**UNIVERSIDADE FEDERAL DE SÃO CARLOS**

CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

# DETECÇÃO DE POSTAGENS COM INFORMAÇÕES FALSAS SOBRE A PANDEMIA DO COVID-19 NA REDE SOCIAL INSTAGRAM

**MATEUS OLIVEIRA CABRAL**

**ORIENTADOR PROF. DR. RICARDO RODRIGUES CIFERRI**

São Carlos – SP

Julho/2021

**UNIVERSIDADE FEDERAL DE SÃO CARLOS**

CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

# DETECÇÃO DE POSTAGENS COM INFORMAÇÕES FALSAS SOBRE A PANDEMIA DO COVID-19 NA REDE SOCIAL INSTAGRAM

**MATEUS OLIVEIRA CABRAL**

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de São Carlos, como parte dos requisitos para a obtenção do título de Mestre em Ciência da Computação, área de concentração: Engenharia de Software, Banco de Dados e Interação Humano-Computador.

Orientador Prof. Dr. Ricardo Rodrigues Ciferri

São Carlos – SP

Julho/2021

Mateus Oliveira Cabral
Detecção de Postagens com Informações Falsas Sobre a Pandemia do COVID-19 na Rede Social Instagram/ Mateus Oliveira Cabral. – São Carlos – SP, Julho/2021-
109 p. : il. (algumas color.) ; 30 cm.

Orientador Prof. Dr. Ricardo Rodrigues Ciferri

– Universidade Federal de São Carlos, Julho/2021.

1. Conteúdo Falso. 2. Fake News. 3. Redes Sociais. 4. Instagram. 5. COVID-19. I. Prof. Doutor Ricardo Rodrigues Ciferri. II. Universidade Federal de São Carlos. III. Centro de Ciências Exatas e de Tecnologia. IV. Detecção de Notícias Falsas da Pandemia do COVID-19 na Rede Social Instagram.

**UNIVERSIDADE FEDERAL DE SÃO CARLOS**

Centro de Ciências Exatas e de Tecnologia
Programa de Pós-Graduação em Ciência da Computação

## Folha de Aprovação

Defesa de Dissertação de Mestrado do candidato Mateus Oliveira Cabral, realizada em 05/07/2021.

### Comissão Julgadora:

Prof. Dr. Ricardo Rodrigues Ciferri (UFSCar)

Profa. Dra. Marcela Xavier Ribeiro (UFSCar)

Prof. Dr. Thiago Alexandre Salgueiro Pardo (USP)

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

O Relatório de Defesa assinado pelos membros da Comissão Julgadora encontra-se arquivado junto ao Programa de Pós-Graduação em Ciência da Computação.

*Este trabalho é dedicado aos meus pais e minha irmã.*

# AGRADECIMENTOS

Primeiramente aos meus pais, em especial a minha mãe que sempre me incentivou a correr atrás dos meus objetivos, colocando muitas vezes como objetivos dela também. Gostaria de agradecer também ao meu orientador Prof. Dr. Ricardo Rodrigues Ciferri e também aos professores Msc. Pablo Freire Matos e Dr. Pedro Nobile por me auxiliarem na orientação dessa pesquisa de Mestrado com diversas sugestões e criticas construtivas para que o trabalho tivesse um conteúdo de qualidade. Também não poderia deixar de lembrar dos meus amigos Natalie, Beatriz, Núbia, Isabela, Hellen, Rodrigo, Matheus e Laís que sempre me ajudaram de todas as formas possíveis, mesmo que muito deles nem sejam da área de Ciência da Computação. Por último e não menos importante, agradeço aos membros do grupo de pesquisa de Banco de Dados da UFSCar que durante as reuniões que tivemos contribuíram com sugestões para a condução desse trabalho. Agradeço a todos aqui citados direta ou indiretamente, pois todos vocês contribuíram para o sucesso deste trabalho.

