UNIVERSIDADE ESTADUAL DA PARAÍBA
CAMPUS VII
CENTRO CCEA
CURSO DE BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

Natan Bento Cavalcante

AVALIAÇÃO DE ALGORITMOS DE APRENDIZADO DE MÁQUINA NA CLASSIFICAÇÃO DE *FAKE NEWS* EM LÍNGUA PORTUGUESA

PATOS
2024

Natan Bento Cavalcante

AVALIAÇÃO DE ALGORITMOS DE APRENDIZADO DE MÁQUINA NA CLASSIFICAÇÃO DE *FAKE NEWS* EM LÍNGUA PORTUGUESA

Trabalho de Conclusão de Curso apresentado ao Programa de Graduação em Ciência da Computação da Universidade Estadual da Paraíba, como requisito parcial à obtenção do título de Bacharel em Ciência da Computação.

Área de concentração: Aprendizagem de Máquina.

Orientadora: Ma. Keila Lucas dos Santos.

PATOS
2024

C376a Cavalcante, Natan Bento.

Avaliação de algoritmos de aprendizado de máquina na classificação de *fake news* em Língua Portuguesa [manuscrito] / Natan Bento Cavalcante. - 2024.

54 p. : il. colorido.

Digitado.

Trabalho de Conclusão de Curso (Graduação em Computação) - Universidade Estadual da Paraíba, Centro de Ciências Exatas e Sociais Aplicadas, 2024.

"Orientação : Profa. Ma. Keila Lucas dos Santos, Coordenação do Curso de Ciências Exatas - CCEA. "

1. Fake News. 2. Análise de Dados. 3. Aprendizagem de Máquina. 4. Aprendizagem Profunda. 5. Desinformação. 6. Notícias falsas. I. Título

21. ed. CDD 070.4

Elaborada por Camile de A. Gomes - CRB - 15/559 COBIB/UEPB

NATAN BENTO CAVALCANTE

AVALIAÇÃO DE ALGORITMOS DE APRENDIZADO DE MÁQUINA NA CLASSIFICAÇÃO DE *FAKE NEWS* EM LÍNGUA PORTUGUESA

Trabalho de Conclusão de Curso apresentado ao Curso de Bacharelado em Ciência da Computação da Universidade Estadual da Paraíba — Campus VII, em cumprimento à exigência para obtenção do grau de Bacharel em Ciência da Computação.

Aprovado em 22/04/2024

BANCA EXAMINADORA

Keila Lucas dos Santos

Prof.ª Ma. Keila Lucas dos Santos
(Orientadora)

José Aldo Silva da Costa

Prof. Dr. José Aldo Silva Costa
(Examinador)

Angélica Felix Medeiros

Prof.ª Ma. Angélica Felix Medeiros
(Examinadora)

AGRADECIMENTOS

A minha mãe Rita, que sempre apoiou minhas decisões e sempre fez de tudo para garantir meu bem estar.

Ao meu pai Afrânio, a minha avó Vanercilia, pelos conselhos, muitas vezes duros, mas que me fizeram chegar até aqui.

Aos meus tios Efrain e Mizia, que abriram as portas e os corações para mim, que são meu porto seguro aqui nessa cidade.

Aos meus colegas de curso Patrick, Luiz, Caio e Geraldo, que fizeram parte dessa jornada e estiveram dispostos a me dar apoio sempre que precisei, agradeço pela amizade.

Aos professores do Curso de Ciência da Computação do campus-VII da UEPB, em especial, Pablo Roberto, Keila, Fábio Júnior, Angélica e Ingrid que contribuíram e me inspiraram durante a minha formação.

Aos funcionários da UEPB, pela presteza e atendimento sempre que necessário.

Aos demais amigos e familiares que compartilharam alegrias e tristezas comigo.

RESUMO

A detecção automática de notícias falsas (*Fake News*, pelo idioma Inglês) é um dos processos mais importantes para combater a propagação de informações enganosas e potencialmente perigosas nos meios de comunicação. As técnicas de Aprendizagem de Máquina e Processamento de Linguagem Natural vêm sendo aplicadas como estratégia para identificação e filtragem das *Fake News*, proporcionando a revisão automática de informações virtuais. Esta pesquisa tem como objetivo avaliar o desempenho de dois modelos de Aprendizado de Máquina para a classificação de notícias falsas em Língua Portuguesa*. No desenvolvimento do estudo,* os modelos de Regressão Logística e Redes Hierárquicas de Atenção (HAN) foram treinados com os dados textuais da base *Fake.Br*, os quais obtiveram 97% na métrica F1-score. Posteriormente, a generalização dos modelos foi avaliada com os dados da base *FakeTrueBr*, destacando-se o modelo de Regressão Logística, que atingiu uma precisão média de 79,44% contra 72,33% do modelo HAN nos experimentos realizados. Os resultados alcançados nesta pesquisa são promissores, com viabilidade para novas análises de dados e suporte para o desenvolvimento de aplicações para detecção de notícias falsas compartilhadas em mídias sociais.

Palavras-chave: *Fake News*, Análise de Dados, Aprendizagem de Máquina, Aprendizagem Profunda, Desinformação, Notícias Falsas.

ABSTRACT

The automatic detection of Fake News is a critical process in combating the spread of misleading and potentially dangerous information in the media. Machine Learning (ML) and Natural Language Processing (NLP) techniques have been applied as strategies for identifying and filtering Fake News, enabling the automatic review of online information. This research aims to evaluate the performance of two Machine Learning models for classifying Fake News in Portuguese. Logistic Regression and Hierarchical Attention Networks (HAN) were trained using textual data from the Fake.Br database, achieving an F1-score of 97%. The models' generalization capabilities were subsequently assessed using data from the FakeTrueBr database, with the Logistic Regression model achieving an average accuracy of 79.44%, compared to 72.33% for the HAN model. The results obtained in this research are promising, demonstrating the viability for new data analyses and supporting the development of applications to detect Fake News shared on social media.

Key-words: Fake News, Data Analysis, Machine Learning, Deep Learning, Misinformation.

SUMÁRIO

1 INTRODUÇÃO

1.1 Contexto

As *Fake News* são notícias falsas, compartilhadas massivamente, que possuem caráter de desinformação e são potencialmente perigosas no contexto de influência social (Dourado, 2020). As discussões em torno das *Fake News* se popularizaram em meados de 2016, durante as eleições presidenciais dos Estados Unidos da América (EUA), acionando o debate acerca da influência das notícias falsas na intenção de voto dos cidadãos norte-americanos.

Durante o período eleitoral de 2016, a quantidade de notícias falsas favoráveis ao candidato Donald Trump, que foram compartilhadas nas mídias sociais, foi consideravelmente maior do que aquelas em apoio à candidata Hillary Clinton (Allcott; Gentzkow, 2017). Segundo Allcott e Gentzkow (2017), uma significativa parte dos eleitores americanos tiveram contato, bem como, se lembravam de conteúdos das *Fake News* mais populares naquela época, e isso pode ter sido um dos fatores de influência para a vitória de Donald Trump à presidência dos EUA.

No âmbito da política brasileira, especificamente no período eleitoral de 2018, a população nacional vivenciou também um cenário de forte disseminação de *Fake News*. Dourado (2020) menciona em sua pesquisa que 226 (65,31%) de um total de 346 notícias falsas divulgadas nas mídias sociais, eram favoráveis ao candidato Jair Bolsonaro, também eleito à presidência do Brasil.

Carvalho (2020), discute sobre um panorama das mídias sociais e seus impactos na democracia, apontando uma “democracia frustrada” pela influência do conteúdo, principalmente das *Fake News,* que geralmente são disseminadas na Internet, com destacável vantagem sobre o conteúdo verídico e relevante.

Braga (2018) discute a influência das *Fake News* na opinião pública, pela perspectiva da narrativa de discurso de ódio, caracterizado pela segregação social dos indivíduos através de perseguição, insultos ou privação de direitos. Enquanto Quadrado e Ferreira (2020) apresentam as *Fake News* como sendo uma características de destaque no viés do discurso de ódio, apontando os aspectos da interação com as redes sociais e a máscara da liberdade de expressão como

ferramenta de estigmatização, de forma que, os indivíduos se sentem livres para expressar suas opiniões independentemente de consequências. Tais aspectos favorecem o compartilhamento de informações sem que haja a checagem dos conteúdos, o que colabora com a propagação do ódio no âmbito da Internet.

Na área da Saúde, o impacto das *Fake News* está relacionado à propagação de desinformação sobre doenças, sintomas e tratamentos, especialmente em meio a população jovem. Segundo Zanatta *et al.* (2021), o público mais jovem tem mais segurança em utilizar ferramentas online para buscar orientações sobre à saúde. Com a propensão dessa população, cada vez mais, consultar a Internet sobre problemas relacionados à saúde, as notícias falsas têm um grande potencial de prejudicar ações sociais. A exemplo, durante a pandemia de Coronavírus (2020 - 2023), notícias falsas dificultaram o combate a propagação da doença, sendo uma das principais ferramentas dos grupos negacionistas para difundirem idéias distópicas e extremamente prejudicial à saúde coletiva (Marques; Raimundo, 2021).

1.2 Problemática

Diante do contexto das *Fake News*, existe uma necessidade urgente de identificar o que é informação verdadeira e o que é informação falsa nas mídias *on-lines*. As *Fake News* são estruturadas de modo a "simular" notícias reais, dificultando o processo de identificação imediata.

Em agências de checagem – Órgãos que analisam as informações veiculadas na Internet – é realizado um trabalho manual, no qual os avaliadores são responsáveis por avaliar se uma notícia é verdadeira ou não, explicitando o processo de verificação da informação com enfoque nas fontes originais (Spinelli; Santos, 2018).

Para Spinelli e Santos (2018) a classificação manual - realizada por avaliadores humanos - possui pouco alcance de revisão, em virtude do grande volume de novas informações que são publicadas na Internet e também da maneira com que o público interage com elas, sendo necessário mão de obra maior para auxiliar o processo de checagem.

Algumas triagens manuais são integradas a ferramentas que previamente alertam sobre a possível existência de *Fake News* em *websites*. O *Fake News*

Detector[1], por exemplo, é uma extensão que utiliza uma lista restrita de websites analisados, para determinar a possibilidade de uma notícia ser falsa ou um *clickbait*[2].

Em relação à análise automática de notícias falsas, existem ferramentas que são capazes de ponderar sobre as notícias de forma rápida e otimizada, utilizando-se de diferentes técnicas computacionais como Aprendizagem de Máquina e Processamento de Linguagem Natural na avaliação de repositórios de informações (Monteiro *et al.,* 2018; Guarise, 2019; Battisti, 2020)

As alternativas automáticas estão entre as melhores estratégias para o processo de classificação das notícias falsas, oferecendo maior agilidade na identificação de termos suspeitos e construções de informações falaciosas, investir em tecnologias computacionais para a detecção das notícias falsas pode otimizar a checagem dos fatos e ajudar no combate à desinformação (Ciampaglia *et al.*, 2015).

1.3 Motivação

Através da Inteligência Artificial e da Aprendizagem de Máquina, tornou-se possível aplicar algoritmos eficientes para classificar as *Fake News* com alta taxa de precisão.

Oliveira (2019) utiliza duas abordagens de Inteligência Artificial para analisar os padrões encontrados nas *Fake News* e verificar a possibilidade do uso desses métodos na classificação de *Fake News*. Uma delas é voltada para avaliar algoritmos de aprendizagem clássicos e outra para algoritmos modernos de Redes Neurais. Guarise (2019) desenvolveu um classificador de notícias falsas utilizando técnicas de Processamento de Linguagem Natural e Aprendizado Profundo para detecção de *Fake News* em Língua Portuguesa, no qual a precisão de classificação foi de 95,35% nos testes apresentados.

Diante dos estudos citados e dos resultados bem sucedidos envolvendo a aplicação da Aprendizagem de Máquina e classificação de notícias falsas, este estudo pretende avaliar a eficácia dos algoritmos de Regressão Logística e Redes Neurais Hierárquicas (HAN) para a classificação de *Fake News* em Língua

---

[1] Extensão de classificação de *Fake News*, disponível em: https://chrome.google.com/webstore/detail/fake-news-detector/aebaikmeedenaijgjcfmndfknoobahep?hl=pt-BR. Acesso em 11/03/2022.

[2] Postagem demasiadamente chamativa cujo objetivo é atrair o usuário a entrar em um site.

Portuguesa, visando o avanço no desenvolvimento de ferramentas com integração de modelos computacionais que ampliem a detecção em alta escala de informações falsas compartilhadas na Internet.

1.4 Objetivos

1.4.1 Objetivo geral

Avaliar a classificação de *Fake News* em Língua Portuguesa com aplicação da Aprendizagem de Máquina para determinar a veracidade de notícias em mídias digitais.

1.4.2 Objetivos específicos

- Identificar os desafios no processo de classificação de notícias falsas em Língua Portuguesa.
- Avaliar o desempenho dos algoritmos de Regressão Logística e do Modelo HAN no contexto das *Fake News*.
- Promover um processo eficiente para a classificação automática de notícias falsas.

1.5 Justificativa

A integração de um modelo computacional em ferramentas de análise de notícias é capaz de beneficiar o combate à propagação da desinformação em mídias digitais de comunicação. A construção de uma extensão integrada ao navegador, tal como o *Fake News Detector,* que notifica o usuário ao acessar uma notícia falsa, permite um cenário mais confiável de acesso às informações e favorece o processo de interrupção de conteúdo falso e danoso ao público.

