



**UFOP**

Universidade Federal  
de Ouro Preto

**Universidade Federal de Ouro Preto  
Instituto de Ciências Exatas e Aplicadas  
Departamento de Computação e Sistemas**

**Análise de notícias falsas em rede  
social: Uma abordagem utilizando  
transferência de aprendizagem e  
*Transformers***

**Wagner Bianchini Narde**

**TRABALHO DE  
CONCLUSÃO DE CURSO**

**ORIENTAÇÃO:**  
Luiz Carlos Bambirra Torres

**Abril, 2021  
João Monlevade–MG**

**Wagner Bianchini Narde**

**Análise de notícias falsas em rede social: Uma abordagem utilizando transferência de aprendizagem e *Transformers***

Orientador: Luiz Carlos Bambera Torres

Monografia apresentada ao curso de Engenharia de Computação do Instituto de Ciências Exatas e Aplicadas, da Universidade Federal de Ouro Preto, como requisito parcial para aprovação na Disciplina “Trabalho de Conclusão de Curso II”.

**Universidade Federal de Ouro Preto**

**João Monlevade**

**Abril de 2021**

## SISBIN - SISTEMA DE BIBLIOTECAS E INFORMAÇÃO

N223a Narde, Wagner Bianchini .

Análise de notícias falsas em rede social [manuscrito]: uma abordagem utilizando transferência de aprendizagem e transformers. / Wagner Bianchini Narde. - 2021.

46 f.: il.: color., tab..

Orientador: Prof. Dr. Luiz Carlos Bambera Torres.

Monografia (Bacharelado). Universidade Federal de Ouro Preto. Instituto de Ciências Exatas e Aplicadas. Graduação em Engenharia de Computação .

1. Aprendizado de máquina. 2. Aprendizagem por transferência (aprendizagem automática). 3. Confiabilidade. 4. Notícias falsas. 5. Verificação (lógica). I. Torres, Luiz Carlos Bambera. II. Universidade Federal de Ouro Preto. III. Título.

CDU 004.85

Bibliotecário(a) Responsável: Flavia Reis - CRB6-2431



## FOLHA DE APROVAÇÃO

**Wagner Bianchini Narde**

**Análise de notícias falsas em rede social: Uma abordagem utilizando transferência de aprendizagem e Transformers**

Monografia apresentada ao Curso de Engenharia de Computação da Universidade Federal de Ouro Preto como requisito parcial para obtenção do título de Bacharel em Engenharia de Computação

Aprovada em 27 de abril de 2021

### Membros da banca

Doutor - Luiz Carlos Bambirra Torres - Orientador (Universidade Federal de Ouro Preto)  
Mestre - Alexandre Magno de Souza - (Universidade Federal de Ouro Preto)  
Mestre - Carlos Henrique Gome Ferreira - (Universidade Federal de Ouro Preto)  
Doutor - Gustavo Rodrigues Lacerda Silva - (Centro Universitário UNA)

Luiz Carlos Bambirra Torres, orientador do trabalho, aprovou a versão final e autorizou seu depósito na Biblioteca Digital de Trabalhos de Conclusão de Curso da UFOP em 12/05/2021.



Documento assinado eletronicamente por **Luiz Carlos Bambirra Torres, PROFESSOR DE MAGISTERIO SUPERIOR**, em 12/05/2021, às 23:08, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do [Decreto nº 8.539, de 8 de outubro de 2015](#).



A autenticidade deste documento pode ser conferida no site [http://sei.ufop.br/sei/controlador\\_externo.php?acao=documento\\_conferir&id\\_orgao\\_acesso\\_externo=0](http://sei.ufop.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0), informando o código verificador **0170530** e o código CRC **89F6BBD7**.

*Este trabalho é dedicado para maior honra e glória de Nosso Senhor Jesus Cristo.*

# Agradecimentos

Agradeço a Nossa Senhora pela maternal proteção e auxílio para solucionar os problemas que pareciam estar acima da própria capacidade. Agradeço a minha família pelo apoio incondicional. E a minha namorada por toda ajuda, principalmente na reta final desse trabalho.

*“Se a verdade é relativa, é relativa em relação a quê?”*

— G. K. Chesterton (1874 – 1936),  
*in: Daily News, 02 Jun 1906.*

# Resumo

Notícias falsas se propagam mais rápido no Twitter do que notícias comuns, elas podem desde influenciar o resultado de uma eleição a até causar mortes por tratamentos incorretos de doenças. Este trabalho tem por objetivo principal utilizar métodos baseados em transferência de aprendizagem para aprendizado de máquina e a biblioteca *Transformers* para classificação de *tweets* em verdadeiro e falso. Para isso primeiro é revisada a teoria a respeito das causas e consequências das notícias falsas, tecnologias empregadas e trabalhos relacionados. Elaborada uma base de dados rotulada de forma confiável a partir de postagens extraídas do *Twitter*. Treina-se os modelos *Natural Language Processing (NLP)* de classificação baseados em transferência de aprendizado para o idioma português e avalia-os. O modelo de *Bidirectional Encoder Representations from Transformers (BERT)* ajustado a partir do pré treinamento de Souza, Nogueira e Lotufo (2020) obteve o melhor resultado, conseguindo acurácia de 95,1% de acurácia e 95,4% de F1 para verificação de autenticidade de postagens e 94,4% de acurácia e 94,6% de F1 para detecção de notícias falsas. Até o momento da apresentação deste trabalho não foram encontrados resultados superiores na classificação de *tweets* em português, sendo a principal contribuição deste trabalho um modelo no estado da arte para o problema. Além disso, esse trabalho serve de apoio para pesquisas em português que busquem utilizar os modelos disponíveis na biblioteca *transformers* para *NLP* em categorização de textos.

**Palavras-chaves:** *Fake news*. Desinformação. Verificação. *Machine learning*. *Transfer learning*.



# Abstract

False news spreads faster on Twitter than common news, they can from influencing the outcome of an election to even cause deaths for incorrect treatment diseases. This work aims to use methods based on learning transfer for machine learning and librariesformers for rating of tweets in true and false. For this first is revised the theory Arespeito of the causes and consequences of false news, employed technologies and related works. Elaborated a database reliably labeled from deposts extracted dotwitter. Train the Natural Language Processing (NLP) models based on learning transfer for the Portuguese language and evaluate them. The bidirectional model Representations from Transformers (BERT) adjusted from the pre-training of Souza, Nogueira and Lotufo (2020) obtained the improvement, achieving an accuracy of 95.1% of accuracy and 95.4% of F1 for verification of Posts and 94.4% accuracy and 94.6% F1 for false news detection. Until the presentation of this work no higher results were found in the classification of Tweets in Portuguese, being the main contribution of this work a model in the state of art for the problem. In addition, this work serves as support for Portuguese surveys that seek to use the models available in the Transformers library to NLP in categorization of texts.

**Key-words:** Fake-News. Misinformation. Verification. Machine learning. Transfer learning.

# Lista de ilustrações

Figura 1 – Fluxograma de atividades . . . . .	17
Figura 2 – Exemplo de um <i>tweet</i> . . . . .	28
Figura 3 – Nuvem de palavras da base de dados . . . . .	29

# Lista de tabelas

Tabela 1 – Comparação do número de amostras. . . . .	29
Tabela 2 – Configurações e Hiper-parâmetros de treinamento . . . . .	33
Tabela 3 – Resultados ELECTRA (uncased) com pré-treinamento em Português .	37
Tabela 4 – Resultados RoBERTa com pré-treinamento em Português . . . . .	37
Tabela 5 – Resultados XLM-R com pré-treinamento Multi-idioma . . . . .	38
Tabela 6 – Resultados BERT com pré-treinamento Multi-idioma . . . . .	38
Tabela 7 – Resultados BERT com pré-treinamento em Português . . . . .	39
Tabela 8 – Comparação com Cordeiro (2019) . . . . .	39

# Lista de abreviaturas e siglas

**NLP** *Natural Language Processing*

**COVID-19** *Corona Virus Disease 2019*

**GOT** *Get Old Tweets*

**NLTK** *Natural Language ToolKit*

**BoW** *Bag of Words*

**KNN** *K-Nearest Neighbors*

**SVM** *Support Vector Machines*

**BERT** *Bidirectional Encoder Representations from Transformers*

**RoBERTa** *Robustly optimized BERT approach*

**XLNet** *Crosslingual Language Model - RoBERTa*

**Colab** *Google Colaboratory*

**ONG** *Organização não Governamental*

**ONU** *Organização das Nações Unidas*

**OEA** *Organização dos Estados Americanos*

**EFF** *Electronic Frontier Foundation*

**CSV** *Comma Separated Values*

**UTF-8** *8-bit Unicode Transformation Format*

**UOL** *Universo Online*

**MCC** *Matthews correlation coefficient*

**sklearn** *SciKit-learn*

**RSF** *Repórteres Sem Fronteiras*

**STF** *Supremo Tribunal Federal*

**PL** *Projeto de Lei*

**IA** Inteligência Artificial

**MCC** *Matthews Correlation Coefficient*

**UFMG** Universidade Federal de Minas Gerais

**GPU** *Graphics Processing Unit*

# Sumário

<b>1</b>	<b>INTRODUÇÃO</b>	<b>15</b>
1.1	Elaboração do capítulo	15
1.2	O problema de pesquisa	15
1.3	Objetivos	16
1.4	Metodologia	17
1.5	Resultados e contribuições	18
1.6	Organização do trabalho	18
<b>2</b>	<b>REVISÃO BIBLIOGRÁFICA</b>	<b>19</b>
2.1	<b>A tríade calamitosa da informação</b>	<b>19</b>
2.1.1	Notícias falsas, também conhecidas como <i>Fake News</i>	19
2.1.1.1	Tipologia das <i>Fake News</i>	19
2.1.2	Desinformação, também conhecida como <i>Misinformation</i>	20
2.1.3	Pós-verdade, também conhecida como <i>Post-truth</i>	20
2.1.3.1	Negação da Ciência	20
2.1.3.2	Predisposição Cognitiva	21
2.1.3.3	Declínio da Mídia Tradicional	21
2.1.3.4	Ascensão das Redes Sociais	21
2.1.3.5	Relativismo do Pós-modernismo	21
2.1.4	Leituras complementares	21
2.2	<b>Tecnologia na Informação</b>	<b>22</b>
2.3	<b>Privacidade e liberdade de expressão</b>	<b>23</b>
2.4	<b>Trabalhos Relacionados</b>	<b>24</b>
<b>3</b>	<b>DESENVOLVIMENTO</b>	<b>27</b>
3.1	<b>Base de dados: definição e procedimentos</b>	<b>27</b>
3.1.1	Coleta de dados	27
3.1.2	Preparação dos dados	28
3.2	<b>Escolha dos modelos</b>	<b>30</b>
3.3	<b>Implementação dos modelos</b>	<b>31</b>
3.3.1	Configuração do Ambiente	31
3.3.2	Preparação dos dados	31
3.3.3	Configurações, hiper-parâmetros de treinamento e métricas de avaliação	32
3.3.4	Validação Cruzada, Treinamento e Avaliação de épocas	33
<b>4</b>	<b>RESULTADOS</b>	<b>35</b>

4.1	Métricas . . . . .	35
4.2	ELECTRA <i>uncased</i> pré-treinado em Português . . . . .	36
4.3	RoBERTa pré-treinado em Português . . . . .	36
4.4	XLM-R pré-treinado em multi-idíomas . . . . .	37
4.5	BERT pré-treinado em multi-idíomas . . . . .	38
4.6	BERT pré-treinado em Português . . . . .	38
4.7	Melhor modelo . . . . .	39
5	CONCLUSÃO . . . . .	40
	REFERÊNCIAS . . . . .	41

# 1 Introdução

Os veículos de notícia de grande circulação por muito tempo foram jornais, revistas, rádio e televisão. Entretanto, com um alcance global, a Internet assumiu um papel muito importante na circulação de notícias. Essas notícias circulam através de vídeos no Youtube, portais de notícia como o Universo Online (UOL) e também em redes sociais como Facebook, Twitter e WhatsApp. Hoje em dia, segundo [Shearer e Mitchell \(2021\)](#), 53% dos americanos utilizam redes sociais como veículo de informação com alta e moderada frequência.