*Nós somos problemas que querem ser resolvidos, Nós somos crianças que precisam ser amadas!*
*(Pink, "What about us")*

*É um erro grave formular teorias antes de conhecer os fatos. Sem querer, começamos a mudar os fatos para que se adaptem às teorias, em vez de formular teorias que se ajustem aos fatos.*
*(Sherlock Holmes, "Um Escândalo na Boêmia")*

# RESUMO

Esta dissertação aborda a detecção de informações falsas no Instagram, a rede social que vem crescendo cada vez mais em comparação com as demais plataformas de redes sociais. Por se tratar de uma rede social com conteúdo multimídia (imagem, vídeo e texto), mas com ênfase na postagem de fotos, há pouca pesquisa científica dos impactos das postagens com informações falsas que essa rede proporciona na sociedade. Isso acontece principalmente em épocas de eleições políticas ou em acontecimentos históricos, em que existe uma grande demanda sobre informações. Por isso, essa pesquisa de Mestrado teve como domínio a área da saúde com ênfase no assunto da pandemia de COVID-19, assunto de extrema importância e impacto social. Muitos estudos abordam diversas técnicas para identificação de artigos de notícias falsas e/ou postagens falsas em redes sociais como Facebook, Twitter, Youtube e Whatsapp. Alguns estudos enfocam no conteúdo da notícia, outros estudos enfocam no contexto social por meio de informações das redes sociais que envolve análise de sentimento, enquanto para outros estudos o foco é o temporal muito analisado também sobre a dinâmica das postagens na rede social. Nesta pesquisa de Mestrado, a fonte escolhida para extrair dados de estudo, tem uma dinâmica funcional completamente diferente das demais redes sociais. O compartilhamento dos fenômenos que impactam a dispersão das notícias nas redes sociais não funciona da mesma forma no Instagram. Além disso, as imagens postadas podem conter textos dentro das imagens, o que gera a necessidade de utilizar ferramentas baseadas em *Optical Character Recognition* (OCR) para extrair os textos, para somente depois confrontar a informação extraída em postagens em português para classificar se é uma informação falsa ou verdadeira. Outro problema, além da falta de pesquisas sobre informações falsas relacionados ao Instagram, é a existência de poucos conjuntos de dados de conteúdos em português para análises e *benchmark* de modelos de detecção de informações falsas, principalmente que contenham imagens. O objetivo desta pesquisa de Mestrado foi investigar a detecção de postagens em português com informações falsas sobre a pandemia de COVID-19 na rede social Instagram. Nesse sentido, a pesquisa teve como resultado a proposta de um modelo de aprendizado de máquina que permite a detecção de informações falsas. Além disso, esta pesquisa realizou a compilação de um conjunto de dados relacionadas a COVID-19 para ser disponibilizada para futuras investigações sobre conteúdos falsos na rede social Instagram. O modelo foi validado por meio de testes experimentais com dados reais. Os resultados mostraram uma acurácia entre 96% e 99% na detecção de postagens com informações falsas sobre COVID-19.

**Palavras-chave**: conteúdo falso, fake news, redes sociais, Instagram, COVID-19.

# ABSTRACT

This dissertation addresses the detection of false information on Instagram, the social network that has been growing more and more compared to other social media platforms. Because it is a social network with multimedia content (image, video and text), but with an emphasis on posting photos, there are few scientific research on the impacts of posts with false information that this network provides on society. This happens mainly in times of political elections or in historical events, when there is a great demand for information. Therefore, this Master's research had as its domain the health area, with emphasis on the subject of the COVID-19 pandemic, a subject of extreme importance and big social impact. Many studies address various techniques for identifying fake news articles and/or fake posts on social networks such as Facebook, Twitter, Youtube and Whatsapp.Some studies focus on the content of the news, other studies focus on the social context through information from social networks that involves sentiment analysis, while for other studies the focus is on the temporal, which is also very much analyzed on the dynamics of posts on the social network. In this Master's research, the source chosen to extract study data has a functional dynamic that is completely different from other social networks. Sharing the phenomena that impact the dispersion of news on social media does not work in the same way on Instagram. In addition, the posted images may contain text within the images, which creates the need to use Optical Character Recognition (OCR) based tools to extract the texts, and only then compare the extracted information in posts in Portuguese to classify whether it is false or true information. Another problem, in addition to the lack of research on false information related to Instagram, is the existence of few content datasets in Portuguese for analysis and *benchmark* of false information detection models, especially those containing images. The aim of this Master's research was to investigate the detection of posts in Portuguese with false information about the COVID-19 pandemic on the Instagram social network. In this sense, the research resulted in the proposal of a machine learning model that allows the detection of false information. In addition, this research performed the compilation of a dataset related to COVID-19 to be made available for future investigations into fake content on the Instagram social network. The model was validated through experimental tests with real data. The results showed an accuracy between 96% and 99% in detecting posts with false information about COVID-19.