A presente pesquisa tem como intuito validar métodos de Aprendizagem de Máquina na classificação de notícias falsas, verificando a possibilidade de sua utilização em contexto do mundo real, na tentativa de identificar limitações dos

métodos propostos. Os algoritmos de Regressão Logística e Redes Neurais Hierárquicas (HAN) serão avaliados em processos de classificação das notícias e através dos resultados, pretende-se respaldar a eficiência da análise automática de algoritmos no combate às *Fake News*.

2 REFERENCIAL TEÓRICO

Neste capítulo serão apresentados conceitos relevantes acerca do tema de notícias falsas *(Fake News)*, incluindo a apresentação dos principais conceitos sobre Aprendizado de Máquina e classificação de texto por algoritmos computacionais.

2.1 *Fake News*

Uma *Fake News* pode ser considerada como um “tipo específico de informação com potencial de gerar engano ou desinformação, porque faz com que os indivíduos assumam como verdadeiro e real o que é mentiroso e falso” (Dourado, 2020, p. 40). No contexto das notícias divulgadas em mídias digitais, detectar as *Fake News* é muito importante para prevenir compartilhamento de desinformação, que poderiam provocar diversos problemas sociais, como: Informações erradas durante processos eletivos, prejudicar o alinhamento de investimentos empresariais ou a alocação de recursos em caso de crises como ataques terroristas ou desastres naturais, conforme explica Vosoughi, Roy e Aral (2018).

Outra definição associada é que *Fake News* são artigos de notícias falsas escritos propositalmente que podem enganar os leitores, sendo as mídias sociais o maior veículo para sua disseminação (Allcott; Gentzkow, 2017). As notícias falsas têm como princípio manipular as emoções, que por sua vez, pode dificultar aos leitores a tomarem decisões de forma racional sobre a notícia (Sivek, 2018) favorecendo o compartilhamento imediato em meio social, podendo assim, ganhar credibilidade ao ser compartilhado por pessoas conhecidas ou influentes (Carvalho, 2020).

Os *bots* (geradores automáticos de mensagens) são amplamente utilizados na divulgação das *Fake News*, no entanto o compartilhamento de notícias falsas é ampliado principalmente por ações humanas, geralmente grupos de pessoas que agem em conjunto para compartilhar *Fake News* em mídias sociais (Vosoughi; Roy; Aral, 2018). Segundo Quadrado e Ferreira (2020), as *Fake News* fomentam o discurso de ódio nas redes sociais, provocando violação dos Direitos Humanos, uma vez que, as segregações decorrentes dessa narrativa põem em risco a liberdade e direitos de determinados grupos e minorias.

Braga (2018) destaca o uso das *Fake News* como ferramenta para possível obtenção de vantagem política. Sobre a perspectiva do discurso de ódio e da estigmatização de determinados grupos, a veiculação de notícias associativas de políticos a esses grupos - como exemplo, a imagem associada a um grupo terrorista - pode causar a má compreensão eleitoral mediante às repercussões.

Teixeira e Santos (2020) apontam o perigo das *Fake News* na saúde em divulgações de mídia digitais, em especial, os casos de anti-vacinação da febre amarela. Em revisão bibliográfica sobre a influência das notícias falsas na hesitação vacinal contra a COVID-19, Silva *et al.* (2023) discutem os impactos das informações falsas sobre possíveis efeitos colaterais da vacina, gerando receios na população e prejudicando o processo de imunização da sociedade. O compartilhamento de desinformação promove o caos social e consequentemente deixa a sociedade vulnerável a doenças (Teixeira; Santos, 2020).

Em análise às *Fake News* compartilhadas nos primeiros meses da pandemia de COVID-19 em 2020, Barcelos *et al.* (2021) observaram um aumento nas buscas por termos presentes nas notícias falsas, destacando as categorias mais frequentes como política, estatísticas e epidemiologia e prevenção.

Diferente das notícias jornalísticas, que são revisadas e tendem a apresentar fatos reais em seu conteúdo, as *Fake News* são criadas para causar impacto e comoção, sendo caracterizadas pelo uso de termos e recorrência de padrões textuais. Atualmente, o reconhecimento desses padrões pode ser desenvolvido a partir das técnicas de análise de dados, com aplicação de classificadores inteligentes, construídos com Aprendizado de Máquina e Aprendizado Profundo, que favorecem o reconhecimento de padrões em texto, possibilitando a detecção de notícias digitais falsas.

2.2 Aprendizado de Máquina

O Aprendizado de Máquina diz respeito à capacidade do computador em aprender com dados e identificar padrões para a definição de resultados com o mínimo de intervenção humana. O Aprendizado de Máquina é a área da programação de computadores baseada na aprendizagem por dados (Géron, 2019).

O objetivo do Aprendizado de Máquina é fazer o computador aprender através de dados passados, e dessa forma ser capaz de tomar decisões de forma genérica de acordo com alguns exemplos (Faceli *et al.*, 2021). Dessa maneira os computadores podem realizar determinadas tarefas sem que seja diretamente programado para tal, por exemplo, reconhecer padrões em imagens ou classificar documentos.

Para Faceli *et al.* (2021), as tarefas do Aprendizado de Máquina podem ser definidas através de diferentes critérios, sendo um deles o paradigma de aprendizado associado à realização da tarefa. Nesse aspecto, as tarefas podem ser divididas em preditivas, onde o algoritmo tem como objetivo prever uma saída de acordo com rótulos atribuídos ao conjunto de dados, e tarefas descritivas, onde o algoritmo tem como objetivo analisar a base de dados a fim de criar inferências de acordo com critérios pré-estabelecidos.

2.2.1 Paradigmas de Aprendizado de Máquina

O Aprendizado de Máquina é definido por quatro tipos principais de aprendizagem, que caracterizam o modo como o modelo computacional aprenderá sobre o conjunto de dados, sendo eles: Supervisionado, Não Supervisionado, Semissupervisionado e Por Reforço (Géron, 2019).

Aprendizado Supervisionado - Consiste na apresentação dos rótulos das instâncias do conjunto de dados, para o treinamento do algoritmo. Os algoritmos supervisionados analisam os padrões de cada classe e realizam as tarefas preditivas. Os principais algoritmos de aprendizagem supervisionada são:

- **Árvore de Decisão:** Constrói uma estrutura em forma de árvore, onde cada nó interno é um teste de um atributo, cada ramificação corresponde a um resultado possível e cada nó folha é uma classe ou decisão. O objetivo é simplificar o conjunto de dados até que todas as instâncias pertençam à mesma classe.
- ***Naive Bayes*:** Tem como princípio o Teorema de Bayes, estabelecendo uma relação entre a probabilidade condicional entre duas variáveis.

- ***Support Vector Machine* (SVM):** Algoritmo que busca encontrar o hiperplano que melhor separa duas classes, através de um espaço de características.

Aprendizado Não Supervisionado - Neste tipo de aprendizagem, os dados de treinamento não apresentam rótulos, o algoritmo tenta inferir soluções a partir dos padrões identificados no próprio conjunto. Esse paradigma está associado a tarefas descritivas da base de dados, por exemplo a realização de agrupamento ou associação de instâncias.

- **K-means:** É um algoritmo de agrupamento que divide os dados em grupos (*clusters)* com base em um ponto central ou centróide. Para atualização dos *clusters* é calculada a distância entre os dados e a centróide, onde cada ponto de dado é inserido ao grupo com o ponto central mais próximo.
- **Autoencoders:** São Redes Neurais Artificiais treinadas para gerar representações eficientes dos dados de entrada, utilizando dados não rotulados. Tem como objetivo copiar a entrada para sua saída, normalmente de maneira a capturar características relevantes para criar uma cópia aproximada da entrada. Podem ser utilizados para diminuir a dimensionalidade dos dados, redução de ruídos e gerar novos dados no escopo dos dados de treino.

Aprendizado Semissupervisionado - É caracterizado quando há uma parte do conjunto de dados que não possui rótulos, desta maneira o treinamento do modelo é realizado sobre dados rotulados e não rotulados. O objetivo é utilizar os dados com e sem rótulos para criar modelos preditivos mais robustos. Uma das abordagens desse paradigma é utilizar os dados rotulados para criar rótulos de instâncias mais expressivas da base não rotulada.

- **Redes Neurais de Crenças Profundas:** Tem como base modelos estocásticos generativos não supervisionados, denominados Máquinas Restritas de Boltzmann, empilhados uns aos outros. Esses modelos são treinados de maneira não supervisionada, e posteriormente são ajustados utilizando técnicas de Aprendizado Supervisionado.

Aprendizado por Reforço - Essa forma de aprendizado acontece a partir da atribuição de pesos a acertos e erros do algoritmo. Dessa maneira, para uma ação

correta o algoritmo é recompensado, enquanto os erros resultam em uma recompensa reduzida ou até uma penalidade. Os algoritmos desse paradigma aprendem por meio da tentativa e erro, a exemplo de um robô inteligente que é “repreendido” cada vez que erra a tarefa que foi atribuída.

- **Q-learning:** É um algoritmo de aprendizado por Reforço baseado em valor, projetado para aprender a função de valor de ação $Q(s,\ a)$. Essa função representa a recompensa esperada ao tomar uma ação a a partir do estado s. O Q-learning busca encontrar a política que maximize a recompensa total ao longo do tempo.
- **AlphaGo:** Um programa desenvolvido para aprender e jogar o jogo de tabuleiro Go, sendo a primeira aplicação de Inteligência Artificial para derrotar humanos nesse jogo. Ele foi treinado utilizando os dados de milhões de jogos e praticando contra si mesmo, desenvolvendo sua política vencedora através da tentativa e erro.

2.2.2 Regressão Linear e Logística

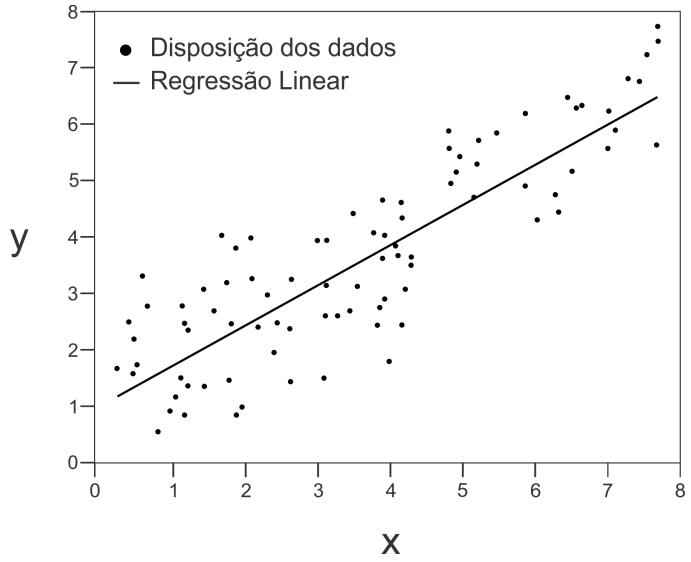
A Regressão é uma técnica estatística aplicada para estimar a condicional de uma variável y, considerando os valores de algumas outras variáveis X (Goodfellow; Bengio; Courville, 2016). A regressão é uma das principais tarefas na área do Aprendizado de Máquina, definida por uma função $f: \mathbb{R}^{n} \rightarrow \mathbb{R}$, para inferir valores y - saída - de acordo com cada entrada de x.

A Regressão Linear é uma das abordagens de regressão para o Aprendizado de Máquina, na qual a saída desse sistema é uma função linear ajustada pelos dados de entrada, definida por: $\hat{y} = w^{T}x$, onde ŷ é o valor predito de y, e w o vetor de parâmetros que funcionam como pesos para determinar a precisão do modelo. Podendo também, em um modelo mais sofisticado possuir um termo de interceptação b, transformando em: $\hat{y} = w^{T}x + b$, esse termo chamado de viés, serve para ajustar o modelo, definindo o ponto inicial da linha de regressão (Goodfellow; Bengio; Courville, 2016).

Na Figura 1 é possível visualizar a disposição dos dados em conjunto com a função de Regressão Linear. Cada ponto no gráfico representa a observação dos

valores reais dos dados, enquanto a linha traçada reflete os valores previstos pelo modelo.

Figura 1 - Representação da Regressão Linear.



Fonte: Autoria própria.

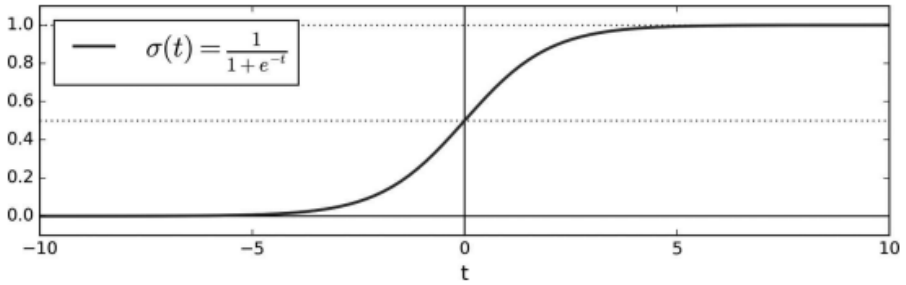
Na Regressão Linear, uma das maneiras de medir a precisão do modelo preditivo é através do cálculo do erro quadrático médio, processo no qual é realizada a soma do quadrado da distância - em relação a função linear - de cada ponto na disposição dos dados no espaço (Goodfellow; Bengio; Courville, 2016). O objetivo do modelo de regressão, é ajustar os pesos *w* para obter um menor valor da soma dos erros quadráticos dos dados de treinamento. James *et al.* (2013) afirmam que a Regressão Linear é uma maneira simples e direta de prever um valor quantitativo de y, considerando uma única variável de predição x - supondo que exista uma relação próxima de linearidade entre elas.

