer Science*, Elsevier, v. 179, p. 704–712, 2021. Citado na página 41.
- OSHIKAWA, R.; QIAN, J.; WANG, W. A survey on natural language processing for fake news detection. 11 2018. Citado na página 22.

- PAKA, W. S. et al. Cross-sean: A cross-stitch semi-supervised neural attention model for covid-19 fake news detection. *Applied Soft Computing*, Elsevier, v. 107, p. 107393, 2021. Citado na página 41.
- PALANI, B.; ELANGO, S.; K, V. V. Cb-fake: A multimodal deep learning framework for automatic fake news detection using capsule neural network and bert. *Multimedia Tools Appl.*, Kluwer Academic Publishers, USA, v. 81, 2022. ISSN 1380-7501. Citado na página 41.
- PENNINGTON, J.; SOCHER, R.; MANNING, C. D. Glove: Global vectors for word representation. In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. [S.l.: s.n.], 2014. p. 1532–1543. Citado na página 32.
- QAWASMEH, E.; TAWALBEH, M.; ABDULLAH, M. *Automatic identification of fake news using deep learning*. 2019. 383–388 p. Citado na página 41.
- RANJAN, V.; AGRAWAL, P. Fake news detection: Ga-transformer and ig-transformer based approach. *2022 12th International Conference on Cloud Computing, Data Science & Engineering*, 2022. Citado na página 41.
- RATHOD, S. Exploring author profiling for fake news detection. 2022. Citado na página 41.
- RECUERO, R.; GRUZD, A. *Cascatas de Fake News Políticas: um estudo de caso no Twitter*. scielo, 2019. 31 - 47 p. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1982-25532019000200031&nrm=iso>. Citado na página 26.
- REUTERS, I.; OXFORD, U. of. *Reuters Institute Digital News Report 2021*. 2021. Disponível em: <https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2021-06/Digital_News_Report_2021_FINAL.pdf>. Acesso em: 18 Maio 2022. Citado na página 21.
- RONG, X. *word2vec Parameter Learning Explained*. 2016. Citado 2 vezes nas páginas 15 e 31.
- ROSA, M. d. A. et al. Detecção de fake news: uma abordagem utilizando redes neurais convolucionais. 2019. Citado 2 vezes nas páginas 48 e 55.
- RUCHANSKY, N.; SEO, S.; LIU, Y. *Csi: A hybrid deep model for fake news detection*. 2017. 797–806 p. Citado na página 41.
- RUSSELL, S.; NORVIG, P. *Artificial Intelligence*. 3rd. ed. [S.l.]: Prentice Hall, 2010. Citado na página 22.
- SADEGHI, F.; BIDGOLY, A. J.; AMIRKHANI, H. Fake news detection on social media using a natural language inference approach. *Multimedia Tools Appl.*, Kluwer Academic Publishers, USA, v. 81, n. 23, 2022. ISSN 1380-7501. Citado na página 41.
- SASTRAWAN, I. K.; BAYUPATI, I.; ARSA, D. M. S. *Detection of fake news using deep learning CNN-RNN based methods*. [S.l.]: Elsevier, 2021. Citado na página 41.
- SAWCZYN, A. et al. *Fact-checking: relevance assessment of references in the Polish political domain*. [S.l.]: Elsevier, 2021. 1285–1293 p. Citado na página 41.
- SEPÚLVEDA-TORRES, R. et al. Headlinestancechecker: Exploiting summarization to detect headline disinformation. *Journal of Web Semantics*, Elsevier, v. 71, p. 100660, 2021. Citado na página 41.
- SHAILENDRA, P. et al. Fake news detection in hindi using embedding techniques. *2022 IEEE Region 10 Symposium (TENSYP)*, 2022. Citado na página 41.