**Keywords**: fake news, social network, Instagram, COVID-19.

# LISTA DE SIGLAS

| | |
|---|---|
| OCR | *Optical Character Recognition* |
| USP | Universidade de São Paulo |
| RNA | Redes Neurais Artificiais |
| HTML5 | *Hypertext Markup Language version 5* |
| SVM | *Support Vector Machine* |
| API | *Application Programming Interface* |
| LIWC | *Linguistic Inquiry and Word Count* |
| PLN | Processamento de Linguagem Natural |
| BRET | *Bidirectional Encoder Representations from Transformers* |
| VGG-19 | *Visual Geometry Group with 19 Weight Layer* |
| NLTK | *Natural Language Toolkit* |
| AP | *Associated Press* |
| EDCM | *Exponential Dirichlet Compound Multinomial* |
| POS | *Part of Speach* |
| NER | *Named Entity Recognition* |
| TF-IDF | *Term Frequency–Inverse Document Frequency* |
| RNN | Redes Neurais Recorrentes |
| AM | Aprendizado de Máquina |
| MBR | Máquinas de Boltzmann Restritas |
| RoBERTa | *Robustly Optimized BERT Pretraining Approach* |
| BoW | *Bag of Words* |

| | |
|---|---|
| ACP | Análise de Componentes Principais |
| t-SNE | *t-Distributed Stochastic Neighbor Embedding* |
| DBNs | *Deep Belief Network* |
| MBR | Máquinas Boltzmann Restritas |
| ROC | *Receiver Operating Characteristic* |
| AUC | *Area Under Curve* |
| LBP | *Local Binary Pattern* |
| HOG | *Histogram of Oriented Gradients* |
| TAS | *Threshold Adjacency Statistics* |
| MLP | *Multi-layer Perceptron* |
| SVC | *Support Vector Clustering* |

# LISTA DE FIGURAS

# LISTA DE TABELAS

# SUMÁRIO

# Capítulo 1
# INTRODUÇÃO

## 1.1 Considerações Iniciais

Neste primeiro capítulo são descritos os objetivos desta pesquisa de Mestrado no contexto da detecção de postagens com informações falsas na rede social Instagram. Com o uso de técnicas do domínio de inteligência artificial é possível classificar com uma boa precisão se a informação postada é de fato verdadeira ou por outro lado é falsa. Antes da detecção da postagem é necessário o uso de um sistema de reconhecimento de caracteres, conhecido também como *Optical Character Recognition* (OCR). Com este sistema torna-se possível trabalhar com as imagens que possuem texto e são compartilhadas nas redes sociais. As imagens podem acrescentar mais uma camada de contexto que vai ajudar na classificação dessas postagens. Na seção 1.2 é detalhada a contextualização da investigação. Na seção 1.3 são apresentadas as motivações para o desenvolvimento desta pesquisa de Mestrado, destacando a necessidade da pesquisa para a solução parcial ou total do problema. Na seção 1.4 são discutidos os objetivos. Por fim, na seção 1.5 é detalhada a estrutura desta Dissertação.