O processo de regressão também pode ser aplicado em tarefas de previsão de classes. Neste objetivo, destaca-se a Regressão Logística, um tipo de regressão utilizada para calcular a probabilidade de uma instância pertencer a uma classe específica, como determinar a chance de um e-mail ser spam (Géron, 2019). O modelo de Regressão Logística, em contexto da classificação binária, calcula a probabilidade de uma das classes ser superior a 50%, para atribuir o rótulo da classe em questão.

A Regressão Logística produz uma função logística - que é função sigmóide em formato de “S”, na qual, $\sigma(\cdot)$ - que varia entre 0 e 1 para calcular a probabilidade de uma entrada pertencer a uma determinada classe. Após estimar a probabilidade

da instância pertencer a classe positiva o algoritmo será capaz de prever o valor de ŷ (Géron, 2019).

Figura 2 - Representação da Regressão Logística.



Fonte: Géron (2019).

Na Figura 2, se t ≥ 0 então σ(t) que é a função de estimativa da probabilidade, vai ser maior ou igual a 0.5, logo ŷ = 1. Caso contrário, para t < 0, ŷ = 0. A representação matemática pode ser observada na Figura 3.

Figura 3 - Previsão do Modelo de Regressão Logística.

$$\hat{y} = \begin{cases} 1, \text{ se } \sigma(w^T x) \geq 0.5, \\ 0, \text{ se } \sigma(w^T x) < 0.5. \end{cases}$$

Fonte: Adaptado de Géron (2019).

Para se adequar a classificação o modelo utiliza a máxima verossimilhança, esse método estima valores para w, tal que, para cada previsão de σ(t), o resultado seja o mais próximo possível dos valores reais observados. Ou seja, atribuir valores a w para que ao estimar a probabilidade dos dados seja possível obter um valor próximo a 1, para as instâncias pertencer a classe, e um valor próximo de 0 para as que não pertencem a classe (James *et al.*, 2013).

Embora a Regressão Linear e Logística sejam bastante eficientes no cenário do Aprendizado de Máquina, suas limitações tornam-se evidentes ao lidar com padrões complexos e não lineares nos dados. Para superar esses desafios e capturar de forma mais eficaz as complexidades das instâncias de dados, o Aprendizado Profundo possibilita soluções poderosas, a partir das Redes Neurais

Artificiais, oferecendo uma abordagem avançada para modelar relações e padrões complexos.

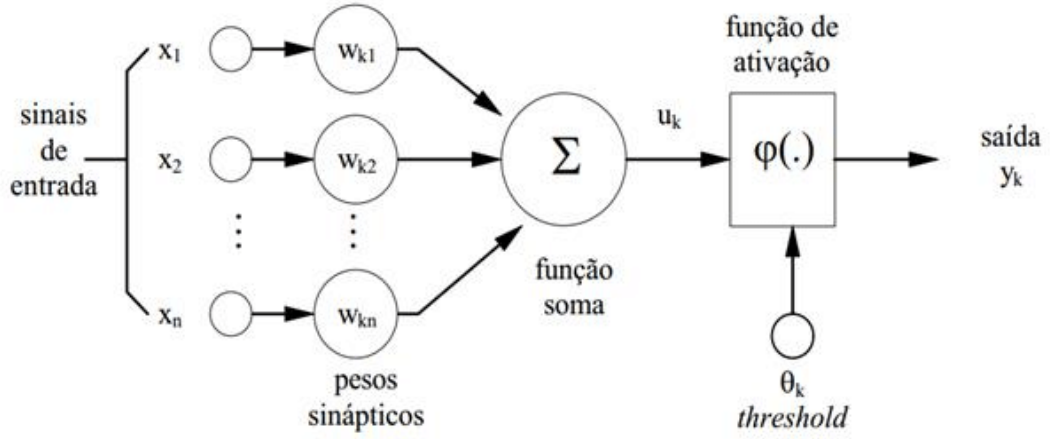
2.3 Aprendizado Profundo

O Aprendizado Profundo (*Deep Learning*) é uma subárea do Aprendizado de Máquina, com uma abordagem de aprendizagem baseada em Redes Neurais Artificiais (RNA) (Aggarwal, 2018). Uma RNA é uma estrutura computacional projetada para processar dados à maneira como o cérebro realiza uma tarefa particular ou função de interesse (Haykin, 2001). Em outras palavras, é uma tentativa de simular o processo de pensamento, se baseando nas características biológicas dos seres vivos.

A Figura 4 representa o modelo de um neurônio artificial, estrutura básica que compõe em organização de camadas uma Rede Neural Artificial. Um neurônio artificial apresenta as seguintes características estruturais:

- Um conjunto ponderado de entradas, denominadas sinapses;
- Uma função de soma para junção das entradas;
- Uma função de ativação para calcular a saída do neurônio.

Figura 4 - Representação de um neurônio artificial.



Fonte: Haykin (2001).

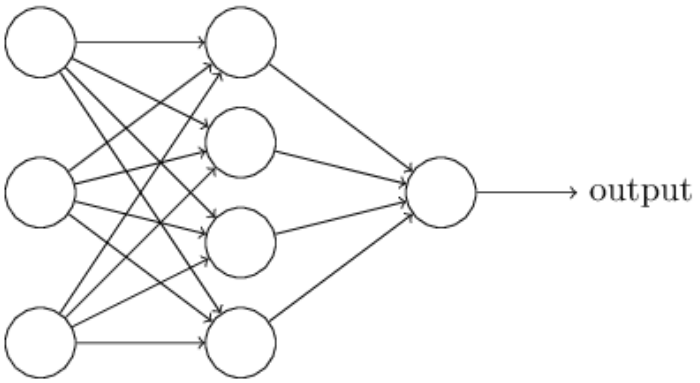
O neurônio artificial recebe as entradas, realiza a soma ponderada dos pesos através da função de soma, e com o resultado a função de ativação determina se a

saída será ativada ou não, para gerar novas entradas de outros neurônios (Nielsen, 2015).

As Redes Neurais Artificiais são compostas por um conjunto de neurônios artificiais, interligados em camadas, criando uma unidade complexa de processamento (Goodfellow; Bengio; Courville, 2016). Haykin (2001) ressalta duas características básicas das Redes Neurais Artificiais, a capacidade de aprender conforme o ambiente e a capacidade de mudar o peso entre as conexões dos neurônios para armazenar o conhecimento. Dessa forma, uma RNA tem a capacidade de evoluir com o processamento de dados.

A Figura 5 ilustra uma RNA formada por uma camada de entrada (que recebe os dados a serem processados), uma camada oculta (camada mais interna da rede) e uma camada de saída, formada pela quantidade de neurônios que for necessária para calcular a saída do processamento de dados (Nielsen, 2015).

Figura 5 - Representação de uma Rede Neural Artificial.



Fonte: Nielsen (2015).

A representação da Figura 5 é um dos modelos heurísticos para a definição das Redes Neurais, denominado Rede Neural *Feedforward (FNN)*. De acordo com Nielsen (2015) e Goodfellow, Bengio e Courville (2016), esse tipo de rede propaga a informação em apenas uma direção, ou seja, não existem *loops* no sistema, os dados são sempre repassados sem que voltem à um ponto anterior.

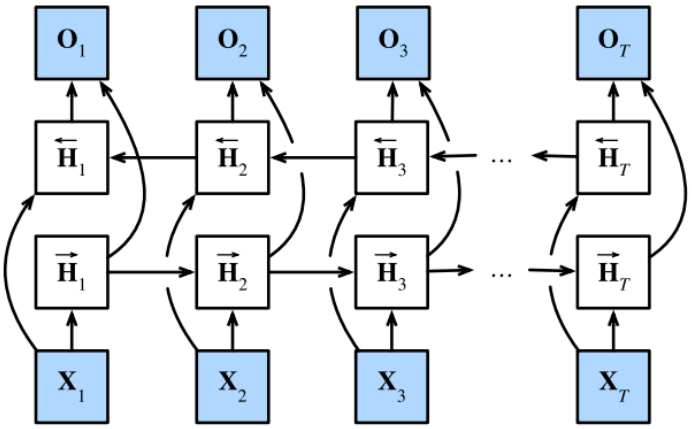
Quando a rede é estruturada para permitir que os dados sejam retroalimentados, então a RNA é definida como uma Rede Neural Recorrente (RNN). Nesse contexto, os neurônios ativam em intervalos temporais, estimulando outros neurônios a dispararem após um certo período. Isso assegura que os dados

sejam atualizados antes de serem reintroduzidos em um nó da rede (Nielsen, 2015), promovendo uma propagação eficiente das informações ao longo do tempo. Esse formato de rede permite uma dimensão mais substancial de entradas, para redes baseadas em sequência, também permitindo a entrada variável dos dados (Goodfellow; Bengio; Courville, 2016).

Considerando que as RNN podem enviar o estado da saída a um ponto no passado da rede, surgiu o modelo das Redes Neurais Recorrentes Bidirecionais que conectam camadas ocultas de direções opostas à mesma saída, podendo obter informações de estados passados e futuros simultaneamente (Zhang *et al.* 2023).

Na Figura 6 apresenta-se uma representação da arquitetura RNN bidirecional.

Figura 6 - Arquitetura de RNN bidirecional.



Fonte: Zhang *et al.* (2023).

Esse modelo proporciona uma compreensão mais abrangente e sofisticada dos dados sequenciais, particularmente eficaz em tarefas onde o contexto global é essencial para a tomada de decisão. A capacidade de uma RNA em processar grande quantidade de dados, favorece a extração de recursos e a identificação de padrões.

Conforme ressalta Nielsen (2015), os algoritmos de Aprendizado de Máquina, em particular as Redes Neurais, são capazes de modelar de forma automática problemas que seriam difíceis de realizar de forma manual, por exemplo, o reconhecimento de padrões em imagens. Isso representa um avanço significativo na capacidade de processamento e análise de dados, possibilitando a resolução de problemas complexos. A convergência entre capacidade de processamento e

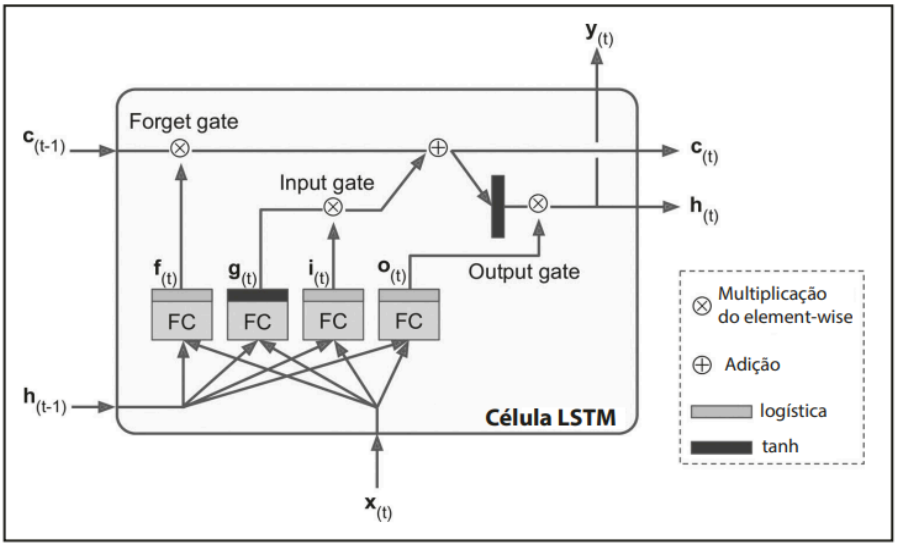
aprendizado automático é o cerne da revolução atual no campo da Inteligência Artificial.

2.3.1 LSTM

A arquitetura Memória Longa de Curto Prazo (LSTM), é uma versão aprimorada das Redes Neurais Recorrentes. Essa arquitetura surgiu para solucionar o problema do gradiente desaparecendo, decorrente das transformações dos dados ao longo do tempo em uma RNN (Géron, 2019). As LSTMs incorporam camadas de estado adicionais, semelhantes às camadas ocultas do modelo. Essas camadas são dedicadas a armazenar informações adicionais ao longo do tempo (Zhang *et al.*, 2023).

As células LSTM se parecem com uma célula comum - equivalente a um neurônio em RNN - a diferença se encontra em dois vetores de estados: $h_{(t)}$ e $c_{(t)}$, onde $h_{(t)}$ sendo o estado de curto prazo e $c_{(t)}$ o estado de longo prazo (Géron, 2019). A ideia desse modelo consiste em aprender quais memórias armazenar, ler ou deletar. Cada camada, representada na Figura 7, é definida pelas seguintes características:

Figura 7 - Modelo de Célula LSTM.



Fonte: Géron (2019).

- **A camada principal** ($g_{(t)}$) - responsável por analisar as entradas de $x_{(t)}$ e $h_{(t)}$;
- **O *forget gate*** ($f_{(t)}$) - capaz de aprender sobre quais entradas devem ser preservadas ou apagadas no intervalo de tempo.
- **O *input gate*** ($i_{(t)}$) - capaz de aprender sobre quais entradas são importantes para adicionar ao estado de longo prazo.
- **O *output gate*** ($g_{(t)}$) - capaz de aprender a controlar quais memórias devem ser lidas quando necessárias.

As LSTMs têm a capacidade única de preservar e acessar informações importantes de longo prazo, permitindo que o modelo capture relações temporais complexas em dados sequenciais de maneira mais eficaz do que as RNNs tradicionais.