Enquanto nos veículos noticiários legados grande parte das notícias eram construídas por jornalistas, a popularização da Internet permitiu que usuários comuns também escrevessem notícias, sem passar portanto por nenhuma verificação. Com isso, a qualidade das notícias em circulação foi afetada, aumentando o número de notícias que promovem desinformação ou espalham informações falsas.

A acessibilidade da Internet tornou enorme o número de informações compartilhadas em redes sociais. Surgiu então a necessidade de se construir ferramentas capazes de classificar textos, para descobrir se a informação que ele carrega é verdadeira ou falsa. Sites como Facebook e Twitter trabalham para evoluir no sentido em que as informações que estão sendo lidas e compartilhadas nas redes sociais possam ser o mais próximo possível dos fatos.

## 1.1 Elaboração do capítulo

Esta monografia aplica técnicas modernas de processamento de linguagem natural ou NLP e aprendizagem de máquina para classificar postagens extraídas do Twitter em notícias verdadeiras ou falsas. Como o número de postagens em redes sociais é muito elevado, torna-se humanamente inviável de serem verificadas manualmente, fazendo-se necessário um modo de verificação computacional. Para este trabalho, foram ajustados modelos de BERT, *Robustly optimized BERT approach* (RoBERTa), *Crosslingual Language Model - RoBERTa* (XLM-R) e ELECTRA, para então escolher a melhor opção.

## 1.2 O problema de pesquisa

As notícias falsas podem causar inúmeras consequências, [Bovet e Makse \(2019\)](#) analisaram a influência de notícias falsas no Twitter durante as eleições presidenciais de 2016. [Islam et al. \(07 Oct. 2020\)](#) analisaram impactos da desinformação sobre *Corona Virus*



*Disease 2019* ([COVID-19](#)) em vários países, sendo que na Turquia chegaram a morrer cerca de 800 pessoas e 5876 foram hospitalizadas devido a um rumor que afirmava que a ingestão de metanol curava [COVID-19](#).

Segundo [Vosoughi, Roy e Aral \(2018\)](#), notícias falsas se propagam 70% mais rápido no Twitter. Com isso, torna-se um problema até mesmo corrigir seus efeitos mesmo depois de ter descoberto uma postagem falsa, pois a correção não consegue alcançar a velocidade de disseminação da notícia falsa. Este fato torna crítica a necessidade de descobrir uma postagem falsa o mais cedo possível, pois o dano será bem menor, já que a postagem se espalha exponencialmente com o tempo. Apesar de todos problemas causados pelas notícias falsas, é inviável colocar pessoas para ficarem checando cada postagem em uma rede social para averiguar se trata-se de informação verdadeira ou falsa. Por isso, é urgente a necessidade de ferramentas capazes de classificar informações de postagens de forma computacional. Mais carente ainda observamos que é tal processamento no idioma português, esse trabalho busca dar alguns passos para preencher esta lacuna no idioma português.

### 1.3 Objetivos

O presente trabalho tem como objetivo geral elaborar um classificador de *tweets* em verdadeiros ou falsos baseado em transferência de aprendizagem para o idioma português. Envolve a revisão da literatura para investigar causas e efeitos das notícias falsas, construção de uma base de dados para treinamento e avaliação de uma ferramenta capaz de analisar de forma computacional se a mensagem de um *tweet* é verdadeira ou falsa. Este trabalho apenas considera os elementos textuais de uma publicação, estão fora do escopo deste trabalho analisar imagens, vídeos, relações sociais, propagação na rede entre outras técnicas. A abordagem aqui empregada lança mão das mais modernas técnicas de [NLP](#) na detecção de notícias falsas.

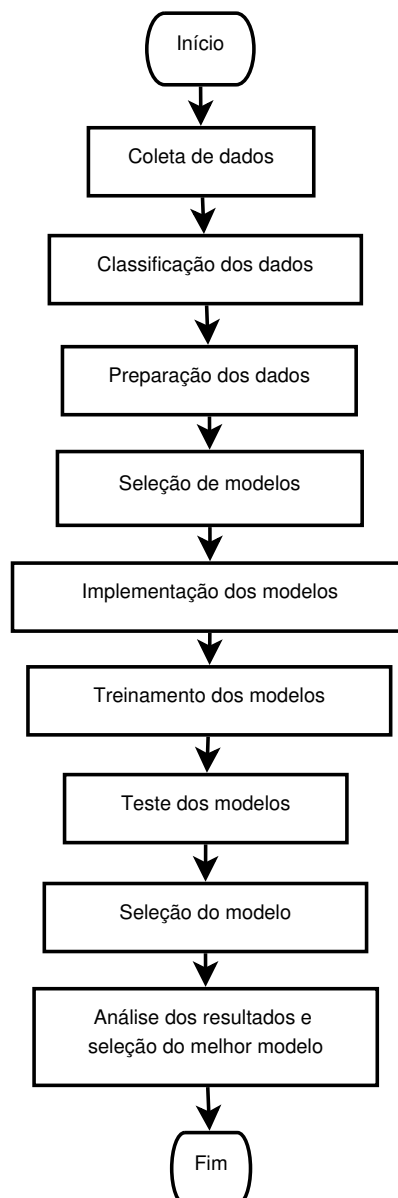
Este trabalho possui os seguintes objetivos específicos:

- Analisar causas e efeitos de notícias falsas através de revisão bibliográfica;
- Criar uma base de dados confiável (amostras falsas são verificadamente falsas e verdadeiras são verificadamente verdadeiras) em português;
- Escolher modelos utilizados no trabalho;
- Treinar os modelos com a base de dados;
- Apresentar o resultado dos modelos;
- Apontar o melhor modelo.

## 1.4 Metodologia

O primeiro passo é a construção de uma base de dados contendo postagens dos usuários da rede social. Além disso, ela precisa ser classificada de forma confiável, pois é utilizada para o treinamento do algoritmo. Ser confiável significa que na realidade todas amostras falsas podem ser verificadas que de fato são falsas e todas as verdadeiras é possível verificar que são verdadeiras. Para construção dessa base de dados, primeiro a ferramenta *Get Old Tweets (GOT)* de Henrique (2018) foi utilizada, assim foi possível obter *tweets* de notícias falsas passadas que são processadas por classificações manuais, ou seja, cada postagem é lida e classificada manualmente. A Figura 1 ilustra o fluxo do trabalho.

Figura 1 – Fluxograma de atividades



Fonte: Elaborado pelo autor

Os textos obtidos passam por uma fase de preparação, tornando-os adequados para servirem de entrada nos modelos. Na fase de seleção de modelos, dentre os modelos de aprendizagem de máquina disponíveis, são escolhidos aqueles que podem ser usados no problema. Na fase seguinte, o código para utilização de cada modelo é implementado. Treina-se os modelos e testa-os para obter resultados. Após os testes, é feita a comparação dos resultados e é elegido o melhor modelo ajustado para resolver o problema.

O objeto de pesquisa deste trabalho é a análise automática de notícias falsas em português. Os passos para execução deste trabalho são assim definidos na [Figura 1](#).

## 1.5 Resultados e contribuições

O modelo de [BERT](#) ajustado a partir do pré treinamento de [Souza, Nogueira e Lotufo \(2020\)](#) obteve o melhor resultado, conseguindo acurácia de 95,1% de acurácia e 95,4% de F1 para verificação de autenticidade de postagens e 94,4% de acurácia e 94,6% de F1 para detecção de notícias falsas. Até o momento da apresentação deste trabalho não foram encontrados resultados superiores na classificação de *tweets* em português, sendo a principal contribuição deste trabalho um modelo no estado da arte para o problema. Além disso, esse trabalho serve de apoio para pesquisas em português que busquem utilizar os modelos disponíveis na biblioteca *transformers* para [NLP](#) em categorização de textos.

## 1.6 Organização do trabalho

O restante deste trabalho é organizado como se segue. O [Capítulo 2](#) apresenta estudos relevantes para entender as notícias falsas em redes sociais e os métodos de [NLP](#). No [Capítulo 3](#), é explicado todo o processo prático da elaboração e utilização da ferramenta. O [Capítulo 4](#) apresenta os resultados obtidos e faz a comparação entre eles. Por fim, no [Capítulo 5](#), apresenta-se os pontos mais importantes obtidos através do trabalho e possibilidades de trabalhos futuros.

## 2 Revisão bibliográfica

As primeiras seções fornecem a fundamentação e os conceitos necessários enquanto a última discute os principais trabalhos relacionados. Na primeira seção, busca-se explicar os conceitos de pós-verdade, desinformação e *fake news*, e explicar suas causas. Prossegue-se com a fundamentação referente às ferramentas utilizadas por este trabalho. Mostra-se uma visão sobre os aspectos jurídicos. Por fim, apresenta-se trabalhos relacionados.

### 2.1 A tríade calamitosa da informação

Pós-verdade (*post-truth*) foi a palavra do ano em 2016 no dicionário Oxford (FOX, 2016). *Fake news* foi a palavra do ano em 2017 no dicionário Collins (COEN, 2017) e desinformação (*misinformation*) foi a palavra do ano em 2018 no site *dictionary.com* (ITALIE, 2018).

#### 2.1.1 Notícias falsas, também conhecidas como *Fake News*

O dicionário Collins (2016) define notícia falsa como “informação falsa, frequentemente sensacional, disseminada sob o propósito de passar-se por notícia” (tradução do autor). Entretanto, é interessante observar que notícias falsas ou *fake news* é um termo que sofreu certa evolução com o passar do tempo. Pode-se dizer que:

as notícias falsas do passado estavam associadas mais à prática jornalística e correspondiam aos falsos acontecimentos que a imprensa porventura viesse a publicar (como no caso do jornalismo amarelo). Hoje, o estado da arte acerca do conceito *fake news* mostra que as notícias falsas acabam sendo desprendidas do jornalismo e aparecem em outras formas de comunicação que se valem do formato reconhecido dos produtos jornalísticos (CARVALHO, 2019, p. 9).

Apesar das considerações anteriores para o entendimento prático deste trabalho, define-se como notícia falsa aquela mensagem que contém informações que não estão de acordo com os fatos<sup>1</sup>. Já como notícia verdadeira, aquela mensagem cujas informações estão de acordo com os fatos e pode ser verificada. A notícia em que a mensagem possui informações que os fatos não podem ser verificados estão fora do escopo deste trabalho.

##### 2.1.1.1 Tipologia das *Fake News*

Uma das tipologias de notícias falsas mais conhecidas é a de Tandoc Jr., Lim e Ling (2018), eles dividiram-nas em sátira, paródia, fabricação, manipulação de foto,

<sup>1</sup> Fato: Informação cuja realidade pode ser comprovada.

anúncio/publicidade e propaganda. Aqui, porém, acredita-se que podem ser definidos menos tipos. Levando em consideração a intenção do autor da postagem falsa, pode-se observar os 3 tipos de notícias falsas a seguir.

Em primeiro lugar, as postagens onde o autor pretende obter o humor. Em segundo lugar, quando o autor pretende alterar a percepção do leitor a respeito de um determinado assunto, ou seja, promover ou difamar um produto/pessoa. Por fim, aquela com a finalidade de crime cibernético, onde uma conta invadida ou um perfil de página/usuário falso são utilizados para divulgar informações falsas a respeito de promoções, benefícios gratuitos, atualizações, entre outras técnicas de engenharia social que são usadas para atrair o leitor a visitar uma página maliciosa na *web* e/ou fornecer informações pessoais.