## 1.2 Contextualização

O conceito de mídias sociais engloba tanto as redes sociais, como os *e-mails*, os *blogs* e outras ferramentas que produzem conteúdo digital para consumo em massa de usuários na internet. A utilização de mídias sociais tem crescido a cada ano, e tem cada vez maior aderência dos grupos de usuários *seniores*, aqueles com faixa etária acima de 60 anos (SHEARER; MATSA, 2018). As mídias sociais são muito utilizadas na transmissão de conhecimento, por exemplo, uma pessoa que saiba um idioma diferente do adotado pelo seu país, pode ajudar outras pessoas a aprenderem o novo idioma por meio de tutorias em mídias sociais, tal como no Youtube ou nas redes sociais Facebook e Instagram. As redes sociais ajudam as pessoas a se conectarem virtualmente uma as outras. Pode-se encontrar amigos antigos que o contato havia sido perdido há algum tempo ou se contactar com parentes que se mudaram para um local distante. Além das conexões, as redes sociais ajudam a disseminar informações, e vem sendo

uma poderosa ferramenta para o consumo de notícias dos mais variados domínios. Portanto, a principal funcionalidade das redes sociais é permitir a interação entre pessoas, em geral de forma assíncrona. De acordo com a *Pew Research Center* (SHEARER; MATSA, 2018), cerca de 69% da população adulta americana ocasionalmente se informa por meio das redes sociais. A principal justificativa utilizada pelos entrevistados é a conveniência e alguns dizem que também é por causa da possibilidade de interação com outros usuários da rede.

O grande problema de se informar por meio das redes sociais é a qualidade das informações que podem ser muito ruins e incorretas. Veículos jornalísticos, sejam eles rádios, TVs, jornais impressos, revistas ou *sites* de empresas jornalísticas tradicionais são todos conduzidos por meio de um sistema organizacional. Esse sistema contém um grupo de diretores, editores e jornalistas que verificam a informação, além de possuírem especialistas jurídicos, especialistas de tecnologias e entre outras áreas que ajudam na consolidação de um artigo, até mesmo de uma matéria mais complexa com inúmeras edições. Com muitas pessoas trabalhando e com um devido processo para a edição de uma notícia, é necessário tempo e investimento, por isso as mídias tradicionais compostas por grandes oligarcas eram a principal fonte de notícias para toda a população. De certa forma, a mídia tradicional pode ser considerada o quarto poder de um estado, sempre crítica e questionando ações de governos.

Devido ao avanço das mídias sociais, qualquer pessoa pode atuar como jornalista e não precisa de uma empresa para poder divulgar as notícias, elas podem ser publicadas em um simples *blog* ou em um *post* de rede social sem limite de caractere, para serem visualizadas ou compartilhadas com outras pessoas. Com essas ações, o alcance da mensagem aumenta, e com o aumento medido por meio dos números de acesso ou interações nas redes sociais, o jornalista pode utilizar apenas as redes sociais como fonte de faturamento por intermédio de artigos publicitários que é uma das principais fontes de renda em uma rede social.

Algumas notícias são propositalmente mal formuladas, muitas vezes tendenciosas ou escondem a parcialidade do veículo de divulgação da informação ou do jornalista que as escreve ou pública. Além desses aspectos, existem muitas notícias falsas escritas para prejudicar pessoas públicas, grupos políticos, partidos e ideologias políticas. O mundo das notícias falsas (ou *Fake News*) começou a se expandir intensamente no ano de 2016 com diversos acontecimentos políticos, três dos principais foram o *BREXIT*, o Impeachment do Presidente Dilma Rousseff e as eleições americanas. Com a ajuda das redes e mídias sociais a difusão das notícias falsas foi tão grande que vários pesquisadores começaram a estudar o fenômeno (BARRETT, 2019).

De acordo com Barrett (2019), as informações falsas podem ser classificadas em quatro grupos básicos: sátira, boatos, *clickbait* e notícias falsas. Sátiras não têm intenção de causar dano, mas tem potencial para enganar por meio do humor ou da falta de conhecimento de quem as lê. Boatos ou rumores, são os casos que misturam tanto a verdade quanto a mentira, além também de se tratar de assuntos aos quais sejam difíceis de contestar. O *clickbait* é o mais comum dentro das redes sociais e da mídia tradicional, utiliza-se do título ou subtítulo da notícia para chamar a

atenção do usuário com frases sensacionalistas ou tendenciosas, porém dentro do próprio corpo do texto o título pode se contradizer. Por último as notícias falsas, ou Fake News como são popularmente conhecidas, são matérias com aspecto jornalístico produzidas para unicamente e exclusivamente desinformar os usuários, para gerar polarização em determinados assuntos ou para apoiar ideias controversas e teorias conspiratórias.