2.3.2 Mecanismos de Atenção

A atenção, em contextos de Aprendizado de Máquina, pode ser direcionada tanto pelo estímulo da tarefa quanto por pistas volitivas associadas a um objetivo específico. Isso significa que, quando há uma tarefa definida, a tendência do foco pode ser influenciada não apenas pela relevância dos elementos de entrada, mas também pela orientação voluntária fornecida pela natureza da tarefa (Zhang *et al.* 2023).

No mundo visual, as características dos objetos podem estimular o foco da atenção, determinadas vezes como um processo de atenção não voluntária, mas quando existe o interesse, a tendência do foco pode ser atrelada a um objetivo.

Os mecanismos de atenção em Redes Neurais tem como princípio distribuir pesos sobre a entrada de dados, dessa maneira, atribuindo valores mais altos sobre os elementos mais relevantes da entrada (Galassi; Lippi; Torroni, 2021). Essa capacidade dinâmica de atribuir pesos, melhora o desempenho em tarefas de Aprendizado de Máquina, especialmente em contextos nos quais partes específicas dos dados são mais informativas para a tarefa em questão.

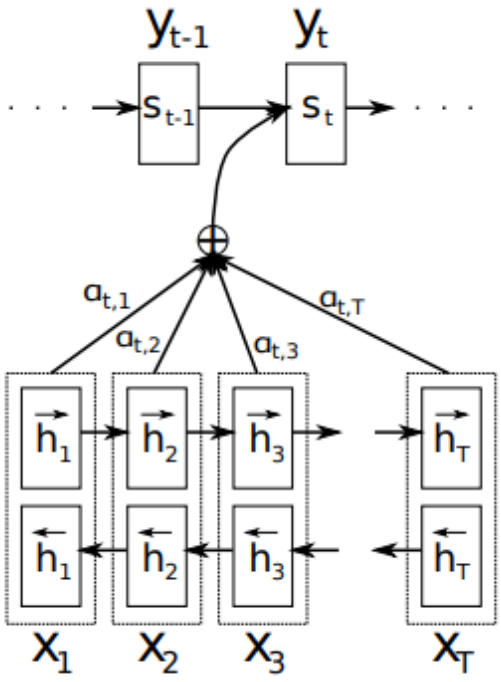
Para o aprimoramento das RNN, Bahdanau, Cho e Bengio (2014) propuseram um modelo de atenção tornando a rede capaz de procurar automaticamente por partes relevantes na sentença de dados. O modelo proposto

tem como arquitetura base um codificador-decodificador para a tradução automática, conforme apresentado na Figura 8.

Alguns dos componentes apresentados nessa arquitetura, são:

- S_{t-1} - **Estado oculto do decodificador na etapa antecedente:** Valor de cada sequência de entrada anterior, $t - 1$;
- h_i - **Estado oculto do codificador**: É uma anotação que possui informações de toda sequência de entrada X_i, focando nas partes que rodeiam a palavra na i-ésima posição da sequência.
- α_{ti} - **Pesos das anotações**: Peso associado a cada sequência de anotações, h_i, durante a etapa atual t.

Figura 8 - Codificador-decodificador RNN com atenção de Bahdanau.



Fonte: Bahdanau, Cho e Bengio (2014).

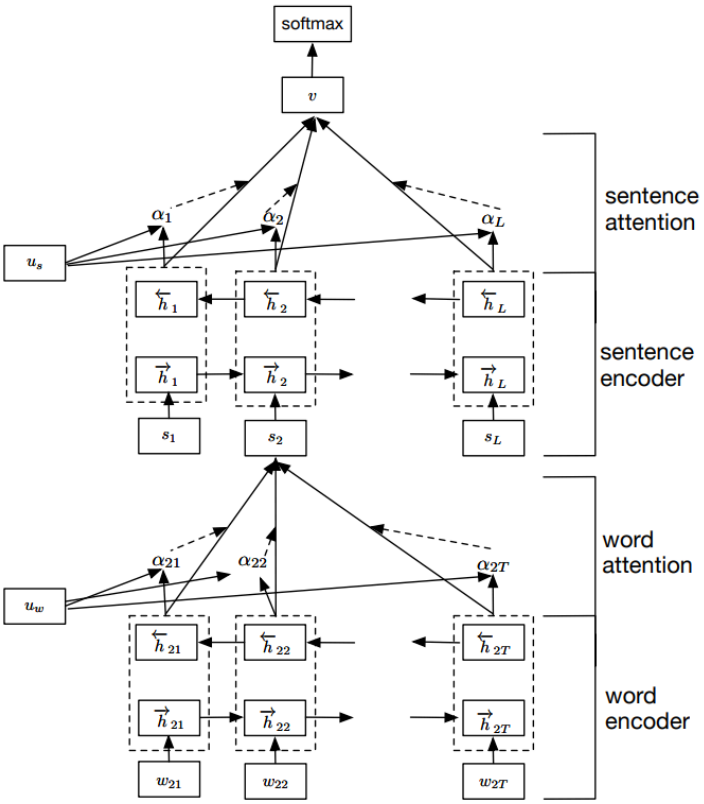
A camada de atenção nesse modelo é uma Rede Neural *Feedforward* capaz de aprender quais entradas são mais importantes para a previsão atual (Bahdanau; Cho; Bengio, 2014). A essência dessa abordagem está na variável de contexto c_t - que é obtida através da soma do produto dos pesos de cada sequência de

anotações, α_{ti}, pelo estado oculto do codificador h_i em qualquer tempo de decodificação t para T números de *tokens*: $c_t = \sum_{i=1}^{T} \alpha_{ti} h_i$.

2.3.3 O Modelo HAN

A Rede Hierárquica de Atenção (HAN), é uma arquitetura de Rede Neural proposta por Yang *et al.* (2016), que utiliza relações hierárquicas do texto para gerar uma representação completa do documento. Partindo do pressuposto que palavras formam sentenças e as sentenças formam documentos, o modelo HAN utiliza camadas de atenção para atribuir pesos em diferentes níveis hierárquicos.

Figura 9 - Estrutura da Rede Hierárquica de Atenção (HAN).



Fonte: Yang *et al.* (2016).

A estrutura do modelo HAN é composta pelas seguintes partes (Conforme apresentado na Figura 9):

- Um **codificador de sequência de palavras** - Para incorporar as palavras em *Embeddings* contextuais;
- Uma **camada de atenção no nível de palavras** - Para filtrar palavras irrelevantes ao significado da sentença;
- Um **codificador de sentenças** - Para construir a representação vetorizada das sentenças;
- E **uma camada de atenção no nível de sentenças** - Para ponderar as sentenças que contribuem na classificação do documento.

A respeito dos *Embeddings* contextuais, eles são uma representação matemática das palavras, realizada através de vetores, também conhecido como um processo de “incorporação de palavras” (Akdogan, 2021). Sendo uma importante técnica no Processamento de Linguagem Natural (NLP).

No processo de representação das palavras, uma das formas mais simples é através de um vetor de palavras quentes (*one-hot*), no qual as palavras são representadas por vetores de tamanho N (com N sendo o tamanho do vocabulário). Cada palavra pode ser representada pela sua posição i no conjunto de palavras, sendo assim, a representação *one-hot* de cada palavra seria um vetor de tamanho N preenchido por “zeros”, e “um” na posição i do vetor, representando sua posição dentro do conjunto de palavras (Zhang *et al.*, 2023).

No entanto, quando o conjunto de palavras é muito grande, essa representação *one-hot* torna-se um desafio. Uma das maneiras de solucionar essa representação de vetores de tamanho N é utilizar vetores de tamanho fixo, e trabalhar o processamento das palavras através de uma Rede Neural, que irá estimar uma boa representação para cada palavra durante o treinamento dos dados (Géron, 2019).

Para superar as limitações da representação one-hot, surgiu a abordagem Word2Vec, uma técnica de embedding desenvolvida por Mikolov et al. (2013), que visa reduzir a complexidade computacional em modelos de representação de palavras.

2.4 *Word2Vec*

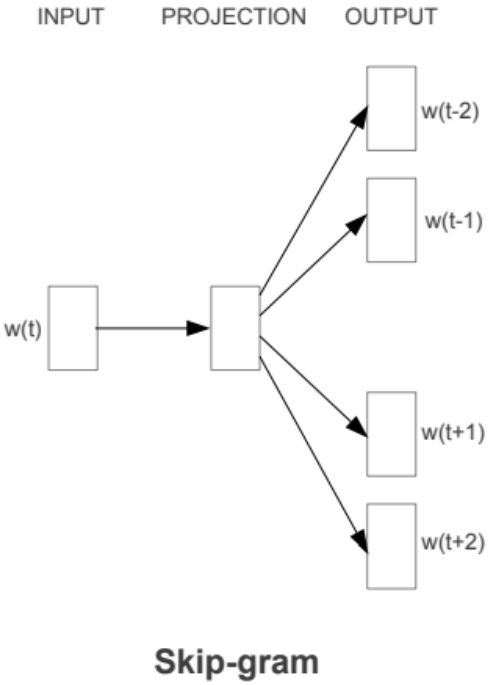
Word2Vec é uma técnica de *embedding* que tem como objetivo minimizar a complexidade computacional em modelos de representação distribuída de palavras.

Essa técnica utiliza pequenos vetores treinados, de números reais para representar as palavras, permitindo que palavras com proximidade lexical fiquem próximas em um espaço vetorial.

Mikolov *et al.* (2013), propôs dois modelos para o treinamento de *Embeddings Word2Vec*, denominados Saco de Palavras Contínuo (CBOW) e *Skip-gram*. Essas representações são distintas na maneira em que as palavras alvos são geradas, se baseando no contexto para gerar uma palavra alvo ou decompondo a palavra para encontrar associações.

O modelo *Skip-gram* tem como princípio prever a partir de uma palavra chave, quais são as palavras que aparecem no mesmo contexto que ela, buscando estabelecer as relações de inferência sobre o contexto. Na Figura 10 tem-se uma representação do modelo *Skip-gram*, que a partir da entrada w(t) o modelo prevê as palavras relacionadas que estão ao redor dela, dessa maneira, a saída consiste de palavras anteriores {w(t-1), w(t-2), …} e palavras posteriores à palavra atual {w(t+1) w(t+2), …}.

Figura 10 - Modelo *Skip-gram* para representação de palavras.

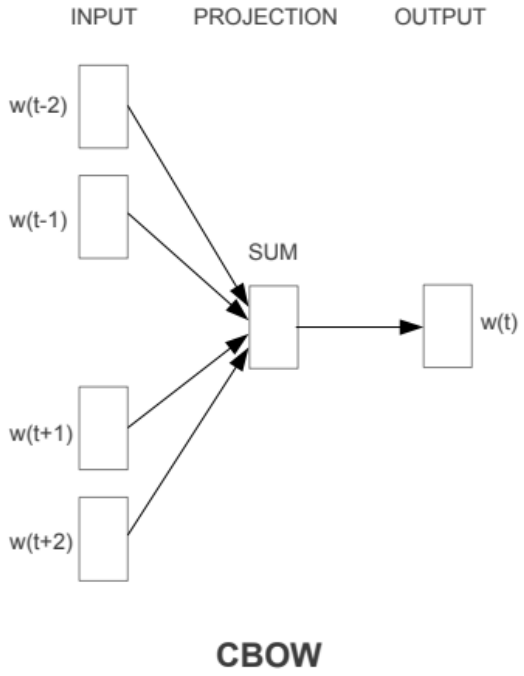


Fonte: Mikolov *et al.* (2013).

O modelo Saco de Palavras Contínuo (CBOW) assume que a palavra é definida com base no contexto, conforme apresentado na Figura 11, funcionando de

forma inversa ao *Skip-gram*, na qual o modelo tem como foco gerar palavras-alvo central tendo como base outras palavras no contexto. Ou seja, o modelo recebe um conjunto de entradas w(n) e projeta uma associação a uma representação de palavra no vetor w.

Figura 11 - Modelo CBOW para representação de palavras.



Fonte: Mikolov *et al.* (2013).

O desempenho das abordagens *Word2Vec* é comparável a modelos de Redes Neurais, sendo assim uma alternativa de menor custo computacional para a representação de palavras em vetores, especialmente no Processamento de Linguagem Natural (Hill *et al.*, 2016).

2.5 TF-IDF

O TF-IDF é uma técnica estatística para mensurar matematicamente a importância das palavras em documentos (Akdogan, 2021). A técnica funciona de forma semelhante ao vetor de palavras quentes, mas no lugar de criar vetores binários de tamanho n para representar cada item no conjunto se é atribuído um valor TF-IDF, esse termo é definido pela multiplicação dos valores de Frequência de Prazo (TF) e (Frequência Inversa do Documento) IDF:

- **Frequência de Prazo** - Mede a frequência que um termo aparece em um documento. Quanto maior o número de vezes que um termo aparece, maior será o valor TF.
- **Frequência Inversa do Documento** - Mede a raridade de uma palavra no conjunto de documentos, através da razão logarítmica das vezes que o termo aparece no documento e o número de documentos que possuem o termo. Palavras que são comuns em muitos documentos terão um IDF mais baixo, enquanto palavras raras terão um IDF mais alto.

Segundo Akdogan (2021), a vetorização do TF-IDF cria um vetor de tamanho n - número de instâncias únicas de palavras - para cada documento. Os vetores são preenchidos com os valores TF-IDF para cada posição i correspondente a disposição da palavra no documento. Dessa maneira é possível mensurar a relevância de uma palavra no documento em relação ao todo, pois através dos valores atribuídos ao vetor de representação, não apenas serão identificadas as palavras presentes no documento, mas a sua relevância contextual.