### 2.1.2 Desinformação, também conhecida como *Misinformation*

O site [Dictionary.com](#) (2018) define desinformação (*misinformation*) como “informação falsa que é espalhada, independentemente de se há intenção de enganar” (tradução do autor). Desinformação é realmente próximo do conceito de *fake news*, sendo que o segundo é um caso do primeiro em que há intenção de enganar. Uma observação importante a respeito da palavra desinformação é sobre o contexto do verbo desinformar: uma pessoa que consumiu desinformação é desinformada no sentido em que está mal informada e não sem informações.

### 2.1.3 Pós-verdade, também conhecida como *Post-truth*

Segundo o dicionário [Oxford](#) (2017), pós-verdade é um adjetivo e está “relacionado a circunstâncias em que as pessoas respondem mais a sentimentos e crenças do que fatos” (tradução do autor). Esse adjetivo é muitas vezes utilizado dizendo que nós estamos na era da pós-verdade, ou seja, uma era em que as pessoas dão mais importância aos sentimentos e crenças pessoais do que à razão. Considerando a atualidade um momento de pós-verdade, torna-se ambiente perigoso para desinformação por facilitar a disseminação de informações falsas.

No livro *Post-truth* de [McIntyre](#) (2018), o autor do livro encontra 5 motivos para explicar o atual fenômeno de pós-verdade, que permite a proliferação da desinformação, e serão explicados a seguir muito superficialmente. Para uma maior compreensão e referências, é recomendada a leitura do livro. Resumidamente:

#### 2.1.3.1 Negação da Ciência

Parte de muitos argumentos, como dizer que os estudos são tendenciosos, afirmar que tal teoria não pode ser comprovada cientificamente e financiar cientistas para produzirem evidências no intuito de gerar dúvida.

### 2.1.3.2 Predisposição Cognitiva

Algumas tendências da nossa mente acabam nos tornando menos racionais do que imaginamos que somos. Tendências como evitar o desconforto psicológico ou desarmonia das crenças (dissonância cognitiva), evitar desconforto social (conformidade social), a ideia de pensar que sua crença inicial está sempre correta (viés de confirmação).

### 2.1.3.3 Declínio da Mídia Tradicional

Alguns fatores levaram a perda de popularidade entre os meios jornalísticos tradicionais, como televisão e jornal. Entre eles, estão a mídia partidária, a ascensão das redes sociais e a promoção de alguns veículos de que pontos de vista diferentes deviam ter a mesma importância, mesmo que um fosse comprovadamente menos confiável que o outro.

### 2.1.3.4 Ascensão das Redes Sociais

A popularização da banda larga, dos computadores pessoais e dos *smartphones* fez com que as pessoas pudessem usar as redes sociais para se comunicarem com muito mais facilidade. Além disso, as redes sociais proporcionaram as pessoas cercarem-se apenas daquilo que elas desejam, vivendo em um ambiente *online* parecido como uma bolha, gerando um ciclo de repetição de confirmação das coisas que elas acreditam, independentemente de serem fatos verdadeiros ou não. Além de toda velocidade praticamente instantânea de disseminação e consumo de informações.

### 2.1.3.5 Relativismo do Pós-modernismo

A relativização da verdade defendida pela cultura pós-modernista. Um dos legados do pós-modernismo para a era atual foi a ideia de que não existe verdade, com isso, toda notícia poderia ser desconstruída do ponto de vista de alguém e cada pessoa que desconstruísse poderia ter uma verdade diferente.

## 2.1.4 Leituras complementares

Como complementação teórica para o leitor que deseja se aprofundar mais no assunto do ponto de vista de ciência da informação, recomenda-se o volume 28, número 3 da revista *El profesional de la información* (POSVERDADE... , 2019), pois esta edição trata apenas de pós-verdade e desinformação. Araújo (2020) investiga sobre o fenômeno de pós-verdade. Um panorama sobre as crenças em notícias falsas e sua proliferação é apresentado por Lazer et al. (2018), passando pelos campos da ciência da informação, psicologia e tecnologia.

Para aprofundar nos aspectos psicológicos, pode-se recorrer a alguns trabalhos. Uma abordagem sobre os aspectos morais e éticos em termos psicológicos é feita por Efron

e Raj (2020). Axt, Landau e Kay (2020) buscaram investigar os fenômenos psicológicos por trás do compartilhamento de desinformação. Já Bago, Rand e Pennycook (2020) analisam o processo de formação de crença no compartilhamento de desinformação. Acerbi (2019) aborda a desinformação das perspectivas da evolução cultural e antropologia cognitiva.

## 2.2 Tecnologia na Informação

Como é inviável empregar esforços humanos na leitura e classificação de cada postagem em redes sociais para averiguar as informações, a tecnologia deve ser usada para fazer isso em massa. Mas como a tecnologia pode fazer isso? Homens e máquinas não falam a mesma língua, para que um algoritmo possa realizar funções sobre um texto, é necessário primeiramente aplicar técnicas de NLP no texto. Este trabalho utiliza o estado da arte de NLP para classificação de textos, aplicando-o em postagens extraídas do Twitter.

Conforme apresentado por Shu et al. (2017), existem alguns traços presentes em notícias falsas que podem ser usados para diferenciá-las, eles podem ser baseados em características de linguística, visual, usuário, postagens, rede de contatos, conhecimento, estilo, postura e em propagação. Através de uma ou mais dessas características, é possível criar ferramentas utilizando diferentes abordagens para resolver o problema. A categorização ou classificação de texto consiste em “dado algum tipo de texto, decida qual conjunto predefinido de classes pertence a ele.” (RUSSELL; NORVIG, 2013, 22.2). Neste trabalho, o tipo de texto são mensagens postadas na rede social Twitter e existem duas classes: conteúdo verdadeiro e conteúdo falso.

Alguns métodos populares usados para classificação de texto são *Bag of Words* (BoW), *K-Nearest Neighbors* (KNN) e *Support Vector Machines* (SVM). Mas esses métodos possuem algumas limitações, como por exemplo não considerarem a posição em que as palavras aparecem numa frase, muitas vezes usando características apenas quantitativas. Por exemplo, nas frases ‘Não, nunca diga sim’ e ‘Sim, nunca diga não’ eles poderiam entendê-las como possuindo significado semelhantes, enquanto o significado delas é praticamente oposto.

Por conta dessa ‘fraqueza’ de métodos populares mais antigos, novos métodos foram criados, onde podemos chamá-los de direcionais. Os modelos direcionais consideram a posição das palavras na frase, seja na direção esquerda pra direita, direita pra esquerda ou até mesmo bidirecionalmente. Porém o custo computacional para essas ferramentas é muito alto. Foi aí que os pesquisadores Vaswani et al. (2017), da Google, propuseram a arquitetura *transformer*, que possibilitava uso de modelos direcionais com um custo significativamente menor do que os que existiam na época, tornando tais modelos mais viáveis para utilização e pesquisa. O campo mais impactado pela custo melhorado pela arquitetura *transformers* foram os modelos que utilizam técnicas de aprendizagem profunda

(*deep learning*).

Aprendizagem profunda é um procedimento muito custoso computacionalmente, em que algoritmos de aprendizagem utilizam um conjunto de dados extremamente grande para obter uma melhor capacidade de generalização. Tais técnicas têm capacidade de obter melhores resultados que o aprendizado de máquina tradicional, mas estão limitadas pela dificuldade de treinar os modelos em profundidade. A boa notícia é que nas redes neurais atuais utiliza-se um conceito de transferência de aprendizagem, com isso, um modelo pode ser treinado uma única vez profundamente e depois ser distribuído já treinado para que outras pessoas possam se beneficiar do aprendizado profundo, mesmo sem ter condições de fazer o treinamento profundo. Com a transferência de aprendizado, os pesos e as características do modelo que passou por aprendizado profundo são exportadas num modelo chamado de pré-treinado. Esse modelo pré-treinado já contara com o conhecimento de um ou vários idiomas (no caso de [NLP](#)) pelo qual tenha passado por aprendizado profundo e estará disponível para ser feito um processo de ajuste fino, para que ele seja ajustado na resolução de problemas específicos, como classificação de textos, tradução e geração de texto.

Usando a arquitetura *transformer* como base, uma série de modelos de aprendizagem profunda para [NLP](#) foram criados. Para facilitar a utilização de tais modelos, [Wolf et al. \(2020\)](#) fizeram a biblioteca *Transformers*, que disponibiliza para a comunidade modelos pré-treinados e modelos ajustados pela comunidade. Além de disponibilizar ampla documentação, tutoriais, bases de dados, implementações de funções necessárias para o uso e avaliação dos modelos entre outros recursos. Graças a esta biblioteca, o uso dos modelos de aprendizagem profunda mais poderosos atualmente é muito mais simples, isto foi um dos facilitadores que possibilitou o uso de tais modelos neste trabalho.

Portanto, os modelos que foram ajustados nesse trabalho são os que dentro da biblioteca *transformers* melhor correspondiam ao problema de análise de notícias falsas, ou seja, estavam prontos pra ser usados em classificação de texto, seu pré-treinamento foi feito em português ou multi-idiomas (se um dos idiomas de pré-treinamento for português) e, por fim, o custo computacional não excedia aos recursos gratuitos disponibilizados pela plataforma *Google Colaboratory* ([Colab](#)).

## 2.3 Privacidade e liberdade de expressão

Diante de tudo que já foi exposto, seja das graves consequências que a desinformação pode causar, ou sobre como a tecnologia deve ser usada para combatê-la, não se pode negligenciar o risco de que, sob o pretexto de combater a desinformação, os direitos constitucionais sejam tolhidos. Especialmente o artigo 5º sobre o direito à liberdade e manifestação do pensamento, e o artigo 220º parágrafo 2 sobre a censura política e



ideológica, presentes na constituição [Brasil \(2020\)](#). O Brasil vive um momento classificado como sensível, ocupando a 107<sup>a</sup> posição no ranking de 180 países sobre liberdade de imprensa feito pela Organização não Governamental (ONG) Repórteres Sem Fronteiras (RSF), muito próximo de passar da situação sensível para difícil pelo ranking [RSF \(2020\)](#).

A preocupação com liberdade *online* é real no Brasil e não uma simples ameaça. Segundo [Lorenzetto e Pereira \(2020\)](#), o polêmico inquérito das *fake news* tocado pelo Supremo Tribunal Federal (STF) é classificado como inconstitucional e na prática censurou pelo menos 2 veículos de comunicação além de contas de usuários em redes sociais. Já no poder legislativo, paira uma completa inexperiência e discordância para tratar o assunto, a situação da divisão entre as opiniões no poder legislativo é explicada como 2 correntes:

uma corrente favorável a criação de leis proibitivas que busquem inibir a produção e circulação de notícias falsas bem como penas severas para seus autores. O prejuízo causado à democracia justificaria o risco em que seria colocado o princípio da liberdade de expressão. Uma outra corrente mostra-se contrária a punições e proibições sob o mesmo argumento, o risco que limites impostos à liberdade de expressão traria à democracia, podendo facilmente estimular e legitimar atos de censura ([THEMUDO; ALMEIDA, 2020](#)).

Prova de que a preocupação com a liberdade *online* é real é que entre os projetos de lei analisados de agosto de 2018 até agosto de 2019 para tratar o assunto das *fake news*, 50% deles foram considerados inconstitucionais por [Macedo e Costa \(2020\)](#). [Rodrigues, Bonone e Mielli \(2020\)](#) trazem posicionamentos da Organização das Nações Unidas (ONU), Organização dos Estados Americanos (OEA), *Electronic Frontier Foundation (EFF)* e outras redes de entidades, contrárias ao dispositivo de coleta excessiva de dados no Projeto de Lei (PL) 2630, aprovado pelo senado em 30 de junho, para o combate de *fake news*, reconhecendo que o PL prejudica a privacidade dos usuários.