## 1.3 Motivação

A filtragem e a classificação de uma informação falsa é uma tarefa complexa que vem sendo discutida de uma maneira multidisciplinar, pois adere as áreas da Ciência da Computação, Ciência Social, Comunicação e áreas específicas dependendo do contexto da informação falsa. Atualmente, a maneira mais utilizada é a de verificação de fatos, realizada por jornalistas e algumas vezes com o suporte de especialistas. Algumas das plataformas mais conhecidas em verificação de fatos são as agências Lupa *(https://piaui.folha.uol.com.br/lupa/)*, Aos Fatos *(https://aosfatos.org/)* e a mais famosa de todas a americana Politifact *(https://www.politifact.com/)*. Essas agencias funcionam de acordo com a demanda, quando uma notícia falsa é reportada e existe uma grande procura sobre ela, os jornalistas dessas agências verificam a notícia. Um fato importante de destaque é que em sua grande maioria as agências que verificam fatos são independentes de mídias jornalísticas e não possuem anúncios comerciais, pois ambos poderiam causar influência em decisões editoriais nas publicações.

Se uma notícia referência uma ação ocorrida no passado ou no presente, um jornalista que trabalha com a verificação dos fatos se debruça nas informações e por meio de um banco de arquivos digitais ou físico (papeis e imagens fotográficas) ele reconstrói a realidade da determinada notícia. Muitas vezes uma notícia falsa apresenta elementos reais, porém com uma visão completamente distorcida dos fatos, utilizando argumentos com muita persuasão para influenciar o leitor a pensar e interpretar de uma outra forma o fato narrado. Fatos não são flexíveis, eles são reais ou não, opiniões podem ser flexíveis dentro de uma notícia.

De maneira geral, os pesquisadores de Ciências Sociais explicam a alta tendência das informações falsas por meio de vários ingredientes, que combinados mostram a eficiência no sucesso de desinformar as pessoas. Um desses ingredientes é o momento de pós-verdade que denota circunstâncias nas quais fatos objetivos têm menos influência em moldar a opinião pública do que apelos à emoção e a crenças pessoais (FABIO, 2016). Outro amplificador são as bolhas ideológicas, algoritmos utilizados pelas redes sociais com o intuito de ser mostrado apenas o que agrada o usuário de acordo com as interações dele em sua rede (SALAS, 2015). O último ingrediente que é o alto escalonamento de uso das redes sociais, que de forma positiva democratizou o acesso a informação das pessoas, porém carece de ferramentas para restringir a circulação de informações falsas que podem vir a prejudicar pessoas ou grupos sociais.

No mundo todo existe esforços e pesquisas para relatar os impactos das informações

falsas e o alcance delas. Os problemas pesquisados variam da detecção de informação falsa ao compartilhamento delas e do alcance nacional ou mundial da informação falsa. A rapidez de propagação em uma rede social e em outras formas de mídias (TV e rádio) pode ser muito grande, devido ao fato que nas redes sociais há ferramentas disponíveis para elevar a dispersão da mensagem como a utilização de robôs, a retransmissão de notícias em cascatas ocasionados por amigos compartilhando postagens de amigos que já estavam também compartilhando textos de um outro lugar.

Segundo a pesquisa do DataSenado (2019), 45% dos entrevistados se sentem influenciados pelas redes sociais na hora de votar, o que é preocupante tendo em vista que há um grande volume de notícias falsas que podem atrapalhar ciclos de eleições federais, estaduais e municipais que no Brasil por exemplo são a cada 2 anos, alternando entre eleições que são eleitos prefeitos e vereadores, e as que elegem deputados, governadores, senadores e presidente. O motivo da adesão de um grande grupo de pessoas as notícias falsas são por meio do compartilhamento de notícias entre pessoas de uma mesma família ou círculo de amizade. Portanto, há um nível de confiança devido as relações interpessoais, logo o usuário dificilmente vai verificar por outros meios se aquela notícia é verdadeira. Além disso, para se realizar a verificação se a mensagem não é uma notícia falsa leva-se algumas horas, algo que pode não ser estimulante as pessoas que não têm muito tempo no seu dia-a-dia para se informar.