Neste Capítulo foram apresentados conceitos acerca das *Fake News* e do seu impacto social, assim como foi ressaltado a importância de aperfeiçoar as ferramentas de detecção das notícias falsas, considerando as aplicações de Aprendizado de Máquina, Redes Neurais e as técnicas de *embeddings* para o processamento de Linguagem Natural, facilitando a análise e interpretação de textos, essenciais para detectar e combater as *Fake News* de maneira eficiente.

No próximo Capítulo será realizada a apresentação dos trabalhos relacionados com os objetivos deste estudo, com destaque para as metodologias e técnicas aplicadas para a análise de notícias falsas.

3 TRABALHOS RELACIONADOS

Diante do objetivo de identificar as *Fake News* e diminuir a disseminação em mídias sociais, este trabalho buscou através de uma pesquisa bibliográfica encontrar trabalhos relacionados ao tema, que abordassem sobre diferentes estratégias e métodos direcionados para a identificação automática de notícias falsas, com destaque de resultados significativos, além da discussão de desafios e lacunas entre as soluções para o problema das *Fake News*.

Oliveira (2019) propôs a detecção de *Fake News* aplicando diferentes algoritmos de Aprendizado de Máquina, como Árvore de Decisão, *Naive Bayes*, *Support Vector Machine* (SVM) e modelos baseados em Redes Neurais Artificiais. No estudo, a classificação dos dados com o processamento das Redes Neurais alcançou uma acurácia de 92,36% e 94,46% de precisão nos testes. Dentre os algoritmos clássicos avaliados, o algoritmo SVM obteve uma acurácia de 90,13% e precisão de 87,53%, se apresentando superior aos demais algoritmos que tiveram desempenho inferior a 80% de acurácia e precisão.

Guarise (2019) desenvolveu uma aplicação de Aprendizado Profundo para a classificação de *Fake News*, com foco em artigos em Língua Portuguesa. No estudo realizado, o autor aplicou um modelo de classificação de texto, especificamente um modelo HAN, para classificar as notícias e nos resultados da tarefa de classificação, o modelo teve uma acurácia de 95,35%, com precisão de 96,80% na classificação de notícias falsas e 93,89% em notícias verdadeiras. O algoritmo também foi testado com 30 notícias de fora da base de dados com 15 instâncias de cada categoria, obtendo uma acurácia de 73,33%, precisão de 60% e 86,67% em classificar notícias verdadeiras e *Fake News* respectivamente.

Battisti (2020) realizou uma abordagem comparativa entre os algoritmos de Regressão Logística e *Long Short-Term Memory* (LSTM) para a classificação de dados da base *Fake.Br* Corpus[3]. No estudo desenvolvido, o melhor resultado foi obtido pelo modelo de Regressão Logística, alcançando 92,8% na métrica F1-score[4], enquanto o melhor modelo de LSTM obteve F1-score de 89,7%. Silva (2022) também utilizou Regressão Logística na classificação de *Fake News*, com o

[3] Conjunto de dados sobre um determinado tema, que serve de base para estudos.
[4] Métrica que combina precisão e recall em um único valor para mensurar a eficácia de modelos de Aprendizado de Máquina.

foco direcionado para avaliar os impactos da etapa de pré-processamento dos dados na tarefa de classificação. Combinando as técnicas de Regressão Logística e *Embeddings*, o resultado chegou a 96% na acurácia.

Monteiro *et al.* (2018) aplicaram diversas técnicas de processamento de texto por meio da implementação de Aprendizado de Máquina, com *Support Vector Machine* (SVM), para avaliar as *Fake News* do Corpus *Fake.Br*. De acordo com os resultados relatados, os métodos que baseados na abordagem de *"bag of words"* apresentaram os resultados mais promissores. O modelo SVM com *"bag of words"* obteve F1-score de 88% para as classes de notícias verdadeiras e falsas, e a combinação com outras representações não melhoraram o desempenho do modelo.

A partir da revisão bibliográfica foi identificada outra base de notícias, a *FakeTrueBr*, desenvolvida por Chavarro *et al.* (2023). Em sua pesquisa, Chavarro *et al.* (2023) manipularam a *FakeTrueBr,* aplicando *embeddings* para encontrar padrões léxicos entre notícias falsas e verdadeiras, gerando assim uma correlação de dados, na qual, para cada notícia falsa armazenada, existe a sua correspondente verdadeira. O conjunto de dados do Corpus *FakeTrueBr* é constituído de 3582 notícias em Língua Portuguesa, divididos em 5 categorias: Brasil, Política, Entretenimento, Saúde e Mundo. Chavarro *et al.* (2023) também aplicaram técnicas de *"bag of words"* com diferentes modelos de Aprendizado de Máquina e n-gramas[5] alcançando 94.5% de F1-score no melhor resultado.

Diante da revisão bibliográfica dos trabalhos relacionados aos objetivos da pesquisa foi possível identificar as principais estratégias e métodos direcionados à identificação automática de notícias falsas, com foco na aplicação dos algoritmos de Aprendizado de Máquina e técnicas de processamento de texto. A partir dos estudos revisados, no próximo capítulo será apresentada a metodologia deste trabalho, com especificação do pré-processamento de dados e da preparação dos modelos de Aprendizado de Máquina para a classificação de notícias em Língua Portuguesa.

[5] Sequência contígua de itens em um texto, que em Processamento de Linguagem Natural pode ser conjuntos de palavras, sílabas, letras ou até caracteres.

4 METODOLOGIA

Este trabalho consiste de uma pesquisa empírica, com abordagem quantitativa, estruturada com objetivos descritivos e explicativos. Segundo Fonseca (2002), a pesquisa quantitativa se centra na objetividade, considerando a análise de dados brutos, recolhidos com o auxílio de instrumentos padronizados e neutros.

A metodologia desenvolvida no trabalho foi baseada no *Design Science Research*, paradigma de pesquisa pragmático proposto por Hevner *et al.* (2004), na qual a resolução de problemas no mundo real é feita por intermédio da criação e análise de artefatos inovadores. O presente estudo teve como objetivo verificar o desempenho de algoritmos de Aprendizado de Máquina para a classificação de *Fake News*, construindo uma relação entre a precisão de treinamento e a generalização de cada algoritmo para novas notícias analisadas.

Foram selecionados dois algoritmos para a tarefa de classificação das *Fake News*, o algoritmo de Regressão Logística e o modelo *Hierarchical Attention Networks* (HAN), tomando como referência a proposta de classificação de Battisti (2020), que aplicou uma abordagem comparativa entre a Regressão Logística e o modelo de *Long Short-Term Memory* (LSTM).

O desenvolvimento do trabalho foi estruturado em duas etapas: A Etapa 1 foi constituída da revisão bibliográfica e obtenção dos dados que foram utilizados na pesquisa. A Etapa 2 foi caracterizada pela implementação e avaliação dos algoritmos para a classificação das notícias. Na segunda etapa do trabalho também foi executada a avaliação dos modelos com dados externos (conjunto secundário de notícias) para se obter um panorama sobre o desempenho das generalizações dos algoritmos.

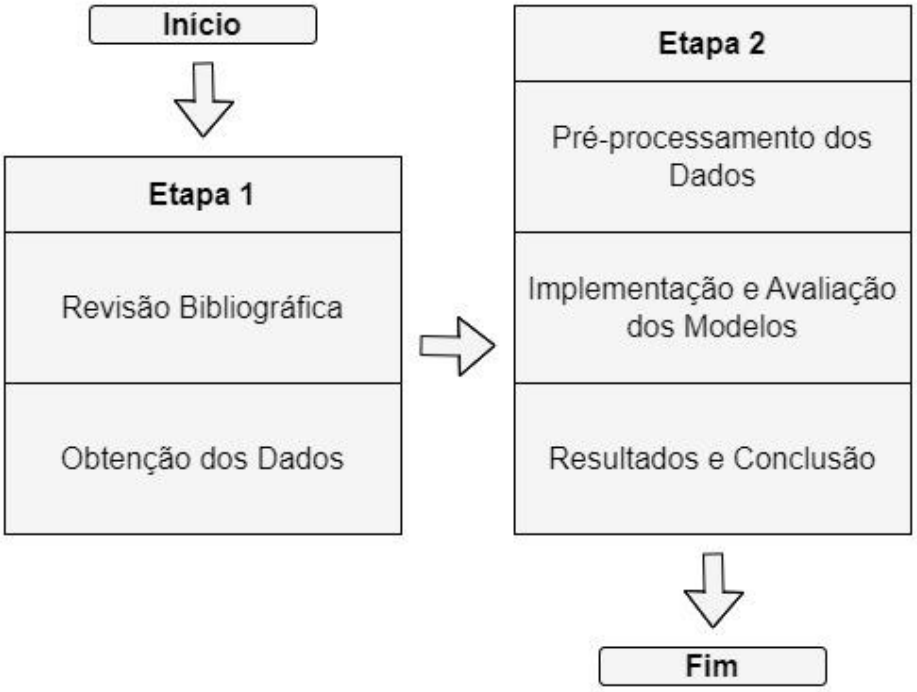
Para a realização dos experimentos foi utilizado Corpus *Fake.br* de Monteiro *et al.* (2018), e uma base extraída do conjuntos de dados de Chavarro *et al.* (2023), o *FakeTrueBr*. Nos *datasets* de trabalho, os dados referentes às notícias falsas e verdadeiras estão separadas por categorias e essas notícias foram pré-processadas antes de serem submetidas no treinamento dos algoritmos para a fase de classificação.

Por fim, os resultados que foram obtidos da classificação foram analisados por meio de uma matriz de confusão, que é um representação gráfica da classificação dos dados, a fim de identificar qual algoritmo obteve o melhor

desempenho, revelando-se a discussão acerca das vantagens e desvantagens de cada algoritmo, conforme os *insigts* da análise dos dados.

Uma visão detalhada da metodologia pode ser observada na Figura 12, na qual se apresenta um diagrama das etapas e a sequência dos procedimentos realizados no desenvolvimento desta pesquisa.

Figura 12 - Diagrama da metodologia proposta.



Fonte: Autoria própria.

4.1 Base de treinamento

A base de dados utilizada neste trabalho foi derivada do conjunto Corpus *Fake.Br*, produzido pelo Núcleo Interinstitucional de Linguística Computacional (NILC). O Corpus original possui uma compilação de 7200 notícias em Português, sendo 3600 notícias verdadeiras e 3600 notícias falsas (Monteiro *et al.,* 2018).

Em relação a qualidade do Corpus *Fake.Br*, essa base de notícias falsas foi estruturada manualmente por Monteiro *et al.* (2018), na qual as notícias falsas foram coletadas de 4 sites de notícias: Diário do Brasil, Folha do Brasil, *The Jornal Brasil*, *Top Five TV*, passando por uma verificação textual para garantir a confiabilidade do laudo de veracidade de cada notícia. Para preencher o restante do conjunto das

notícias verdadeiras, Monteiro *et al.* (2018) coletaram os dados de forma semiautomática das maiores agências de notícias do Brasil no período da pesquisa: G1, Folha de São Paulo e Estadão. As notícias verdadeiras foram filtradas através das palavras chaves contidas no texto e no título de cada *Fake News*, também foi aplicada uma técnica de análise lexical para definir as notícias mais similares das *Fake News*.

Para a estruturação do conjunto de dados principal deste trabalho, as notícias foram analisadas de forma manual para garantir que estavam realmente alinhadas em seus conteúdos. No *dataset* criado, os textos são completos, sem pré-processamento, permitindo a aplicação de novas análises, conforme a demanda ou necessidade de pesquisa. A base *Fake.Br* foi submetida ao processo de organização dos arquivos de textos, para ordenar as notícias aos seus rótulos de identificação. Esses rótulos são marcações numéricas, sendo 1 para identificação das *Fake News* e 0 para as - não *Fake News* - ou seja, notícias verdadeiras.

Tabela 1 - Descrição da base de dados de treinamento.

Categoria	Quantidade de amostras	%
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

Fonte: Adaptado de Monteiro *et al.* (2018).

Conforme apresentado na Tabela 1, a distribuição das categorias não são uniformes, possuindo uma maior quantidade de notícias sobre o tema de política, sendo 4.180 amostras, enquanto as categorias Economia e Religião só possuem 44

amostras. Os dados originais da *Fake.Br* foram coletados a partir do repositório no *Github*[6].

Para o experimento de teste generalista dos modelos, foi utilizada uma amostra dos dados da base *FakeTrueBr* (Chavarro *et al.,* 2023) disponibilizada no Github[7], na qual o conjunto de dados é composto por 1791 notícias falsas e 1791 de suas correspondentes verdadeiras. Na amostra coletada, foram selecionadas 180 notícias, sendo 90 verdadeiras e 90 falsas, separadas manualmente em seis categorias, alinhadas com as categorias originais do Corpus *Fake.Br*.

Na Tabela 2 apresenta-se a quantidade de amostras por categoria selecionadas do *dataset FakeTrueBr*.

Tabela 2 - Subconjunto da base de dados *FakeTrueBr*.

Categoria	**Quantidade de amostras**	**%**
Política	30	16,67
TV e celebridades	30	16,67
Sociedade e notícias diárias	30	16,67
Ciência e tecnologia	30	16,67
Religião	30	16,67
Saúde	30	16,67

Fonte: Autoria própria.

A base de dados obtida de *FakeTrueBr* foi estruturada para possuir notícias verdadeiras e falsas relacionadas à organização de dados do Corpus *Fake.Br*, e dessa forma foi possível prosseguir para a fase de implementação dos modelos de Aprendizado de Máquina.