- SHARMA, K. et al. Combating fake news: A survey on identification and mitigation techniques. *ACM Trans. Intell. Syst. Technol.*, Association for Computing Machinery, New York, NY, USA, v. 10, n. 3, apr 2019. ISSN 2157-6904. Disponível em: <<https://doi.org/10.1145/3305260>>. Citado 3 vezes nas páginas 22, 26 e 32.
- SHISHAH, W. Jointbert for detecting arabic fake news. *IEEE Access*, v. 10, 2022. Citado na página 41.
- SHU, K. et al. *defend: Explainable fake news detection*. 2019. 395–405 p. Citado na página 41.
- SOUZA, F. et al. Detecção de fake news em língua portuguesa combinando redes neurais convolucionais e algoritmos de aprendizagem de máquina. *Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos*, SBC, Porto Alegre, RS, Brasil, 2022. ISSN 2177-9384. Citado 5 vezes nas páginas 33, 34, 41, 46 e 47.
- SOUZA, J. V. de et al. A systematic mapping on automatic classification of fake news in social media. *Social Network Analysis and Mining*, Springer, v. 10, n. 1, p. 1–21, 2020. Citado 2 vezes nas páginas 48 e 55.
- SOUZA, M. P. de et al. *A linguistic-based method that combines polarity, emotion and grammatical characteristics to detect fake news in portuguese*. 2020. 217–224 p. Citado na página 41.
- TESTONI, G. A. et al. *Um método linguístico que combina polaridade, emoção e aspectos gramaticais para detecção de fake news em inglês*. 2021. 151–162 p. Citado na página 41.
- UNESCO. *Teorias da conspiração: ligadas à literatura*. 2021. Disponível em: <<https://pt.unesco.org/courier/2021-2/teorias-da-conspiracao-ligadas-literatura>>. Acesso em: 02 Maio 2022. Citado na página 26.
- VEYSEH, A. P. B. et al. *Rumor detection in social networks via deep contextual modeling*. 2019. 113–120 p. Citado na página 41.
- VIEIRA, R.; LOPES, L. Processamento de linguagem natural e o tratamento computacional de linguagens científicas. *EM CORPORA*, p. 183, 2010. Citado na página 27.
- WANG, Y. et al. *Eann: Event adversarial neural networks for multi-modal fake news detection*. 2018. 849–857 p. Citado na página 41.
- WU, F.; LUO, X.; DEMERTZIS, K. English text recognition deep learning framework to automatically identify fake news. *Intell. Neuroscience*, Hindawi Limited, London, GBR, v. 2022, 2022. ISSN 1687-5265. Citado na página 41.
- ZERVOPOULOS, A. et al. Deep learning for fake news detection on twitter regarding the 2019 hong kong protests. *Neural Comput. Appl.*, Springer-Verlag, Berlin, Heidelberg, v. 34, 2022. ISSN 0941-0643. Citado na página 41.
- ZHANG, L.; HUANG, G.; KHAN, M. A hybrid approach for network rumor detection based on attention mechanism and bidirectional gru model in big data environment. *Mob. Inf. Syst.*, IOS Press, NLD, v. 2022, 2022. ISSN 1574-017X. Citado 2 vezes nas páginas 41 e 46.
- ZHOU, X. et al. Fake news early detection: A theory-driven model. *Digital Threats: Research and Practice*, ACM New York, NY, USA, v. 1, n. 2, p. 1–25, 2020. Citado na página 41.
- ZHOU, X.; ZAFARANI, R. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Comput. Surv.*, Association for Computing Machinery, New York, NY, USA, v. 53, n. 5, sep 2020. ISSN 0360-0300. Disponível em: <<https://doi.org/10.1145/3395046>>. Citado na página 25.

- ZIANI, A. et al. Deceptive opinions detection using new proposed arabic semantic features. *Procedia Computer Science*, Elsevier, v. 189, p. 29–36, 2021. Citado na página 41.
- ZUBIAGA, A. et al. Detection and resolution of rumours in social media: A survey. *ACM Comput. Surv.*, Association for Computing Machinery, New York, NY, USA, v. 51, n. 2, feb 2018. ISSN 0360-0300. Disponível em: <<https://doi.org/10.1145/3161603>>. Citado na página 25.