Exposta a problemática, o trabalho aqui presente não pretende entrar em questões jurídicas de como é ou como deveria ser a lei para lidar com a desinformação. Contudo, acredita-se que o conteúdo compartilhado nas redes sociais deve ter o emissor do discurso responsabilizado analogamente a como acontece quando o mesmo é dito na cadeira do bar ou na mesa de um restaurante, sendo as redes sociais apenas mais um ambiente em que as pessoas socializam.

## 2.4 Trabalhos Relacionados

Este trabalho representa um esforço em produzir avanços na detecção de notícias falsas no idioma português através de NLP na tarefa de categorização de texto e aproveitando transferência de aprendizagem. Contribui com a aplicação do estado da arte em NLP aplicado ao idioma português. De todos trabalhos encontrados na revisão da

literatura, o trabalho que melhor corresponde ao aqui presente é a dissertação de [Cordeiro \(2019\)](#), pois foi o principal trabalho encontrado de classificação de *tweets* em *fake news* no idioma português brasileiro. Assim como o presente trabalho, ele também envolveu a construção de uma base de dados de *tweets* e todas etapas até a aplicação de algoritmos para classificação.

O livro organizado por [Shu et al. \(2020\)](#) reúne vários artigos cobrindo todo conteúdo sobre desinformação em redes sociais. Começa desde a conceituação, até caracterização, identificação, classificação, usuários, estudos de casos, revisão do estado da arte e ect. O livro busca cobrir todos aspectos sobre o assunto. Uma caracterização de *fake news* com revisão do estado da arte sobre a detecção delas é apresentada por [Zhang e Ghorbani \(2020\)](#). Uma exploração sobre as principais características para detecção de notícias falsas é mostrada por [Reis et al. \(2019\)](#). Uma discussão a respeito do problema das notícias falsas, bem como das técnicas de identificação e mitigação, é feita por [Sharma et al. \(2019\)](#). Uma análise de como a Inteligência Artificial (IA) pode ser usada para derrotar as *fake news* é apresentada por [Cybenko e Cybenko \(2018\)](#).

Análise de detecção de notícias falsas nas perspectivas de conhecimento falso da notícia, estilo de escrita, padrões de propagação e credibilidade do autor, é feita por [Zhou e Zafarani \(2020\)](#). Já [Balestrucci e De Nicola \(2020\)](#) investigam o comportamento de usuários do Twitter propícios a compartilhar notícias falsas pelo seu alto envolvimento com usuários *bots* em um estudo de caso real. [Shu et al. \(2020\)](#) propõem uma abordagem de rotulação de amostras por supervisão social fraca para solucionar o problema de obter bases de dados que passaram por rotulação confiável, pois esta última demanda muito mais esforço para construção. Uma base de dados para detecção de imagens fora de contexto, bem como um modelo para classificação, é apresentado por [Sabir et al. \(2018\)](#).

No estudo de [de Oliveira, Medeiros e Mattos \(2020\)](#), foram utilizados cerca de 33 mil *tweets* em inglês para realizar o treinamento do modelo de classificação. Uma base de dados em filipino é construída pelos autores [Cruz, Tan e Cheng \(2020\)](#), que ajustam classificadores conseguindo até 96% de precisão em um modelo multi-tarefa através de transferência de aprendizado. O *corpus* FAKE.BR em português para detecção de notícias falsas, bem como avaliação de alguns métodos de classificação, são apresentados por [Silva et al. \(2020\)](#).

O trabalho de [Apuke e Omar \(2021\)](#) investiga o compartilhamento de notícias falsas sobre a COVID-19 na Nigéria. [Bal et al. \(2020\)](#) analisaram postagens e compartilhamentos de *tweets* falsos sobre câncer. [Cinelli et al. \(2020\)](#) fizeram um amplo estudo com cerca de 400 mil *tweets* analisados durante o período de eleição do parlamento europeu em 2019.

Um modelo matemático para predição do espalhamento de notícias falsas no Twitter é proposto por [Murayama et al. \(2020\)](#). [Zellers et al. \(2019\)](#) analisam um modelo utilizado para geração de texto que é utilizado para a geração de notícias falsas e também medem a

capacidade de detectar notícias falsas geradas artificialmente. Uma estrutura usando *block chain* é proposta por Qayyum et al. (2019), bem como um levantamento dos desafios e vantagens da utilização de *block chain* para a prevenção de disseminação de desinformação. Gupta et al. (2018) fazem a modelagem de câmaras de eco dentro de redes sociais utilizando matriz de tensores e aplicam fatoração de tensor na detecção de notícias falsas.

Ghanem, Ponzetto e Rosso (2020) apresentam uma ferramenta para detecção de notícias falsas utilizando um modelo neural recorrente e uma variedade de características linguísticas e semânticas. Murayama, Wakamiya e Aramaki (2020) partem do pressuposto de características universais de notícias falsas para criar um sistema não supervisionado de detecção de notícias falsas multi-idioma. Uma estrutura para classificação de notícias baseada em contexto social é apresentada por Shu, Wang e Liu (2019). Borges, Martins e Calado (2019) propuseram um modelo que utiliza redes neurais recorrentes bidirecionais, aplicando-o ao *corpus Fake News Challenge - 1*. Bozarth e Budak (2020) propõem como e com quais métricas avaliar os modelos atuais de classificação de notícias falsas.

Existem também alguns trabalhos e com as mais variadas técnicas voltadas para área de verificação, prevenção e combate a desinformação que estão a nível de utilização por usuários comuns. Santos et al. (2019) propuseram um *bot* para Telegram que ajuda as pessoas verificarem notícias. Souza (2019) fez como dissertação de mestrado um aplicativo *Android* que auxilia na checagem de *links* de notícias. Botnevik, Sakariassen e Setty (2020) fizeram uma extensão de navegador para verificação de *fake news*.

## 3 Desenvolvimento

Este capítulo descreve o desenvolvimento do trabalho. Para realizar este trabalho, todas as etapas descritas na [Figura 1](#) foram realizadas. A começar pela construção de uma base de dados, depois o treinamento de modelos de aprendizagem de máquina a partir dessa base para treinar o classificador de notícias falsas, por fim avaliar qual foi o melhor classificador.

### 3.1 Base de dados: definição e procedimentos

Neste trabalho se escolheu utilizar a rede social Twitter e o idioma português. Por uma questão computacional de lidar com uma menor quantidade de dados (máximo de 280 caracteres por postagem) e facilidade de obter as postagens, foi escolhida a rede social Twitter. Escolhida a rede social, pode-se dar continuidade à coleta de dados para o preenchimento da base de dados.

#### 3.1.1 Coleta de dados

A parte de coleta de dados é uma das mais desgastantes e exaustivas do processo de construção da base de dados, pois ela envolve pesquisar, obter, selecionar e rotular todos os dados selecionados para que possam popular a base. Os processos de pesquisar e obter podem nem sempre retornar os resultados esperados, e o volume de dados retornados pode ser imenso. Selecionar e rotular manualmente os dados obtidos é exaustivo, pois exige a leitura individual de cada amostra para avaliar se está num formato útil ou se é melhor descartar, e, se útil, rotular qual classe a amostra pertence individualmente.

O processo para obter *tweets* consistiu em pesquisar no site da agência de checagem as notícias falsas que foram verificadas e pesquisar por elas usando a ferramenta [GOT](#)<sup>1</sup>, configurando os filtros para a data aproximada da circulação daquele boato. Então, feito o *download* de 100 *tweets* pela ferramenta, cada um era manualmente selecionado e rotulado pelo autor, indexando-os num arquivo *Comma Separated Values (CSV)* que já era o arquivo da base de dados em si. Assim foi feito para todas as amostras de notícias falsas e poucas amostras verdadeiras. Para as amostras verdadeiras, simplesmente foram selecionadas notícias (apenas fatos, e não opiniões) da conta oficial do portal G1 no Twitter e indexadas no arquivo da base de dados até o número de amostras verdadeiras ser balanceado com o

<sup>1</sup> Para buscar e baixar os *tweets*, foi utilizada a ferramenta [GOT](#) de [Henrique \(2018\)](#). A [GOT](#) provê um conjunto de recursos para buscar e fazer downloads de *tweets*, com o código fonte em *python*, é possível selecionar os *tweets* com uma sessão de filtros como data, assunto, *tags*, usuário, número de *tweets* e apenas *tweets* de usuários populares.

número de falsas. Na Figura 2, vê-se um exemplo de como é uma postagem do Twitter originalmente.

Figura 2 – Exemplo de um *tweet*



Fonte: Conta oficial do portal G1 no Twitter

### 3.1.2 Preparação dos dados

Depois de obtida a base de dados foram realizados alguns pré-processamentos descritos a seguir. Amostras repetidas foram excluídas da base de dados. Amostras que os usuários complementavam a postagem original foram mantidas, pois tal complementação poderia tornar uma notícia verdadeira ou falsa dependendo do que o usuário adicionasse na postagem original. Por exemplo, se uma pessoa citar o texto de uma postagem original afirmando que ela é falsa.

Todas as *hashtags* foram removidas, pois a simples presença de uma *hashtag* já poderia significar por exemplo que uma notícia era falsa ou verdadeira. Além da remoção das *hashtags*, removeu-se também todos os *links*, uma vez que a análise de *links* de notícias está fora do escopo desse trabalho e os *links* não possuem valor para análise da mensagem postada.

Foram mantidos acentos, letras maiúsculas e minúsculas, cedilha, aspas e pontuação. Imagens ou vídeos anexados ao conteúdo da postagem estão fora do escopo do nosso trabalho e foram descartados. É importante observar também que a codificação que a base de dados foi salva é a *8-bit Unicode Transformation Format (UTF-8)*. No arquivo final da base de dados restaram apenas 2 colunas, a primeira contém o texto entre aspas duplas e a segunda, o rótulo de verdadeiro ou falso, sendo a primeira linha do arquivo o cabeçalho.

Abaixo, pode-se conferir como o exemplo da [Figura 2](#) é encontrado no arquivo da base de dados.

"Morre Lee Kun-hee, presidente da Samsung", True

A base de dados terminou com 162 amostras, balanceada, sendo as 81 primeiras falsas e as próximas 81 verdadeiras. Conforme exibido na [Tabela 1](#) e comparado com [Cordeiro \(2019\)](#).

Tabela 1 – Comparação do número de amostras.

	Verdadeiras	Falsas
<a href="#">Cordeiro (2019)</a>	93	206
Este trabalho	81	81

Fonte: Elaborado pelo autor.

A seguir uma nuvem de palavras da base de dados é exibida na [Figura 3](#).

Figura 3 – Nuvem de palavras da base de dados



Fonte: Elaborado pelo autor.

## 3.2 Escolha dos modelos

A princípio, a intenção era utilizar a biblioteca em *python Natural Language ToolKit* (NLTK) (BIRD; KLEIN; LOPER, 2009), ela contém métodos para pré-processamento de texto e que permitem obter diversos valores numéricos de um texto, ou seja, realizar processamento de linguagem natural. Todavia, tal biblioteca era uma faca de dois gumes, trazia o benefício de sua simplicidade e facilidade de uso, porém não era uma ferramenta tão poderosa em termos de NLP. Como solução para esse problema, foi proposta a utilização da biblioteca spaCy (HONNIBAL et al., 2020), sendo esta uma biblioteca mais poderosa do que a NLTK, principalmente para obter mais qualidade na classificação no idioma português. Porém, antes mesmo de ser aplicada a spaCy, soube-se de uma ferramenta ainda mais recente e com resultados impressionantes que estavam causando uma revolução no estado da arte de NLP, que foi o *Google BERT* (DEVLIN et al., 2019).