[6] Repositório do Corpus *Fake.Br*. Disponível em: https://github.com/roneysco/Fake.Br-Corpus, acesso em 10/04/2024.
[7] Repositório da *FakeTrueBr*. Disponível em: https://github.com/jpchav98/FakeTrue.Br, acesso em 10/04/2024.

4.2 Implementação dos modelos

Após a revisão dos trabalhos relacionados e com base nas metodologias de Battisti (2020) e Guarise (2019), foram selecionados dois algoritmos para construir os modelos de classificação das notícias, os algoritmos de Regressão Logística e *Hierarchical Attention Networks* (HAN). A seleção destes algoritmos foi fundamentada através dos melhores resultados apresentados anteriormente pelos autores, também para o processo de detecção de *Fake News*.

Toda a fase de implementação e testes dos modelos foi realizada através da plataforma *Google Colaboratory*, com uso de bibliotecas Python para executar todos os procedimentos de pré-processamento e implementação dos modelos de classificação. As principais bibliotecas utilizadas foram:

- ***Numpy*** - Biblioteca de funções e operações matemáticas para manipulação de arrays multidimensionais de forma otimizada.
- **Pandas -** Biblioteca destinada a manipulação e análise de dados.
- ***Seaborn* -** Biblioteca de visualização de dados baseada em matplotlib, para facilitar a criação de visualizações gráficas.
- ***Matplot* -** Biblioteca para criação de gráficos 2d.
- **NLTK -** Biblioteca para processamento de linguagem natural, possui diversos métodos de manipulação, tais como: tokenização, stemming, lematização, análise sintática, etc.
- ***Sklearn* -** Biblioteca que implementa diversos algoritmos de Aprendizado de Máquina, assim como, de pré-processamento e de avaliação dos modelos.
- **Keras -** Biblioteca que implementa diversos métodos voltados ao Aprendizado Profundo.

4.2.1 Implementação do modelo de Regressão Logística

Para implementar o modelo de Regressão Logística foi necessário executar o pré-processamento dos dados, ajustando-os para a fase de treinamento e validação do algoritmo. Durante o pré-processamento, foi realizada uma limpeza no conjunto de dados, principalmente para eliminar caracteres especiais e acentos das palavras. Além disso, os textos foram convertidos para letras minúsculas, possibilitando um

reconhecimento melhor dos *tokens*. Adicionalmente, foram removidas as chamadas "*stopwords*", utilizando a biblioteca NLTK, visto que essas palavras não exercem influência significativa nas sentenças e sua exclusão contribui para otimizar o desempenho do modelo de classificação. Uma amostra dos dados pré-processados pode ser visualizada na Figura 13.

Figura 13 - Notícia antes e depois do pré-processamento de dados.

Antes:

Ator petista e senador acusado na Lava-Jato protagonizam baixaria através do twitter. No último dia de 2014, o ator José de Abreu foi surpreendido com uma resposta ao ataque que fez ao senador Randolfe Rodrigues (Rede-AP) através do twitter. Abreu criticou o senador, que foi citado em um dos processos da Lava Jato e acusado por um delator de receber R$ 200 mil em propina. "E o Randolfe, hein? Outro hipócrita safado" , escreveu José de Abreu em em twitter. Passado algum tempo, o senador respondeu ao ator petista: "Caro José de Abreu, seu baixo nível já é notório na Rede Globo, que já lhe chamou até a atenção. Não me inclua na sua laia, seu submundo sujo". O ator não ficou calado e continuou o quebra pau com o senador: "Seu Ran [...]

Depois:

ator petista senador acusado lava jato protagonizam baixaria atraves twitter ultimo dia ator jose abreu surpreendido resposta ataque fez senador randolfe rodrigues rede ap atraves twitter abreu criticou senador citado processos lava jato acusado delator receber r mil propina randolfe hein outro hipocrita safado escreveu jose abreu twitter passado algum tempo senador respondeu ator petista caro jose abreu baixo nivel ja notorio rede globo ja chamou ate atencao nao inclua laia submundo sujo ator nao ficou calado continuou quebra pau senador randolfe deixa mentiroso jamais ocorreu contrario veja papel novela h reflita alias randolfe ha anos acusado estado deem google disse jose abreu tambem atacou senador aecio neves psdb mg sorte randolfe aecio [...]

Fonte: Dados da pesquisa.

A representação de vetores para os *tokens* utilizada foi o TF-IDF, que foi treinado na própria base de notícias para calcular a frequência e estimar a relevância das entradas. Os valores TF-IDF foram dimensionados através do valor máximo absoluto, que consiste em ajustar os dados para que o maior valor absoluto em cada vetor se torne 1, dessa maneira garantindo que os termos possam ser comparados de maneira justa e que o modelo de Regressão Logística não seja influenciado por valores TF-IDF muito altos ou baixos. Essa estratégia é utilizada na tentativa de obter uma melhor performance do modelo.

Em relação a partição do conjunto de dados, as amostras da *Fake.Br* foram divididas na proporção de 70% para o treinamento (5040 amostras) e os 30% restante foi dividido meio a meio, num total de 1080 amostras para a validação e 1080 amostras de teste.

O modelo de Regressão Logística utilizado encontra-se na biblioteca *Sklearn* (ou *scikit-learn*), ela oferece uma implementação de fácil utilização, além de ser eficiente, permitindo a aplicação de técnicas avançadas de modelagem de dados de maneira simples.

4.2.2 Implementação do modelo HAN (*Hierarchical Attention Networks*)

Para a implementação do modelo HAN, o pré-processamento dos dados foi concentrado no processo de substituição de alguns caracteres como "-" e "/" por espaços, remoção de caracteres especiais dos textos e outras pontuações (exceto pontuação de fim de sentença), para que durante a tokenização, o texto fosse dividido em sentenças limpas. Por fim, os textos foram completamente convertidos em letras minúsculas.

O modelo de *embeddings* escolhido foi o CBOW, com base na metodologia de Monteiro *et al*. (2018), enquanto os vetores de palavras já treinados para Língua Portuguesa foram obtidos no *NILC-Embeddings*[8]. O pré-processamento dos textos, permitiu que os *tokens* reconhecidos pelo *embedding* após a transformação da base de dados representassem 88% do número total de palavras no conjunto, servindo para minimizar o número de *embeddings* vazias (*tokens* não reconhecidos pelo modelo Word2Vec treinado), permitindo que o modelo reconhecesse o máximo de instâncias possível dentro da base de treinamento.

Ainda na fase de implementação do modelo, as camadas LSTM bidirecionais da biblioteca Keras foram combinadas com uma camada de atenção com contexto[9] que caracterizam as Redes Hierárquicas de Atenção na classificação de documentos.

Por fim, é importante ressaltar um desafio a respeito do treinamento da rede, considerando a limitação de recursos na versão gratuita da plataforma Google Colab, devido ao custo de tempo na geração de *embeddings* para o conjunto de treinamento e no processo de treinamento do modelo, pois dependendo dos parâmetros utilizados, o tempo de execução da máquina virtual extrapola o limite permitido para uso, sendo necessário realocar a máquina virtual, baixar algumas das

[8] Repositório do NILC destinado a armazenar e compartilhar vetores de palavras treinados para Língua Portuguesa. Disponível em: http://nilc.icmc.usp.br/nilc/index.php/repositorio-de-word-embeddings-do-nilc, acesso em 10/04/2024.
[9] https://gist.github.com/cbaziotis/7ef97ccf71cbc14366835198c09809d2.

dependências, importar o *Embedding* e os dados do Google Drive, e então recomeçar o treinamento do modelo.

5 RESULTADOS

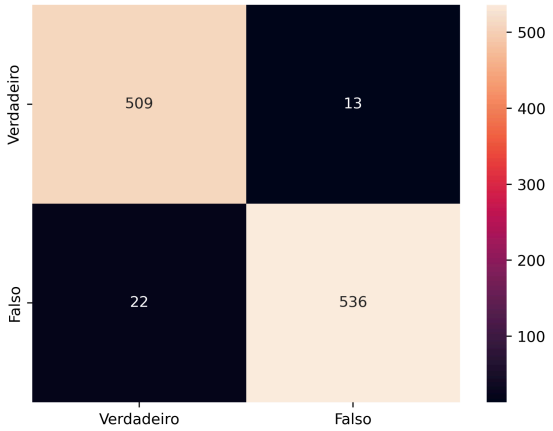
Neste Capítulo serão apresentados e discutidos os resultados obtidos a partir do treinamento e dos testes realizados nos modelos Regressão Logística e HAN em relação a classificação das *Fake News* em Língua Portuguesa. O desempenho dos modelos foi avaliado por meio das métricas de acurácia, precisão, *recall* e F1-*score,* além da construção de matrizes de confusão para detalhar o número de acerto e erros na classificação das notícias.

Na avaliação do modelo de Regressão Logística, foi identificada uma acurácia de 96,76% no conjunto de teste (*Fake.Br*), correspondente a uma classificação correta de 1045 das 1080 notícias avaliadas. Este resultado é comparável aos resultados apresentados por por Battisti (2020) e Silva (2022).

O treinamento do modelo de Regressão Logística, através da plataforma Colab, durou cerca de 5 minutos, incluindo o processo TF-IDF e o treinamento da classificação. A Figura 14 apresenta a Matriz de Confusão do modelo, na qual é possível visualizar que 509 notícias verdadeiras foram classificadas corretamente, assim como 536 notícias falsas foram classificadas corretamente como *Fake News.*

Uma Matriz de Confusão é uma forma gráfica de visualizar o desempenho da classificação de um modelo de Aprendizado de Máquina. Nesta visualização, para uma classificação binária, uma matriz 2x2 é criada e nela os resultados da previsão correta do algoritmo são exibidos na diagonal principal, denominados de Verdadeiros Positivos e Verdadeiros Negativos. Enquanto na diagonal secundária ficam as classificações incorretas, denominadas de Falsos Positivos e Falsos Negativos.

Figura 14 - Matriz de Confusão do modelo de Regressão Logística.



Fonte: Dados da pesquisa.

Em relação às métricas de acurácia e precisão do modelo de Regressão Logística, tem-se o seguinte relatório de valores:

$$\text{Acurácia total} = \frac{509 + 536}{1080} \approx 96{,}76\%$$

$$\text{Precisão de verdadeiros} = \frac{509}{509+13} \approx 97{,}5\%$$

$$\text{Precisão de falsos} = \frac{536}{536+22} \approx 96{,}05\%$$

O modelo de Regressão Logística obteve maior precisão que as encontradas em Battisti (2020), considerando que o autor relatou uma precisão de verdadeiros 92,91% e precisão de falsos 92,63% em sua abordagem de classificação das notícias.

Em relação a avaliação do modelo HAN, foi identificada uma precisão máxima de 97% para a classificação das notícias. O modelo acertou 1049 das 1080 previsões de teste. Ressaltando que o modelo foi treinado com 70% da base de dados *Fake.Br* e o processo de teste foi realizado com a porcentagem reservada dos dados da *Fake.Br*.

O treinamento do modelo HAN foi realizado em 10 épocas, em lotes de 30 entradas. Cada época de treino durou em média 13,7 minutos, totalizando aproximadamente 2 horas e 28 minutos de treinamento. O tempo de treinamento registrado foi mais rápido que o treinamento da abordagem de Guarise (2019), que relatou 115 horas e 12 minutos para treinar um modelo HAN, utilizando uma plataforma dedicada.

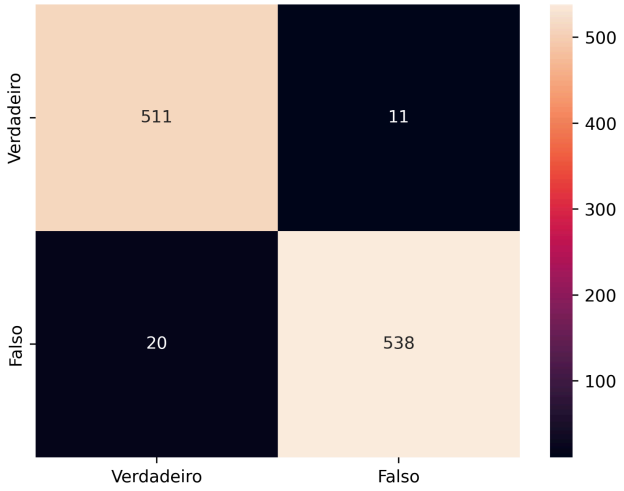
Na Figura 15 apresenta-se a Matriz de Confusão do modelo de HAN, na qual é possível visualizar a classificação correta de 511 notícias verdadeiras e 538 notícias falsas. Em relação às métricas de acurácia e precisão, foram obtidos os seguintes registros:

$$\text{Acurácia total} = \frac{511 + 538}{1080} \approx 97{,}12\%$$

$$\text{Precisão de verdadeiros} = \frac{511}{511+11} \approx 97{,}89\%$$

$$\text{Precisão de falsos} = \frac{538}{538+20} \approx 96{,}41\%$$

Figura 15 - Matriz de confusão do modelo HAN.



Fonte: Dados da pesquisa.

De acordo com o desempenho alcançado pelo modelo HAN, identificou-se que a precisão obtida na classificação de verdadeiros foi maior que a apresentada por Guarise (2019). Considerando que nesta abordagem o modelo obteve 97,89% na precisão de verdadeiros, enquanto Guarise (2019) relatou uma porcentagem de 93,89% nessa categoria.

Os modelos treinados produziram resultados aproximados e através do relatório de classificação (*classification_report)* do *sklearn,* apresentado na Figura 16, é possível observar o valor das métricas de precisão, *recall* e pontuação F1 referente à classificação das notícias.