Uma pesquisa proposta pelo Monitor do Debate Político no Meio Digital, da Universidade de São Paulo (USP) (GRAGNANI, 2018) concluiu que uma grande parte da reprodução de notícias falsas sobre o assassinato da vereadora carioca Marielle Franco, ocorreu em grupos de família da rede social WhatsApp. Esses boatos foram espalhados na mesma noite em que ela foi assassinada, e nos dias posteriores começaram a ser dispersos nas redes sociais Twitter e Facebook. O foco da pesquisa foi a investigação de padrões de distribuição da notícia, como o aplicativo é de mensagens privadas e não tem caráter público, é extremamente difícil rastrear a fonte da notícia, e avaliar o alcance no WhatsApp. Do grupo de pessoas que participaram dessa entrevista 51% responderam ter recebido o texto no WhatsApp em grupos de família, 32%, em grupos de amigos e 9% em grupos de colegas de trabalho.

A principal motivação para se combater as notícias falsas deve ser a busca pelos fatos e também para combater o mal que algumas dessas informações podem causar. A respeito dos fatos cabe a cada pessoa buscar a sua verdade quando o fato é passível de interpretação. Porém para se obter uma fonte confiável de informação é extremamente necessário haver uma apuração dos detalhes por jornalistas, cientistas especialistas quando se tratar de casos específicos. Entretanto, com um grande volume de dados que são enviados cotidianamente nas redes sociais, a tarefa pode não ser tão simples e mecanismos devem ser criados para facilitar as apurações. As redes sociais começaram a sofrer com os impactos negativos causados pelos atos anteriormente citados que de certa forma influenciam o mundo real. Por exemplo, em um dos vários casos o comitê de inteligência dos Estados Unidos confirmou a interferência de agentes russos nas eleições

presidenciais americanas devido a artigos publicitários do Facebook, muito deles utilizando notícias falsas para deteriorar relações de grupos sociais uns com os outros (JALONICK, 2018). Outro exemplo foi a rede social WhatsApp confirmando que houve disparos em massas de notícias falsas em grupos da sua rede social nas eleições presidências brasileiras, algo que ainda está sendo investigado desde meados de 2018 pela imprensa e desde 2019 pelo tribunal eleitoral (CURY, 2019). O impacto do compartilhamento dessas informações podem ser dos mais variados possíveis. Em março de 2018, a vereadora do Rio de Janeiro Marielle Franco foi executada com vários tiros dentro do seu carro juntamente com seu motorista no centro do Rio de Janeiro. A morte da vereadora foi altamente divulgada por ser um crime bárbaro e a vereadora ser uma pessoa pública. Várias notícias começaram a circular nas redes sociais, algumas com informações sobre o caso, outras sobre a trajetória da vereadora e algumas notícias falsas. De acordo com a pesquisa de, divulgada também pelo jornal O Globo, a notícia mais compartilhada na internet sobre a morte da vereadora era uma notícia falsa. A fonte da notícia falsa saiu de uma página chamada Ceticismo Político, pagina que na época era ligado ao grupo político Movimento Brasil Livre. A notícia vinculava a imagem de Marielle a um traficante da facção do comando vermelho, acusava a vereadora de ser usuária de drogas e que o motivo do homicídio era que a vereadora havia descumprido ordens do comando vermelho e por isso havia sido executada. As alegações vieram de uma desembargadora do tribunal de justiça do Rio de Janeiro. Por meio de uma matéria o portal G1 rebateu todas as afirmações prestadas pela desembargadora, que foi a fonte da notícia da página Ceticismo Político (G1, 2018). Um outro caso que chocou a comunidade de Morrinhos em Guarujá. Uma mulher foi confundida por moradores com outra mulher que supostamente estaria sequestrando crianças para rituais de magia negra. Houve um linchamento por moradores dessa comunidade que consideraram que Fabiane de Jesus era a mesma pessoa do retrato falado disponibilizado por uma rede social. Depois da agressão a vítima foi levada ao hospital Santo Amaro, gravemente ferida e morreu dois dias depois do ataque. O ato foi filmado e compartilhado nas redes sociais por meio de celulares. Cinco homens foram condenados pelo homicídio de Fabiane de Jesus, porém a página do Facebook Guarujá Alerta, que publicou o retrato falado e uma foto de uma suposta sequestradora não sofreu represálias, pois na época a justiça não tinha meios legais de penalizar quem incitar atos violentos pela internet ou criação de notícias falsas. (PAULO, 2018).