Para utilização do BERT, já não era o caso da utilização de spaCy. Para isso, uma nova biblioteca foi encontrada, a *Transformers* (VASWANI et al., 2017). Essa biblioteca reúne um conjunto de modelos mantidos por grandes empresas e pela comunidade, implementados e prontos para utilização. Vários modelos disponibilizados na biblioteca já passaram pelo processo de pré-treinamento, possibilitando assim pessoas que não teriam como utilizar os benefícios do aprendizado profundo agora pudessem utilizá-lo. Este trabalho buscou nessa biblioteca os modelos que poderiam ser melhor (modelos que receberam algum pré-treinamento em português, para aproveitar a transferência de aprendizagem) utilizados, como resultado optou por ajustar os seguintes modelos: BERT base pré-treinado em português brasileiro por Souza, Nogueira e Lotufo (2020), BERT base pré-treinado em 104 idiomas, incluindo português, por Devlin et al. (2019), BR\_BERTo que é um modelo RoBERTa (LIU et al., 2019) base pré-treinado com um *corpus* de 6,9 milhões de frases em português, disponibilizado pelo usuário rdenadai da comunidade, XLM-R base que foi pré-treinado por Conneau et al. (2020) em 100 idiomas, inclusive português, e, por fim, um modelo ELECTRA (CLARK et al., 2020) *uncased* pré-treinado em português disponibilizado, pelo usuário dlb da comunidade.

Apesar de BERT ter causado uma revolução no estado da arte de NLP, ele não permaneceu muito tempo como o algoritmo que possui os melhores resultados para a tarefa. Tomando o BERT como base, outros modelos foram feitos, como RoBERTa (LIU et al., 2019) e XLM-R (CONNEAU et al., 2020). Mais recentemente Clark et al. (2020) propuseram ELECTRA, que além de ser computacionalmente mais barato do que BERT, chegou a superar os modelos do estado da arte em muitas tarefas.

## 3.3 Implementação dos modelos

Após ter escolhido os modelos utilizados, foi necessário criar um programa que convertesse a base de dados para o formato adequado, fizesse o treinamento e avalie o modelo. Para executar os modelos, o ambiente *online Colab* foi escolhido. Um computador pessoal não possui poder computacional para executar o treinamento e classificação dos modelos, enquanto o ambiente *Colab* possui um poder computacional significativo que permite executar algumas configurações dos modelos escolhidos. Chegou-se a cogitar utilizar um laboratório da Universidade Federal de Minas Gerais (UFMG) para realizar o experimento, porém devido à praticidade e poder computacional da plataforma da *Google*, o *Colab* foi escolhido. Esta plataforma foi criada pela *Google* com o objetivo de incentivar as pesquisas na área de IA e funcionou muito eficientemente neste trabalho, embora algumas vezes seja alocada uma máquina com 8GB de memória de vídeo, fazendo-se necessário reiniciar o *Colab* até obter-se uma máquina com 16GB de memória de vídeo, pois essa máquina com 8GB de memória de vídeo frequentemente não possuía memória suficiente para treinar e avaliar os modelos.

### 3.3.1 Configuração do Ambiente

A linguagem utilizada foi o *Python*, o ambiente funciona no estilo *Jupyter Notebooks*. *PyTorch* foi utilizado nesse trabalho, porém é possível fazer o mesmo utilizando *TensorFlow*. É importante observar que o acelerador de *hardware* do ambiente deve estar configurado para utilizar *Graphics Processing Unit (GPU)*.

Após a instalação da biblioteca *transformers* e importação dos pacotes necessários, foi utilizado o Drive para armazenar a base de dados, então o programa acessa o Drive para pedir autorização em ler e escrever no Drive da conta *Google* do usuário. Após colar o código de autorização, a base de dados é lida do Drive e carregada para o programa.

### 3.3.2 Preparação dos dados

Os modelos utilizam objetos da classe *Dataset* do *PyTorch* como entrada, fazendo-se necessário adequação da base, já que a base de dados original é um arquivo *CSV*. Após carregada na memória, a base se transforma num objeto *DataFrame* do pacote *pandas*. O programa então converte a coluna da categoria do exemplo de *True* para 1 e de *False* para 0. Após isso, divide o *DataFrame* em 2 listas, uma lista de *strings* contendo todos os textos e uma lista de rótulos contendo todos rótulos.

O primeiro passo realmente importante acontece no momento da *tokenização* dos dados. *Tokenizar* os dados é uma expressão que se refere a separar cada palavra e símbolo por um espaço, cada elemento do texto entre os espaços é chamado de *token*, além disso, novos *tokens* como delimitadores de início e fim de sentenças e máscaras podem



ser inseridos. No final do processo, cada *token* é substituído por um número que é o identificador numérico dele. Então é muito importante instanciar o objeto que irá *tokenizar* as amostras a partir do mesmo modelo que será utilizado para treinamento e classificação dos dados. Pois durante o pré-treinamento do modelo, na construção de seu vocabulário, cada *token* recebe um identificador, portanto, modelos diferentes provavelmente terão identificadores diferentes pra cada *token*. Então, não instanciando o *tokenizador* do modelo que será utilizado para treinamento e classificação, os *tokens* colocados não farão sentido quando correlacionados com o vocabulário do classificador, e o resultado do classificador será comprometido.

O objeto responsável por *tokenizar* as amostras precisa ter o mesmo vocabulário do modelo a ser treinado e avaliado, portanto é importante instanciar esses objetos com o mesmo valor no parâmetro de selecionar o modelo da biblioteca.

Após o processo de *tokenização*, segue-se com o processo de construção do *Dataset*. Uma subclasse de *Dataset* é implementada para a própria base de dados, essa classe é que irá converter a lista de textos, agora chamada de amostras codificadas, e a lista de rótulos. Essa classe, além de construir o *Dataset*, conta com uma função de pegar um item de um *Dataset* e ver o tamanho do *Dataset*. O que resulta num *Dataset* com todos os 162 exemplos da base de dados devidamente codificados e prontos para servirem de entrada para o modelo.

### 3.3.3 Configurações, hiper-parâmetros de treinamento e métricas de avaliação

Ainda precisam ser definidos 2 itens importantes. Um objeto com as configurações que serão utilizadas para treinar o modelo e uma função com as métricas utilizadas para avaliar o modelo. Ela que será chamada após o teste do modelo e calculará o valor das métricas.

Um objeto *TrainingArguments* da biblioteca *transformers* foi utilizado para passar as configurações a serem utilizadas no treinamento e avaliação do modelo. O primeiro e o último parâmetro são endereços para salvar logs e resultados. O parâmetro de número de épocas para treinamento deve informar quantas épocas serão utilizadas no treinamento, nesse trabalho esse valor variou de 3 a 10. A terceira linha da tabela está relacionada ao número de amostras que seriam utilizadas ao mesmo tempo no dispositivo durante o treinamento, quanto maior o valor mais memória o dispositivo precisa ter, foi mantido o valor padrão de 16. A quarta linha da tabela informa o número de amostras por vez no dispositivo para avaliação do modelo, foi mantido o valor padrão de 64. A quinta linha da tabela informa a cada quantos passos do treinamento a taxa de aprendizagem seria alterada, escolheu-se o valor 50. Por fim a penúltima linha da tabela informa a taxa de decaimento da taxa de aprendizagem, foi mantido o valor padrão de 0,01. Os valores utilizados podem ser conferidos na [Tabela 2](#).

Tabela 2 – Configurações e Hiper-parâmetros de treinamento

Hiper-parâmetro/configuração	Valor
output_dir	'./results'
num_train_epochs	N
per_device_train_batch_size	16
per_device_eval_batch_size	64
warmup_steps	50
weight_decay	0.01
logging_dir	'./logs'

Fonte: Elaborado pelo autor.

A função que calcula as métricas recebe todos os textos do conjunto de teste e a classificação que o modelo atribuiu ao texto. Cada uma das métricas do trabalho serão explicadas mais adiante na [seção 4.1](#) As métricas mais importantes utilizadas nos trabalhos de detecção de notícias falsas são acurácia e F1. Além dessas, implementou-se precisão, sensibilidade (*recall*) e *Matthews Correlation Coefficient* (**MCC**). Para o cálculo das métricas, foi utilizada a biblioteca *SciKit-learn* (**sklearn**) ([PEDREGOSA et al., 2011](#)).

### 3.3.4 Validação Cruzada, Treinamento e Avaliação de épocas

Separar uma parte do *Dataset* para treinamento e o restante para testes é simples e rápido de fazer, porém esses testes podem não garantir uma boa avaliação do modelo. Para melhor avaliar a capacidade de generalização do modelo, uma técnica frequentemente empregada é a validação cruzada. Com esta técnica, todo o conjunto de dados é utilizado para testes pelo menos 1 vez. Isso melhora a compreensão da capacidade de generalização do modelo. Neste trabalho, foram utilizadas 10 iterações para processar os dados, sempre utilizando um conjunto de teste diferente do anterior. Para aplicar a validação cruzada, foi utilizada a biblioteca **sklearn**. Como as amostras não estão distribuídas aleatoriamente no *Dataset*, mas sim as 81 primeiras são falsas e as 81 seguintes são verdadeiras, utilizou-se do recurso de montar o conjunto de teste não de forma sequencial, mas de forma aleatória. A biblioteca oferece esse recurso de forma que a taxa de repetição de elementos no conjunto de teste é nula ou muito baixa. Em testes, não verificou-se repetições, porém, segundo a documentação da biblioteca, poderiam ocorrer. Devido ao fato de a validação cruzada gerar 10 conjuntos de teste diferentes, é necessário treinar e testar o modelo 10 vezes pra cada configuração de número de épocas, isso garante mais segurança nas métricas calculadas a partir do modelo. Utilizou-se 10 iterações por recomendação da própria biblioteca, que recomendava 10 ou 5 iterações, que são valores padrões estabelecidos e reconhecidos pela comunidade acadêmica em todo mundo. Além disso, escolhendo 10

divisões para a validação cruzada permitiu-se que fossem ter mais amostras no conjunto de treinamento do que se tivesse utilizado 5.

Antes de iniciar as iterações de treino e avaliação, o índice dos elementos de cada um dos 10 conjuntos de treinamento e teste são gravados em listas de índices de treinamento e teste respectivamente. O primeiro elemento armazena uma lista de índices que identificam as amostras daquela iteração para treinamento e teste. O segundo elemento armazena os índices das amostras para a segunda iteração e assim sucessivamente. Após a divisão dos índices, o modelo é instanciado e o objeto da classe *Trainer* da biblioteca *transformers* é construído com os devidos argumentos. Então inicia-se as iterações, chama-se o método de treinamento que inicia o treinamento. Após o treinamento, o método de avaliação classifica as amostras do conjunto de teste. As predições do conjunto de teste são utilizados para calcular as métricas, que depois de calculadas são incluídas numa lista que armazena os resultados das métricas para cada iteração.

Após a conclusão da 10<sup>a</sup> iteração, a lista com os resultados está completa, então uma média aritmética de cada métrica é feita e impressa. O resultado impresso é salvo numa planilha para avaliação do melhor modelo. Para cada alteração, seja no modelo ou no número de épocas, é necessário salvar numa nova entrada da tabela.

Cada um dos 5 modelos foram executados com validação cruzada várias vezes. Cada vez com um número de épocas diferente. Observa-se que **RoBERTa** e **XLM-R** são notadamente mais pesados computacionalmente. O número de amostras carregadas por vez para treinamento foi 16, porém nenhum dos dois consegue executar com valores padrão em algumas configurações de máquinas gratuitas do **Colab**, que possuíam 8 GB de memória de vídeo. Portanto, fez-se necessário reiniciar a máquina para obter novas configurações, até que fosse possível utilizá-los nas máquinas. Por isso, para os dois modelos foi utilizado um treinamento de 8 amostras por vez. Enquanto a quantidade de amostras do conjunto de testes por vez foi diminuída de 64 para 32 nesses 2 modelos. Para todos eles, o número de épocas que foi utilizado para treinamento foram de 3 até 10, para evitar subaproveitamento do modelo e também o sobre-ajuste (*Overfitting*). Além disso, o atributo usado para alterar a taxa de aprendizagem foi configurado para 50 passos, em vez dos 500 passos que vêm por padrão, pois o valor padrão é muito elevado para uma base de dados pequena como a que foi utilizada, não obtendo uma taxa de aprendizagem adequada para o ajuste fino dos modelos. Testes empíricos com valores como 100, 25, 10 e 5 foram feitos para constatar ser razoável esse valor de 50.