Figura 16 - Relatório de Classificação dos modelos de Regressão Logística e HAN.

classification_report Regressão Logística

	precision	recall	f1-score	support
0	0.96	0.98	0.97	522
1	0.98	0.96	0.97	558
accuracy			0.97	1080
macro avg	0.97	0.97	0.97	1080
weighted avg	0.97	0.97	0.97	1080

classification_report HAN

	precision	recall	f1-score	support
0	0.96	0.98	0.97	522
1	0.98	0.96	0.97	558
micro avg	0.97	0.97	0.97	1080
macro avg	0.97	0.97	0.97	1080
weighted avg	0.97	0.97	0.97	1080
samples avg	0.97	0.97	0.97	1080

Fonte: Dados da pesquisa.

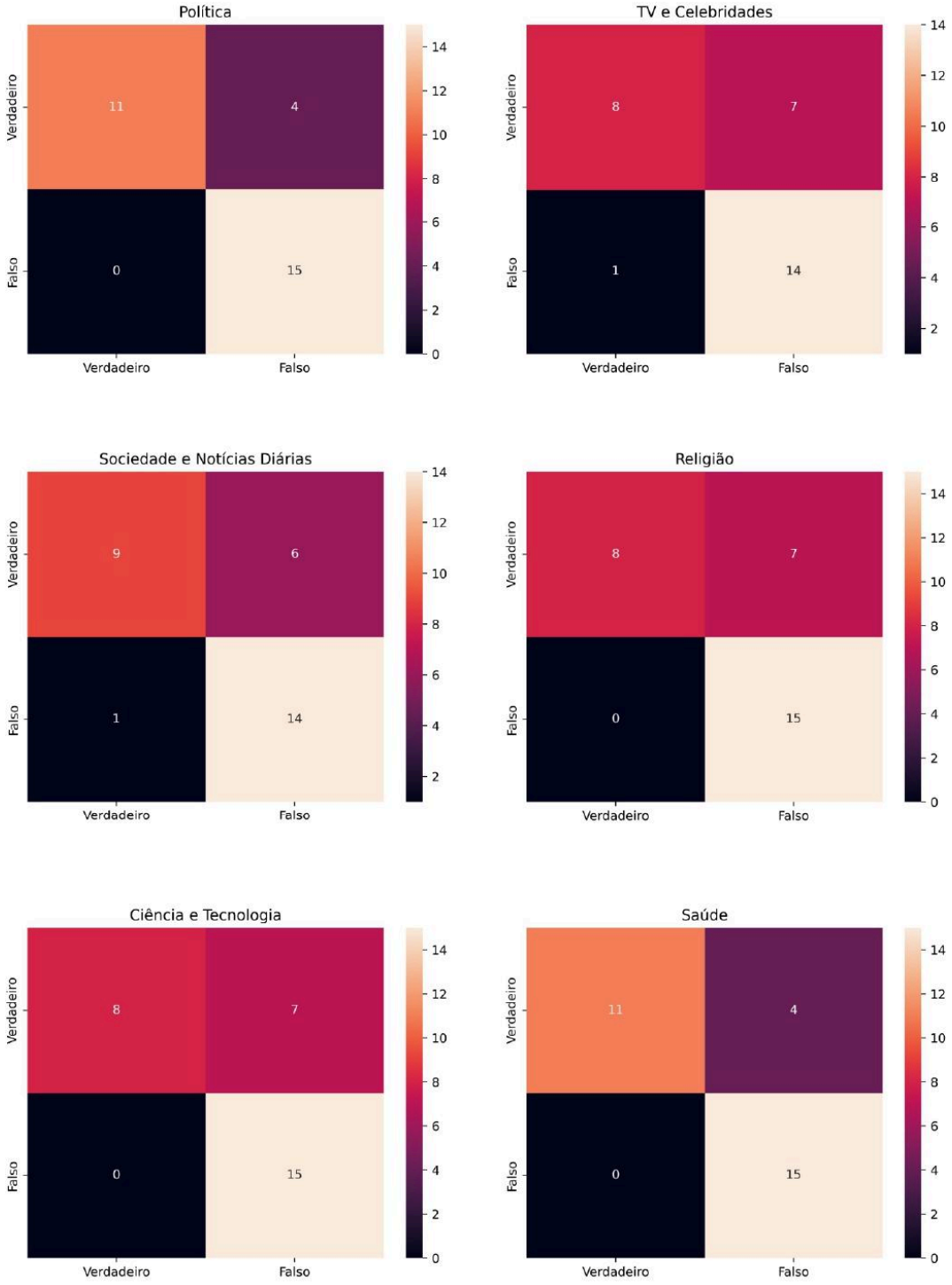
Um ponto importante a ser considerado nos resultados de classificação de notícias é que o número de *tokens* nas notícias verdadeiras é bem maior que o total de *tokens* presentes nas amostras de notícias falsas. E neste aspecto, Monteiro *et al*. (2018) e Silva (2022) tiveram resultados diferentes entre o uso de uma base

truncada e uma base não truncada, evidenciando que a base truncada apresentou uma porcentagem menor na precisão dos algoritmos implementados em cada pesquisa. Entretanto, nesta pesquisa não foi possível realizar experimentos com uma base truncada.

Após o treinamento com os dados da *Fake.Br*, ambos os modelos de classificação foram submetidos ao processo de avaliação com os dados da base *FakeTrueBr*, para o experimento externo de generalização. As matrizes apresentadas nas Figuras 17 e 18 fornecem o registro de desempenho dos modelos, permitindo a identificação de verdadeiros positivos, falsos positivos, verdadeiros negativos e falsos negativos para cada categoria específica.

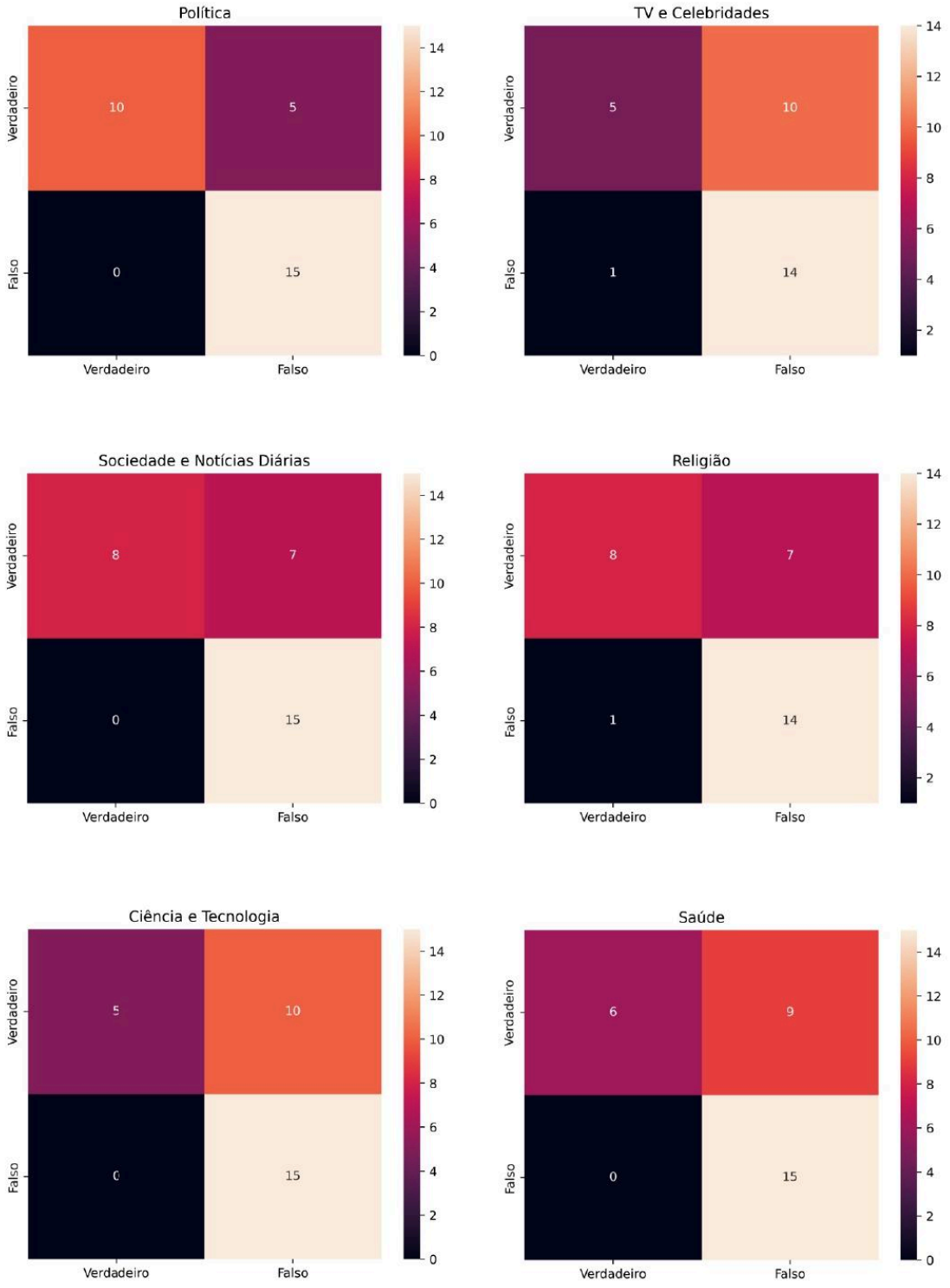
A partir das matrizes de confusão do modelo de Regressão Logística (Figura 17) observa-se que no experimento de generalização o modelo alcançou mais de 90% de acerto para as notícias falsas, no entanto, ocorreu uma alta taxa de falsos negativos para as notícias verdadeiras, fato que também foi verificado nos resultados do modelo HAN (Figura 18). No modelo HAN, o número elevado de falsos negativos ficou mais evidente nas categorias de TV e Celebridades; Ciência e Tecnologia e na categoria de Saúde.

Figura 17 - Matrizes de confusão do modelo Regressão Logística.



Fonte: Dados da pesquisa.

Figura 18 - Matrizes de confusão do modelo HAN.



Fonte: Dados da pesquisa.

No Quadro 1 apresenta-se a acurácia (Acc) dos modelos para cada categoria de notícia, a precisão para a categoria Verdadeiro Positivo (TP) para notícia falsa - e a taxa de Verdadeiro Negativo (TN) para notícia verdadeira.

Quadro 1 - Precisão dos modelos por categoria de notícia.

		Resultados (%)					
		Regressão Logística			HAN		
Categoria	Total de Amostras	Acc	TP	TN	Acc	TP	TN
Política	30	86,67	100,00	73,33	83,33	100,00	66,67
TV e celebridades	30	73,33	93,33	53,33	63,33	93,33	33,33
Sociedade e notícias diárias	30	76,67	93,33	60,00	76,67	100,00	53,33
Ciência e tecnologia	30	76,67	100,00	53,33	66,68	100,00	33,33
Religião	30	76,67	100,00	53,33	73,33	93,33	53,33
Saúde	30	86,67	100,00	73,33	70,00	100,00	40,00
Total	180	79,44	97,78	61,11	72,33	97,78	46,67

Fonte: Dados da pesquisa.

De acordo com as métricas de avaliação, o modelo de Regressão Logística apresentou melhor desempenho na tarefa de classificação de notícias do que o modelo HAN, entretanto ambos os modelos tiveram dificuldades em classificar as notícias verdadeiras. Na classificação de notícias verdadeiras, o desempenho do modelo HAN foi inferior ao modelo de Regressão Logística em todas as categorias.

Outro aspecto observado nos resultados da Classificação é que o modelo de Regressão Logística teve a mesma precisão de classificação para as categorias Política e Saúde, mesmo que o treinamento do modelo tenha sido realizado com uma quantidade menor de amostras para a classe Saúde.

A Figura 19 apresenta uma amostra dos dados utilizados no experimento de generalização. Mesmo com propósitos diferentes, a contraparte tem como objetivo desmentir a notícia falsa. Sendo assim, a contraparte pode possuir partes da notícia falsa dentro de seu conteúdo, e este é um fator crítico na classificação de notícias, pois amostras podem ser classificadas erroneamente pela similaridade textual. No

entanto, esse problema pode ser solucionado com um treinamento mais robusto de noticiais, para que os modelos de Aprendizado de Máquina consigam detectar de forma aprimorada as incorporações de texto.

Figura 19 - Amostra de notícia da base de avaliação.

Falsa

estes chineses são demais.! isto é uma boneca. esta é uma boneca android fabricada na china. é tão real o que a próxima [...]

Contraparte Verdadeira

tem bombado nas redes sociais um vídeo com imagens de bebês acompanhado de uma legenda que diz esta é uma boneca android fabricada na china [...]

Fonte: Dados da pesquisa.

A presença de palavras de cunho sensacionalista pode ser um termômetro da classificação da notícia como *Fake News,* conforme Guarise (2019) ressalta, pois essas palavras denotam o impacto do contexto para a classificação da notícia. A análise restrita de palavras sensacionalistas pode ser uma estratégia útil para aprimorar o processo de classificação das notícias falsas, e este procedimento pode ser adotado como uma etapa de extração de recursos para aperfeiçoar o treinamento dos classificadores.

Considerando a avaliação realizada neste trabalho, é possível afirmar que o modelo de Regressão Logística teve um desempenho melhor para a classificação de notícias, tanto para a categoria de notícias falsas, quanto na precisão geral do conjunto. Nos testes de generalização com dados da base *FakeTrueBr*, o modelo HAN obteve um resultado de acurácia aproximada aos resultados de Guarise (2019), que também realizou experimento semelhante com a arquitetura HAN e obteve 73,33% de acurácia.