Todo o código fonte deste trabalho foi feito a partir das adaptações da documentação da biblioteca *transformers*, que apesar de ser em inglês, é bastante completa, clara e exemplificada. Além disso, o código fonte, bem como a base de dados, deste trabalho estão disponibilizados no *GitHub* do autor<sup>2</sup>.

<sup>2</sup> <https://github.com/WagnerNarde/ML-Transformers-Tweets-falsos>

## 4 Resultados

Os resultados deste trabalho estão divididos em 5 tabelas, uma para cada modelo utilizado. As métricas utilizadas foram Acurácia, F1, Precisão, Sensibilidade e MCC. Acurácia e F1 são frequentemente utilizadas para este tipo de trabalho, algumas vezes os trabalhos são publicados apenas com essas duas métricas, porém para uma visão mais profunda dos resultados, foram selecionadas algumas outras métricas além delas. Observe que as métricas F1, precisão e sensibilidade dos modelos que não obtiveram o melhor resultado são calculadas para classificação de notícias verdadeiras, enquanto o melhor modelo foram calculadas para ambas as classes (falsas e verdadeiras).

### 4.1 Métricas

A acurácia mostra o quanto o resultado do classificador se aproximou do valor real esperado, ou seja, o quanto o classificador acertou na sua classificação. Uma acurácia alta implica em baixa taxa de erro total. A acurácia é calculada através da fórmula:

$$Ac = \frac{VP + VN}{VP + VN + FP + FN} \quad (4.1)$$

Em que VN são amostras de notícias verdadeiras que foram classificadas como verdadeiras. VP são amostras de notícias falsas que foram classificadas como falsas. FN são amostras de notícias falsas que foram classificadas como verdadeiras e FP são amostras de notícias verdadeiras que foram classificadas como falsas.

Precisão mostra o quanto os exemplos classificados como notícias falsas eram notícias falsas de fato. Um classificador com alta precisão significa que, quando ele classificar a notícia falsa como falsa, então é porque ela é realmente falsa. Observe que essa métrica não leva em consideração as notícias falsas que ele deixou de classificar como falsa, apenas mostra que quando classificou como falsa, se era falsa mesmo. Sua fórmula é:

$$Prec = \frac{VP}{VP + FP} \quad (4.2)$$

Sensibilidade (*recall*) mostra o quanto as notícias falsas realmente foram classificadas, ou seja, se o modelo tiver uma sensibilidade alta, significa que a maioria das notícias falsas estão sendo detectadas. Um modelo com baixa sensibilidade deixa muitas notícias falsas se passarem por verdadeiras, mesmo se ele tiver alta precisão. Pode ser calculada pela seguinte equação:

$$Sens = \frac{VP}{VP + FN} \quad (4.3)$$

F1 é a média harmônica de precisão e sensibilidade. Esta métrica, por combinar as duas anteriores, busca representar não apenas o quanto um classificador consegue detectar de notícias falsas, como também o quanto do que ele detectou é realmente falso. Em suma, pra obter uma alta F1 é necessário tanto alta precisão como alta sensibilidade no classificador, caso uma das duas seja baixa, irá puxar a F1 pra baixo. Calcula-se como:

$$F1 = 2 \cdot \frac{Prec \cdot Sens}{Prec + Sens} \quad (4.4)$$

Conforme apresentado na maioria dos trabalhos relacionados, acurácia e F1 são suficientes para avaliar os modelos. Ainda assim, resolveu-se calcular o MCC como uma métrica de desempate, caso necessário, pois este também é uma métrica para avaliar qualidade de classificadores binários. Pode ser calculado a partir da equação:

$$MCC = \frac{VP \cdot VN - FP \cdot FN}{\sqrt{(VP + FP)(VP + FN)(VN + FP)(VN + FN)}} \quad (4.5)$$

## 4.2 ELECTRA *uncased* pré-treinado em Português

O único modelo *uncased*<sup>1</sup> utilizado foi o ELECTRA. Conforme a literatura, modelos que não diferenciam letras maiúsculas e minúsculas geralmente têm resultados piores dos que os que diferenciam. Sendo um modelo desse tipo, era esperado um resultado abaixo dos outros classificadores pré-treinados apenas em português. Mesmo assim, acredita-se que o modelo não conseguiu generalizar bem o problema. Por conseguir bater o estado da arte em algumas tarefas de NLP, esperava-se mais desse modelo. Uma das razões para o baixo desempenho do modelo também pode ser que os pesos do pré-treinamento do modelo da comunidade não ofereceram ganho de eficiência, mas sim queda de desempenho para o ajuste da tarefa. Os resultados para o modelo ELECTRA podem ser vistos na Tabela 3, em que 10 épocas se sobressaíram em eficiência no ajuste.

## 4.3 RoBERTa pré-treinado em Português

RoBERTa é notadamente mais pesado do que BERT em consumo de recursos computacionais, além de ser um aprimoramento do mesmo, portanto, excelentes resultados eram esperados desse modelo. O desempenho ficou melhor que o dos resultados da seção 4.2 e quase empatado com o modelo da seção 4.4, como já era esperado por se tratar de um modelo que diferencia letras maiúsculas e minúsculas. Com 7 épocas foi mostrado na Tabela 4 o modelo que melhor se ajustou ao problema. Sendo RoBERTa um dos mais modernos algoritmos para NLP, um resultado pelo menos melhor que os modelos multi-idiomas era esperado. O resultado atual porém pode ser reflexo que o pré-treinamento em

<sup>1</sup> Não é case sensitive, ou seja, não diferencia letras maiúsculas de minúsculas

Tabela 3 – Resultados ELECTRA (uncased) com pré-treinamento em Português

Épocas	Acurácia	F1	Precisão	Sensibilidade	MCC
3	0.833	0.826	0.844	0.822	0.668
4	0.808	0.789	0.799	0.803	0.624
5	0.834	0.820	0.821	0.829	0.660
6	0.852	0.846	0.887	<b>0.836</b>	0.690
7	0.803	0.796	0.816	0.812	0.632
8	0.816	0.813	0.833	0.817	0.640
9	0.845	0.836	0.843	<b>0.836</b>	0.675
10	<b>0.864</b>	<b>0.848</b>	<b>0.883</b>	0.824	<b>0.720</b>

Fonte: Elaborado pelo autor.

português pelo qual o modelo passou não transferiu aprendizado útil à resolução desta tarefa.

Tabela 4 – Resultados RoBERTa com pré-treinamento em Português

Épocas	Acurácia	F1	Precisão	Sensibilidade	MCC
3	0.815	0.826	0.771	0.919	0.658
4	0.833	0.829	0.854	0.831	0.678
5	0.869	0.872	0.833	0.921	0.721
6	0.876	0.863	0.860	0.883	0.740
7	<b>0.901</b>	0.897	0.852	<b>0.962</b>	<b>0.812</b>
8	0.883	0.866	0.845	0.917	0.781
9	0.889	<b>0.902</b>	<b>0.865</b>	0.955	0.789
10	0.884	0.885	0.855	0.929	0.779

Fonte: Elaborado pelo autor.

#### 4.4 XLM-R pré-treinado em multi-idíomas

XLM-R foi um modelo originalmente proposto para a tradução de idiomas, por isso ele está disponível em versão multi-idíomas pré-treinado em vários idiomas, inclusive português. Apesar dele não ter sido originalmente proposto para classificação de texto, alguns trabalhos da revisão bibliográfica conseguiram bons resultados através da transferência de aprendizagem em modelos multi-tarefas, o que fazia valer a pena o ajuste do modelo ao problema, dada a chance de se obter um bom desempenho. O desempenho do modelo foi bem superior ao modelo da [seção 4.2](#) e ligeiramente superior ao modelo da [seção 4.3](#), sendo capaz de ficar entre os três melhores modelos deste trabalho. Verificou-se na [Tabela 5](#) que com 9 épocas o modelo teve seu melhor desempenho.

Tabela 5 – Resultados XLM-R com pré-treinamento Multi-idioma

Épocas	Acurácia	F1	Precisão	Sensibilidade	MCC
3	0.826	0.832	0.848	0.870	0.703
4	0.858	0.841	0.856	0.878	0.731
5	0.883	0.876	<b>0.903</b>	0.869	0.776
6	0.882	0.884	0.844	0.940	0.772
7	0.884	0.883	0.854	0.919	0.770
8	0.882	0.870	0.859	0.890	0.767
9	<b>0.903</b>	<b>0.898</b>	0.883	0.922	<b>0.804</b>
10	0.895	0.887	0.850	<b>0.950</b>	0.787

Fonte: Elaborado pelo autor.

## 4.5 BERT pré-treinado em multi-idiomas

Este modelo foi o primeiro disponibilizado que possibilitava a aplicação dos ganhos de eficiência em [NLP](#) que o [BERT](#) trazia, no idioma português. Sendo portanto o primeiro modelo a ser escolhido para compor este trabalho, esperava-se dele que fosse capaz de ultrapassar a pontuação dos resultados atuais disponibilizados na literatura. Apesar de ser o modelo mais antigo, cumpriu muito bem sua tarefa e conseguiu uma acurácia e F1 acima de 90% em sua melhor opção, que foi mostrado na [Tabela 6](#) utilizando 6 épocas. Apesar do pré-treinamento ser dividido com outros idiomas, este resultado é muito expressivo.

Tabela 6 – Resultados BERT com pré-treinamento Multi-idioma

Épocas	Acurácia	F1	Precisão	Sensibilidade	MCC
3	0.853	0.860	0.844	0.895	0.724
4	0.869	0.848	0.859	0.853	0.741
5	0.889	0.885	0.858	0.921	0.781
6	0.896	0.891	0.898	0.913	0.810
7	0.864	0.871	0.847	0.918	0.746
8	0.889	0.871	0.865	0.890	0.767
9	0.883	0.878	0.893	0.878	0.772
10	<b>0.914</b>	<b>0.918</b>	<b>0.900</b>	<b>0.944</b>	<b>0.825</b>

Fonte: Elaborado pelo autor.

## 4.6 BERT pré-treinado em Português

O modelo [BERT](#) pré-treinado em português foi o que obteve o melhor resultado entre todos os modelos. Sua acurácia e F1 foram superiores a todos os outros modelos, mostrando que o pré-treinamento em português feito por [Souza, Nogueira e Lotufo \(2020\)](#)

foi muito eficiente e contribuiu positivamente e muito para o bom desempenho do modelo. Tal benefício de pré-treinamento, porém, não foi observado na [seção 4.2](#) e na [seção 4.3](#). Este modelo foi capaz de classificar corretamente as notícias, com excelentes resultados. O modelo que utilizou 6 épocas foi a configuração melhor ajustada ao problema, conforme mostrado na [Tabela 7](#). Observe que as tabelas anteriores continham apenas os valores das métricas calculadas para notícias verdadeiras, enquanto esta foram calculadas as métricas igualmente para notícias verdadeiras, porém uma nova execução foi feita para calcular as métricas para notícias falsas, por isso aparecem (V), (F), acurácia média e MCC média apenas nesta tabela. Não foi feito isso para os modelos anteriores pois por questão de tempo e consumo de recursos computacionais seria inviável reproduzir todos os experimentos anteriores mais uma vez.

Tabela 7 – Resultados BERT com pré-treinamento em Português

Épocas	$\overline{\text{Acurácia}}$	F1 (V)	F1 (F)	Prec. (V)	Prec. (F)	Sens. (V)	Sens. (F)	$\overline{\text{MCC}}$
3	0.923	0.903	0.940	0.904	0.922	0.937	<b>0.965</b>	0.852
4	0.917	0.917	0.907	0.935	0.927	0.921	0.902	0.846
5	0.925	0.912	0.935	0.903	0.949	0.945	0.931	0.861
6	<b>0.944</b>	<b>0.955</b>	0.928	0.944	0.948	<b>0.971</b>	0.927	<b>0.887</b>
7	0.938	0.932	<b>0.946</b>	0.918	<b>0.950</b>	0.957	0.945	0.879
8	0.935	0.941	0.927	0.950	0.932	0.942	0.931	0.877
9	0.942	0.954	0.919	<b>0.953</b>	0.909	0.957	0.933	0.871
10	0.938	0.937	0.938	0.928	0.940	0.951	0.938	0.875

Fonte: Elaborado pelo autor.

## 4.7 Melhor modelo

Apesar das bases de dados serem diferentes, a [Tabela 8](#) podemos verificar uma comparação do o melhor modelo de [Cordeiro \(2019\)](#) com o melhor modelo deste trabalho. Tal comparação evidenciou que, apesar de uma base de dados menos populosa, o modelo ajustado neste trabalho obteve um resultado significativo em relação ao estudo anterior.

Tabela 8 – Comparação com [Cordeiro \(2019\)](#)

Modelo	Épocas	Acurácia Média	F1 Média	Precisão Média
<i>Complement Naive Bayes</i>	-	-	0.74	0.74
BERT base pt-br	6	<b>0.944</b>	<b>0.941</b>	<b>0.946</b>

Fonte: Elaborado pelo autor.



## 5 Conclusão

Este trabalho apresentou uma revisão sobre as notícias falsas, investigando a literatura sobre suas causas e consequências, bem como uma verificação sobre o estado da arte de sua detecção. Construiu-se uma base de dados balanceada com *tweets* classificados de forma confiável. A base possibilitou treinamento de modelos para detecção de notícias falsas. O classificador proposto foi capaz de superar os resultados existentes e estabelecer-se como o estado da arte para detecção de *tweets* falsos em português através de NLP. Sendo que o modelo BERT com 6 épocas foi o melhor de todos, o que pode ser visualizado pela Tabela 8.

Os modelos com os resultados apresentados na seção 4.2 e seção 4.3 foram pré-treinados pela comunidade, sendo os únicos modelos presentes neste trabalho que não foram submetidos a comunidade científica em nenhuma publicação até o momento. Observou-se neles um resultado muito abaixo do esperado, pode-se levantar a questão de que se tais modelos passaram pelo processo de pré-treinamento de modo correto, devido à ausência de material científico a respeito deles e por seus resultados abaixo do esperado.

A principal contribuição deste trabalho foi o ajuste de um modelo do estado da arte em NLP para a detecção de notícias falsas em português. Pois, conforme foi evidenciado pelos resultados, os algoritmos modernos, apesar de computacionalmente custosos, oferecem resultados bem superiores aos métodos clássicos que são utilizados em praticamente todos trabalhos em português. Além disso, a disponibilização do código fonte comentado e as orientações contidas neste trabalho ajudarão outras pessoas para que também aproveitem o benefício da eficiência desses modelos para esse e outros problemas em língua portuguesa. Afinal, é extremamente escassa a disponibilidade de materiais dessas tecnologias em idioma português.

Como trabalhos futuros, uma base de dados maior pode ser utilizada e modelos mais recentes podem ser ajustados para superarem os resultados deste trabalho. Ainda com os presentes modelos, um ambiente de execução com mais recursos computacionais possibilitaria o teste de mais possibilidades de configurações. Além disso, a criação de modelos para outras redes sociais e sites de notícias pode ser uma possibilidade. Pode-se tentar usar transferência de aprendizagem para isso, do modelo atual para esses modelos, buscando obter melhor aproveitamento.

## Referências

- ACERBI, A. Cognitive attraction and online misinformation. *Palgrave Communications*, v. 5, n. 1, p. 15, Feb 2019. ISSN 2055-1045. Disponível em: <<https://doi.org/10.1057/s41599-019-0224-y>>. Citado na página 22.
- APUKE, O. D.; OMAR, B. Fake news and covid-19: modelling the predictors of fake news sharing among social media users. *Telematics and Informatics*, v. 56, p. 101475, 2021. ISSN 0736-5853. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S0736585320301349>>. Citado na página 25.
- ARAÚJO, C. A. O fenômeno da pós-verdade e suas implicações para a agenda de pesquisa na ciência da informação. *Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação*, v. 25, p. 01–17, maio 2020. Disponível em: <<https://periodicos.ufsc.br/index.php/eb/article/view/1518-2924.2020.e72673>>. Citado na página 21.
- AXT, J. R.; LANDAU, M. J.; KAY, A. C. The psychological appeal of fake-news attributions. *Psychological Science*, v. 31, n. 7, p. 848–857, 2020. PMID: 32672128. Disponível em: <<https://doi.org/10.1177/0956797620922785>>. Citado na página 22.
- BAGO, B.; RAND, D. G.; PENNYCOOK, G. Fake news, fast and slow: Deliberation reduces belief in false (but not true) news headlines. *Journal of experimental psychology: general*, American Psychological Association, v. 149, n. 8, 2020. Citado na página 22.
- BAL, R. et al. Analysing the extent of misinformation in cancer related tweets. *Proceedings of the International AAAI Conference on Web and Social Media*, v. 14, n. 1, p. 924–928, May 2020. Disponível em: <<https://ojs.aaai.org/index.php/ICWSM/article/view/7359>>. Citado na página 25.
- Balestrucci, A.; De Nicola, R. Credulous users and fake news: a real case study on the propagation in twitter. In: *2020 IEEE Conference on Evolving and Adaptive Intelligent Systems (EAIS)*. [S.l.: s.n.], 2020. p. 1–8. Citado na página 25.
- BIRD, S.; KLEIN, E.; LOPER, E. (Ed.). *Natural Language Processing with Python: Analyzing text with the natural language toolkit*. 1. ed. [S.l.]: O'Reilly Media, Inc., 2009. ISBN 9780596516499. Citado na página 30.
- BORGES, L.; MARTINS, B.; CALADO, P. Combining similarity features and deep representation learning for stance detection in the context of checking fake news. *J. Data and Information Quality*, Association for Computing Machinery, New York, NY, USA, v. 11, n. 3, jun. 2019. ISSN 1936-1955. Disponível em: <<https://doi.org/10.1145/3287763>>. Citado na página 26.
- BOTNEVIK, B.; SAKARIASSEN, E.; SETTY, V. Brenda: Browser extension for fake news detection. In: *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York, NY, USA: Association for Computing Machinery, 2020. (SIGIR '20), p. 2117–2120. ISBN 9781450380164. Disponível em: <<https://doi.org/10.1145/3397271.3401396>>. Citado na página 26.

- BOVET, A.; MAKSE, H. A. Influence of fake news in twitter during the 2016 us presidential election. *Nature communications*, Nature Publishing Group, v. 10, n. 1, p. 1–14, 2019. Citado na página 15.
- BOZARTH, L.; BUDAK, C. Toward a better performance evaluation framework for fake news classification. *Proceedings of the International AAAI Conference on Web and Social Media*, v. 14, n. 1, p. 60–71, May 2020. Disponível em: <<https://ojs.aaai.org/index.php/ICWSM/article/view/7279>>. Citado na página 26.
- BRASIL. *Constituição da República Federativa do Brasil*. Brasília, DF: Presidência da República, 2020. Disponível em: <[www.planalto.gov.br/ccivil\\_03/Constituicao/Constituicao.htm](http://www.planalto.gov.br/ccivil_03/Constituicao/Constituicao.htm)>. Acesso em: 10 mar 2021. Citado na página 24.
- CARVALHO, R. L. V. R. Notícias falsas ou propaganda?: Uma análise do estado da arte do conceito fake news. *Questões Transversais*, v. 7, n. 13, p. 21–30, 2019. Citado na página 19.
- CINELLI, M. et al. The limited reach of fake news on twitter during 2019 european elections. *PLOS ONE*, Public Library of Science, v. 15, n. 6, p. 1–13, 06 2020. Disponível em: <<https://doi.org/10.1371/journal.pone.0234689>>. Citado na página 25.
- CLARK, K. et al. ELECTRA: Pre-training text encoders as discriminators rather than generators. In: *ICLR*. [s.n.], 2020. Disponível em: <<https://openreview.net/pdf?id=r1xMH1BtvB>>. Citado na página 30.
- COEN, S. 'Fake news' is named words of the year (honestly!): Dictionary to include term after its usage increased by 365% during 2017. Mail Online, 2017. Disponível em: <[www.dailymail.co.uk/news/article-5040897/Fake-news-named-words-year-honest.html](http://www.dailymail.co.uk/news/article-5040897/Fake-news-named-words-year-honest.html)>. Acesso em: 26 fev 2021. Citado na página 19.
- COLLINS, H. *Definition of 'fake news'*. Collins English Dictionary Online, 2016. Disponível em: <[www.collinsdictionary.com/dictionary/english/fake-news](http://www.collinsdictionary.com/dictionary/english/fake-news)>. Acesso em: 26 fev 2021. Citado na página 19.
- CONNEAU, A. et al. Unsupervised cross-lingual representation learning at scale. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, 2020. p. 8440–8451. Disponível em: <<https://www.aclweb.org/anthology/2020.acl-main.747>>. Citado na página 30.
- CORDEIRO, P. R. d. S. *Processo de construção de corpus de tweets para verificação automática de rumores em língua portuguesa*. Dissertação (Mestrado) — Universidade de Fortaleza, Fortaleza, nov. 2019. Disponível em: <<http://dspace.unifor.br/handle/tede/113593>>. Citado 4 vezes nas páginas 10, 25, 29 e 39.
- CRUZ, J. C. B.; TAN, J. A.; CHENG, C. Localization of fake news detection via multitask transfer learning. In: *Proceedings of the 12th Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, 2020. p. 2596–2604. ISBN 979-10-95546-34-4. Disponível em: <<https://www.aclweb.org/anthology/2020.lrec-1.316>>. Citado na página 25.
- Cybenko, A. K.; Cybenko, G. Ai and fake news. *IEEE Intelligent Systems*, v. 33, n. 5, p. 1–5, 2018. Citado na página 25.

de Oliveira, N. R.; Medeiros, D. S. V.; Mattos, D. M. F. A sensitive stylistic approach to identify fake news on social networking. *IEEE Signal Processing Letters*, v. 27, p. 1250–1254, 2020. Citado na página 25.

DEVLIN, J. et al. BERT: Pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, 2019. p. 4171–4186. Disponível em: <<https://www.aclweb.org/anthology/N19-1423>>. Citado na página 30.

DICTIONARY.COM. *Definition of 'misinformation'*. Dictionary.com, 2018. Disponível em: <[www.dictionary.com/browse/misinformation](http://www.dictionary.com/browse/misinformation)>. Acesso em: 26 fev 2021. Citado na página 20.

EFFRON, D. A.; RAJ, M. Misinformation and morality: Encountering fake-news headlines makes them seem less unethical to publish and share. *Psychological Science*, v. 31, n. 1, p. 75–87, 2020. PMID: 31751517. Disponível em: <<https://doi.org/10.1177/0956797619887896>>. Citado na página 22.

FOX, K. 'Post-truth' named word of the year by Oxford Dictionaries. CNN, 2016. Disponível em: <[www.edition.cnn.com/2016/11/16/world/word-of-the-year-post-truth-oxford/index.html](http://www.edition.cnn.com/2016/11/16/world/word-of-the-year-post-truth-oxford/index.html)>. Acesso em: 26 fev 2021. Citado na página 19.