Para trabalhos futuros ressalta-se a importância de adaptar o conjunto de notícias para treinamento dos algoritmos, assim como desenvolver novos ajustes para a configuração do modelo HAN. Evidencia-se também a pretensão de

investigar o desempenho de outras arquiteturas de Redes Neurais Artificiais para o processamento de texto.

5.1 Limitações da Pesquisa

Dentre as principais limitações desta pesquisa, relata-se a dificuldade de encontrar bases de dados que atendessem aos objetivos do trabalho, além do desafio de lidar com o pré-processamento das notícias para desenvolver o aprendizado dos algoritmos.

Também ressalta-se a complexidade no treinamento da Rede Neural do modelo HAN, considerando a limitação do recurso computacional no uso da versão básica do *Google Colaboratory*, pois o serviço é limitado para o processamento nas máquinas virtuais, quando não está definida uma assinatura *premium*. Dessa forma, o processo de treinamento dos algoritmos, incluindo o carregamento da base de dados, carregamento dos *embeddings*, *tokens* e demais procedimentos precisou ser reiniciado a cada vez que o tempo de execução expirava, ampliando o tempo de desenvolvimento dos modelos de classificação.

Por fim, declara-se a importância de desenvolver em trabalhos futuros estratégias para aprimorar a abordagem de detecção de notícias falsas, considerando os resultados alcançados nesta pesquisa e superar as limitações da metodologia aplicada.

6 CONCLUSÃO

O presente trabalho teve como propósito avaliar algoritmos de Aprendizado de Máquina aplicados ao processo de detecção das *Fake News*. Os objetivos de pesquisa foram alcançados a partir dos resultados obtidos com o treinamento e avaliação dos modelos de classificação, com destaque para o modelo de Regressão Logística que apresentou o melhor desempenho no processo de identificação das *notícias*.

De acordo com a avaliação dos modelos de classificação, identificou-se que o modelo de Regressão Logística obteve melhor desempenho, alcançando uma acurácia total de 79,44% contra 72,33% do modelo HAN, no experimento de generalização da base *FakeTrueBr*.

Em observação às categorias de Política e Saúde, o algoritmo de Regressão Logística atingiu uma acurácia de 86,67%. Nessas categorias, ambos os modelos de classificação alcançaram 100% de precisão para identificar as notícias falsas (TP). No entanto, a precisão para identificar notícias verdadeiras (TN) é relativamente maior para o modelo de Regressão Logística.

As falhas de classificação dos modelos em relação às notícias verdadeiras podem estar relacionadas às especificidades da base de dados, mas esse comportamento de erro dos classificadores precisa ser melhor avaliado em futuros experimentos, com um conjunto balanceado de notícias para as classes consideradas.

6.1 Trabalhos futuros

Os resultados apresentados nesta pesquisa proporcionam direcionamentos para novos estudos acerca das notícias falsas. Torna-se importante compreender em novas análises como o pré-processamento de texto pode ser adaptado para melhorar a precisão de classificação, buscando identificar os aspectos dos dados que são capazes de aprimorar a generalização dos classificadores.

Declara-se também a necessidade de aprimorar as atuais bases de dados de notícias para o treinamento de algoritmos de Aprendizado de Máquina e Aprendizado Profundo. Por fim, pretendem desenvolver em futuros estudos a implementação de outras arquiteturas de Redes Neurais e outras técnicas de

Embeddings para viabilizar uma abordagem mais aprimorada para a identificação e classificação das *Fake News.*

REFERÊNCIAS

AGGARWAL, C. C. **Neural Networks and Deep Learning : A Textbook**. Cham: Springer International Publishing, 2018.

AKDOGAN, A. Word Embedding Techniques: Word2Vec and TF-IDF Explained. **Towards Data Science [online]**, [S. l.], jul. 2021. Disponível em: <https://towardsdatascience.com/word-embedding-techniques-word2vec-and-tf-idf-explained-c5d02e34d08>. Acesso em 15 de out. 2023.

ALLCOTT, H.; GENTZKOW, M. Social Media and Fake News in the 2016 Election. **Journal of Economic Perspectives**, [S. l.], v. 31, n. 2, p. 211–236, mai. 2017. Disponível em: https://www.aeaweb.org/articles?id=10.1257/jep.31.2.211. Acesso em: 17 mai. 2024.

BAHDANAU, D.; CHO, K.; BENGIO, Y. Neural machine translation by jointly learning to align and translate. In: International Conference on Learning Representations, 2014, Banff. **Proceedings eletrônicos** [...]. Banff, Canada: ICLR, 2014. Disponível em: https://arxiv.org/abs/1409.0473. Acesso em: 18 out. 2023.

BARCELOS, T. N. et al. Análise de *Fake News* veiculadas durante a pandemia de COVID-19 no Brasil. **Revista Panamericana de Salud Pública**, [S. l.], v. 45, 2021. Disponível em: https://www.scielosp.org/article/rpsp/2021.v45/e65/pt/. Acesso em 3 de mar. 2024.

BATTISTI, Fernando. **Detecção de notícias falsas utilizando aprendizado de máquina**. TCC (graduação) - Centro Tecnológico, Universidade Federal de Santa Catarina, Florianópolis, 2020.

BRAGA, R. M. A indústria das *Fake News* e o discurso de ódio. In: PEREIRA, R. V. (Org.). **Direitos políticos, liberdade de expressão e discurso de ódio.** 1. vol. Belo Horizonte: Instituto para o Desenvolvimento Democrático, p. 203-220, 2018.

CARVALHO, Lucas Borges de. A democracia frustrada: *Fake News*, política e liberdade de expressão nas redes sociais. **Revista Internet & Sociedade**, [S.l.], v. 1, n. 1, 2020, p. 172-179. Disponível em: https://revista.internetlab.org.br/a-democracia-frustrada-fake-news-politica-e-liberdade-de-expressao-nas-redes-sociais. Acesso em: 17 mai. 2024.

CHAVARRO, J. P. et al. FakeTrueBr: Um corpus brasileiro de notícias falsas. In: XVIII Escola Regional De Banco De Dados (ERBD), 2023, Palmas. **Anais [...]**. [S.l.]: Sociedade Brasileira de Computação, 2023, p. 108-117. Disponível em: https://sol.sbc.org.br/index.php/erbd/article/view/24352. Acesso em: 17 mai. 2024.

CIAMPAGLIA, G. L. et al. Computational fact checking from knowledge networks. **PloS one**, [S. l.], v. 10, n. 6, jun. 2015. Disponível em: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0128193. Acesso em: 2 mar. 2024.

DOURADO, Tatiana. ***Fake News* Na Eleição Presidencial De 2018 No Brasil**. 2020. Tese (Programa de Pós-Graduação em Comunicação e Cultura Contemporâneas) - Faculdade de Comunicação, Universidade Federal da Bahia, Salvador, 2020.

Faceli, K. et al. **Inteligência artificial: uma abordagem de aprendizado de máquina**. 2. ed. [S. l.]: LTC, 2021.

FONSECA, J. J. S. **Metodologia da pesquisa científica**. Fortaleza: UEC, 2002.

GALASSI, A.; LIPPI, M.; TORRONI, P. Attention in Natural Language Processing. **IEEE Transactions On Neural Networks And Learning Systems**, [S. l.], v. 32, n. 10, p. 4291-4308, out. 2021. Disponível em: https://www.researchgate.net/publication/344203247_Attention_in_Natural_Language_Processing. Acesso em: 17 mai. 2024.

GÉRON, A. **Mãos à obra Aprendizado de Máquina com Scikit-Learn & TensorFlow**: Conceitos, Ferramentas e Técnicas Para a Construção de Sistemas Inteligentes. 1. ed. Rio de Janeiro: Alta Books, 2019.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. **Deep Learning**. 1. ed. [S. l.]: MIT Press, 2016.

GUARISE, Lucas. **Detecção de notícias falsas usando técnicas de Deep Learning**. 2019. Monografia (Bacharel em Engenharia de Computação) – Instituto De Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2019.

HAYKIN, S. **Redes Neurais:** Princípios e Prática. 2. ed. Porto Alegre: Bookman, 2001.

HEVNER, A. R.; MARCH, S. T.; PARK, J.; RAM, S. **Design Science in Information Systems Research**. MIS Quarterly, vol. 28, n. 1, pp. 75-105. 2004.

HILL, F. et al. Learning to understand phrases by embedding the dictionary. **Transactions of the Association for Computational Linguistics**, Massachusetts, v. 4, p. 17–30, dez. 2016. Disponível em: https://aclanthology.org/Q16-1002/. Acesso em: 17 mai. 2024.

JAMES, G. et al. **An introduction to statistical learning: with applications in Python**. 1. ed. [S. l.]: Springer, 2023.

MARQUES, R.; RAIMUNDO, J. A. O negacionismo científico refletido na pandemia da COVID-19. **Boletim de Conjuntura (BOCA)**, Boa Vista, v. 7, n. 20, p. 67–78, ago. 2021. Disponível em: https://revista.ioles.com.br/boca/index.php/revista/article/view/410. Acesso em: 17 mai. 2024.

MIKOLOV, T. et al. Efficient estimation of word representations in vector space. In: International Conference on Learning Representations, 2013, Arizona. **Proceedings**

eletrônicos [...]. Scottsdale, Arizona: ICLR, 2013. Disponível em: https://arxiv.org/abs/1301.3781. Acesso em: 17 mai. 2024.

MONTEIRO, R. A. et al. Contributions to the study of Fake News in Portuguese: New corpus and automatic detection results. **Lecture Notes in Computer Science**. Cham: Springer International Publishing, v. 11122, p. 324-334, ago. 2018.

NIELSEN, Michael A. **Neural Networks and Deep Learning**. [S. l.]: Determination Press, 2015, E-book.

OLIVEIRA, Leandro. **Inteligência Artificial aplicada à detecção de *Fake News***. 2019. Monografia (Bacharel em Engenharia da Computação) - Centro de Ciências Exatas e Tecnologia, Universidade Federal do Maranhão, São Luís, 2019.

Quadrado, Jaqueline; Ferreira, Ewerton. Ódio e intolerância nas redes sociais digitais. **Revista Katálysis [online]**, [S. l.], v. 23, n. 03, p. 419-428, dez. 2020. Disponível em: https://www.scielo.br/j/rk/a/3LNyLswf9rkhDStZ9v4YT3H/. Acesso em: 2 mar. 2024.

SILVA, Denis de Padua. **Impacto da etapa de pré-processamento na classificação de *Fake News***. 2022. Monografia (Bacharel em Ciência da Computação) - Instituto de Ciência de Tecnologia, Universidade Federal de São Paulo, São José dos Campos, 2022.

SILVA, G. M. et al. Desafios da imunização contra COVID-19 na saúde pública: das *Fake News* à hesitação vacinal. **Ciência & Saúde Coletiva**, [S. l.], v. 28, n. 3, p. 739-748, mar. 2023. Disponível em: https://www.scielo.br/j/csc/a/dVVfKrCWD7sPp8TNp8xcngN/abstract/?lang=pt. Acesso em: 17 mar. 2024.

SIVEK, S. C. Both facts and feelings: Emotion and news literacy. **Journal of Media Literacy Education,** [S. l.], v. 10, n. 2, p. 123–138, ago. 2018. Disponível em: https://digitalcommons.uri.edu/jmle/vol10/iss2/7. Acesso em: Acesso em: 24 mar. de 2022.

SPINELLI, E. M.; SANTOS, Jéssica de A. JORNALISMO NA ERA DA PÓS-VERDADE: *fact-checking* como ferramenta de combate às *Fake News*. **Revista Observatório**, [S. l.], v. 4, n. 3, p. 759-782, abr. 2018. Disponível em: https://sistemas.uft.edu.br/periodicos/index.php/observatorio/article/view/4629. Acesso em: 16 mai. 2024.

TEIXEIRA, A.; SANTOS, R. C. *Fake News* colocam a vida em risco: a polêmica da campanha de vacinação contra a febre amarela no Brasil. RECIIS - **Revista Eletrônica de Comunicação, Informação & Inovação em Saúde**, [S. l.], v. 14, n. 1, p. 72-89, mar. 2020. Disponível em: https://www.reciis.icict.fiocruz.br/index.php/reciis/article/view/1979. Acesso em: 16 mai. 2024.

VOSOUGHI, S.; ROY, D.; ARAL, S. The spread of true and false news online. **Science**, Nova York, v. 359, n. 6380, p. 1146–1151, mar. 2018. Disponível em:

https://www.science.org/doi/10.1126/science.aap9559?url_ver=Z39.88-2003&rfr_id=ori:rid:crossref.org&rfr_dat=cr_pub%20%200pubmed. Acesso em: 16 mai. 2024.

YANG, Z. et al. Hierarchical attention networks for document classification. *In*: Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2016, Califórnia. **Proceedings eletrônicos** [...]. San Diego, Califórnia: Association for Computational Linguistics, jun. 2016. p. 1480–1489. Disponível em: https://aclanthology.org/N16-1174/. Acesso em: 16 mai. 2024.

ZANATTA, E. et al. Fake News: the impact of the internet on population health. **Revista da Associação Médica Brasileira**, [S. l.], v. 67, n. 7, p. 926-930, jul. 2021. Disponível em: https://www.scielo.br/j/ramb/a/KwCzQCqPkYwdKHYgkzrXPtb/. Acesso em: 16 mai. 2024.

ZHANG, A. et al. **Dive into Deep Learning**. Cambridge, Inglaterra: Cambridge University Press, 2023.