GHANEM, B.; PONZETTO, S. P.; ROSSO, P. Factweet: Profiling fake news twitter accounts. In: ESPINOSA-ANKE, L.; MARTÍN-VIDE, C.; SPASIĆ, I. (Ed.). *Statistical Language and Speech Processing*. Cham: Springer International Publishing, 2020. p. 35–45. ISBN 978-3-030-59430-5. Citado na página 26.

Gupta, S. et al. Cimtdetect: A community infused matrix-tensor coupled factorization based method for fake news detection. In: *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. [S.l.: s.n.], 2018. p. 278–281. Citado na página 26.

HENRIQUE, J. *Get Old Tweets Programmatically*. 2018. Repository on GitHub. Disponível em: <[www.github.com/Jefferson-Henrique/GetOldTweets-python](https://www.github.com/Jefferson-Henrique/GetOldTweets-python)>. Acesso em: 27 out 2019. Citado 2 vezes nas páginas 17 e 27.

HONNIBAL, M. et al. *spaCy: Industrial-strength Natural Language Processing in Python*. Zenodo, 2020. Disponível em: <<https://doi.org/10.5281/zenodo.1212303>>. Citado na página 30.

ISLAM, M. S. et al. Covid-19?related infodemic and its impact on public health: A global social media analysis. *The American Journal of Tropical Medicine and Hygiene*, The American Society of Tropical Medicine and Hygiene, Arlington VA, USA, v. 103, n. 4, p. 1621 – 1629, 07 Oct. 2020. Disponível em: <<https://www.ajtmh.org/view/journals/tpmd/103/4/article-p1621.xml>>. Citado na página 15.

ITALIE, L. *Dictionary.com chooses 'misinformation' as word of the year*. AP News, 2018. Disponível em: <[www.apnews.com/article/e4b3b7b395644d019d1a0a0ed5868b10](http://www.apnews.com/article/e4b3b7b395644d019d1a0a0ed5868b10)>. Acesso em: 26 fev 2021. Citado na página 19.

- LAZER, D. M. J. et al. The science of fake news. *Science*, American Association for the Advancement of Science, v. 359, n. 6380, p. 1094–1096, 2018. ISSN 0036-8075. Disponível em: <<https://science.sciencemag.org/content/359/6380/1094>>. Citado na página 21.
- LIU, Y. et al. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. Citado na página 30.
- LORENZETTO, B. M.; PEREIRA, R. d. R. O Supremo Soberano no Estado de Exceção: a (des)aplicação do direito pelo STF no Inquérito das "Fake News"(Inquérito n. 4.781). *Sequência (Florianópolis)*, scielo, p. 173 – 203, 08 2020. ISSN 2177-7055. Disponível em: <[http://www.scielo.br/scielo.php?script=sci\\_arttext&pid=S2177-70552020000200173&nrm=iso](http://www.scielo.br/scielo.php?script=sci_arttext&pid=S2177-70552020000200173&nrm=iso)>. Citado na página 24.
- MACEDO, J. C. d.; COSTA, R. V. A (in)constitucionalidade dos projetos de lei apresentados na câmara dos deputados do brasil sobre *Fake News*. *Internet & Sociedade*, InternetLab, v. 1, n. 2, p. 102 – 125, 12 2020. ISSN 2763-5244. Disponível em: <<https://revista.internetlab.org.br/wp-content/uploads/2020/12/A-InConstitucionalidade-dos-Projetos.pdf>>. Citado na página 24.
- MCINTYRE, L. C. *Post-truth*. Cambridge: MIT Press, 2018. (The MIT Press essential knowledge). ISBN 9780262535045. Citado na página 20.
- MURAYAMA, T.; WAKAMIYA, S.; ARAMAKI, E. *Universal Fake News Collection System using Debunking Tweets*. 2020. Citado na página 26.
- MURAYAMA, T. et al. *Modeling and Predicting Fake News Spreading on Twitter*. 2020. Citado na página 25.
- OXFORD. *Definition of 'post-truth'*. Oxford Learner's Dictionaries, 2017. Disponível em: <[www.oxfordlearnersdictionaries.com/us/definition/english/post-truth?q=post-truth](http://www.oxfordlearnersdictionaries.com/us/definition/english/post-truth?q=post-truth)>. Acesso em: 26 fev 2021. Citado na página 20.
- PEDREGOSA, F. et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, v. 12, p. 2825–2830, 2011. Citado na página 33.
- POSVERDADE y desinformación. *El profesional de la información (EPI)*, v. 28, n. 3, 2019. Citado na página 21.
- Qayyum, A. et al. Using blockchain to rein in the new post-truth world and check the spread of fake news. *IT Professional*, v. 21, n. 4, p. 16–24, 2019. Citado na página 26.
- Reis, J. C. S. et al. Supervised learning for fake news detection. *IEEE Intelligent Systems*, v. 34, n. 2, p. 76–81, 2019. Citado na página 25.
- RODRIGUES, T. C. M.; BONONE, L. M.; MIELLI, R. Desinformação e crise da democracia no brasil: é possível regular *fake news*? *Confluências*, Programa de Pós-Graduação em Sociologia e Direito da Universidade Federal Fluminense, v. 22, n. 3, p. 30 – 52, 12 2020. ISSN 1678-7145. Disponível em: <<https://periodicos.uff.br/confluencias/article/view/45470/27124>>. Citado na página 24.
- RSF. *Classificação Mundial da Liberdade de Imprensa 2020*. Repórteres sem Fronteiras, 2020. Disponível em: <[https://rsf.org/pt/classificacao\\_dados](https://rsf.org/pt/classificacao_dados)>. Acesso em: 11 mar 2021. Citado na página 24.

RUSSELL, S.; NORVIG, P. *Inteligência artificial*. 3rd. ed. Rio de Janeiro: Elsevier, 2013. ISBN 978-85-352-3701-6. Citado na página 22.

SABIR, E. et al. Deep multimodal image-repurposing detection. In: *Proceedings of the 26th ACM International Conference on Multimedia*. New York, NY, USA: Association for Computing Machinery, 2018. (MM '18), p. 1337–1345. ISBN 9781450356657. Disponível em: <<https://doi.org/10.1145/3240508.3240707>>. Citado na página 25.

SANTOS, W. et al. Trendsbot: Verificando a veracidade das mensagens do telegram utilizando data stream. In: *Anais Estendidos do XXXVII Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos*. Porto Alegre, RS, Brasil: SBC, 2019. p. 65–72. ISSN 2177-9384. Disponível em: <[https://sol.sbc.org.br/index.php/sbrc\\_estendido/article/view/7771](https://sol.sbc.org.br/index.php/sbrc_estendido/article/view/7771)>. Citado na página 26.

SHARMA, K. et al. Combating fake news: A survey on identification and mitigation techniques. Association for Computing Machinery, New York, NY, USA, v. 10, n. 3, abr. 2019. ISSN 2157-6904. Disponível em: <<https://doi.org/10.1145/3305260>>. Citado na página 25.

SHEARER, E.; MITCHELL, A. *News Use Across Social Media Platforms in 2020: Facebook stands out as a regular source of news for about a third of americans*. Pew Research Center, 2021. Disponível em: <[www.journalism.org/2021/01/12/news-use-across-social-media-platforms-in-2020/](http://www.journalism.org/2021/01/12/news-use-across-social-media-platforms-in-2020/)>. Acesso em: 04 mar 2021. Citado na página 15.

Shu, K. et al. Detecting fake news with weak social supervision. *IEEE Intelligent Systems*, p. 1–1, 2020. Citado na página 25.

SHU, K. et al. Fake news detection on social media: A data mining perspective. *SIGKDD Explor. Newsl.*, Association for Computing Machinery, New York, NY, USA, v. 19, n. 1, p. 22–36, set. 2017. ISSN 1931-0145. Disponível em: <<https://doi.org/10.1145/3137597.3137600>>. Citado na página 22.

SHU, K. et al. (Ed.). *Disinformation, Misinformation, and Fake News in Social Media: Emerging research challenges and opportunities*. 1. ed. Springer Nature Switzerland AG: Springer International Publishing, 2020. (Lecture Notes in Social Networks). ISBN 9780262535045. Citado na página 25.

SHU, K.; WANG, S.; LIU, H. Beyond news contents: The role of social context for fake news detection. In: *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*. New York, NY, USA: Association for Computing Machinery, 2019. (WSDM '19), p. 312–320. ISBN 9781450359405. Disponível em: <<https://doi.org/10.1145/3289600.3290994>>. Citado na página 26.

SILVA, R. M. et al. Towards automatically filtering fake news in portuguese. *Expert Systems with Applications*, v. 146, p. 113199, 2020. ISSN 0957-4174. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S0957417420300257>>. Citado na página 25.

SOUZA, A. C. F. d. *Aplicativo Verific.AI-automatização de checagem de links de notícias no combate ao ecossistema da desinformação*. Dissertação (Mestrado) — Universidade Católica de Pernambuco, Recife, ago. 2019. Departamento de Pós-Graduação. Disponível em: <<http://tede2.unicap.br:8080/handle/tede/1184>>. Citado na página 26.

SOUZA, F.; NOGUEIRA, R.; LOTUFO, R. BERTimbau: pretrained BERT models for Brazilian Portuguese. In: *9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23 (to appear)*. [S.l.: s.n.], 2020. Citado 4 vezes nas páginas 7, 18, 30 e 38.

TANDOC JR., E. C.; LIM, Z. W.; LING, R. Defining “fake news” a typology of scholarly definitions. *Digital journalism*, Taylor & Francis, v. 6, n. 2, p. 137–153, 2018. Citado na página 19.

THEMUDO, T. S.; ALMEIDA, F. C. d. Direito, cultura e sociedade em tempos de fake news. *Revista de Direitos e Garantias Fundamentais*, v. 21, n. 3, p. 209–236, dez. 2020. Disponível em: <<https://sisbib.emnuvens.com.br/direitosegarantias/article/view/1653>>. Citado na página 24.

VASWANI, A. et al. Attention is all you need. In: . Red Hook, NY, USA: Curran Associates Inc., 2017. (NIPS’17), p. 6000–6010. ISBN 9781510860964. Citado 2 vezes nas páginas 22 e 30.

VOSOUGHI, S.; ROY, D.; ARAL, S. The spread of true and false news online. *Science*, American Association for the Advancement of Science, v. 359, n. 6380, p. 1146–1151, 2018. ISSN 0036-8075. Disponível em: <<https://science.sciencemag.org/content/359/6380/1146>>. Citado na página 16.

WOLF, T. et al. Transformers: State-of-the-art natural language processing. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics, 2020. p. 38–45. Disponível em: <<https://www.aclweb.org/anthology/2020.emnlp-demos.6>>. Citado na página 23.

ZELLERS, R. et al. Defending against neural fake news. In: WALLACH, H. et al. (Ed.). *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2019. v. 32. Disponível em: <<https://proceedings.neurips.cc/paper/2019/file/3e9f0fc9b2f89e043bc6233994dfcf76-Paper.pdf>>. Citado na página 25.

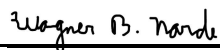
ZHANG, X.; GHORBANI, A. A. An overview of online fake news: Characterization, detection, and discussion. *Information Processing & Management*, v. 57, n. 2, p. 102025, 2020. ISSN 0306-4573. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S0306457318306794>>. Citado na página 25.

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, set. 2020. ISSN 0360-0300. Disponível em: <<https://doi.org/10.1145/3395046>>. Citado na página 25.

# TERMO DE RESPONSABILIDADE

Eu, **Wagner Bianchini Narde** declaro que o texto do trabalho de conclusão de curso intitulado “*Análise de notícias falsas em rede social: Uma abordagem utilizando transferência de aprendizagem e Transformers*” é de minha inteira responsabilidade e que não há utilização de texto, material fotográfico, código fonte de programa ou qualquer outro material pertencente a terceiros sem as devidas referências ou consentimento dos respectivos autores.

João Monlevade, 27 de abril de 2021



---

Wagner Bianchini Narde