Muitas pesquisas vêm sendo desenvolvidas nas redes sociais comentadas nesse capítulo, entretanto o volume delas variam bastante. O Facebook e Twitter possuem investigações concluídas e sendo conduzidas para a identificação de padrões de compartilhamento das notícias com o objetivo de criar um modelo computacional que faça esse trabalho automaticamente. Outras pesquisas tentam classificar uma notícia em falsa ou verdadeira com auxílio da maneira como foi escrita a notícia, utilizando sintaxe, semântica e outros recursos da linguagem que podem concluir por meio de pontuações que classifica todo o artigo da notícia. Algumas pesquisas tentam encontrar o nível de difusão de uma notícia e os caminhos em que elas podem percorrer por toda a rede social, e a forma que os usuários reagem a tal postagem, ou se os usuários são

usuários reais, pois alguns usuários podem ser robôs virtuais programados por uma pessoa para interagir de forma automática com as postagem aumentando a exposição da mensagem.

O Instagram é uma rede social que não é amplamente utilizada como foco de pesquisa segundo um estudo do departamento de direitos humanos da Universidade de Nova Iorque feito no final de 2019 (BARRETT, 2019). Eles entendem que redes sociais como Facebook, Twitter e Youtube tem tido um maior interesse para pesquisas sobre as eleições americanas de 2016. Porém, a plataforma Instagram pode ter tido um impacto maior do que elas e ter sido desconsiderada por todo esse tempo. Esta plataforma é do tipo multimídia, na qual o maior interesse dos usuários é postar imagens e vídeos, e não muita exposição de textos ou artigos quanto as demais redes sociais. Portanto existe uma grande necessidade de compreender as relações dos ciclos de notícias dentro do Instagram e analisar o comportamento dos usuários dentro da plataforma, que mesmo sendo uma rede social, possui diferentes mecanismos para o compartilhamento de postagens entre os usuários da própria rede.

## 1.4 Objetivo

Essa pesquisa de Mestrado teve como objetivo principal a detecção de postagens falsas em português a respeito do COVID-19 publicadas na rede social Instagram. Para extrair textos em imagens publicadas no Instagram, foi usada uma ferramenta de Reconhecimento Óptico de Caractere. O escopo da detecção de notícias falsas foi o assunto da pandemia do corona vírus (COVID-19) pela *hashtag* "vírus chinês", que foi amplamente utilizada no primeiro semestre de 2020. Foi separado e filtrado apenas imagens, e os vídeos foram descartados, pois o escopo dessa pesquisa vai envolver apenas a extração de textos em imagens.

Com base nesse objetivo, define-se a seguinte hipótese:

*O uso de OCR e de técnicas de processamento de linguagem natural e aprendizado de máquina permite a detecção de informações falsas, por meio da extração de informação resultante do processamento da imagem e posterior tratamento linguístico sobre o texto na classificação de uma informação falsa.*

Com a investigação da hipótese, é esperado obter as seguintes contribuições:

- Modelar um esquema para o armazenamento dos conteúdos do Instagram que facilite o processamento das postagens e a exportação delas.
- Implementar as tarefas de aprendizado de máquina e Processamento de Linguagem Natural (PLN) para encontrar algoritmos com melhor acurácia de detecção de postagens falsas.
- Criar um conjunto de dados de informações já verificadas (curadas) por *sites* de verificações de fatos.

- Determinar o quanto os textos das imagens podem contribuir em um modelo de aprendizado de máquina.

## 1.5 Estrutura da Dissertação

Esta dissertação de Mestrado está organizada da seguinte forma: o Capítulo 2, descreve os principais conceitos sobre as ferramentas e tecnologias utilizadas neste trabalho, no Capitulo 3 existe uma síntese dos principais trabalhos relacionados dessa pesquisa, abrangendo desde os trabalhos *baselines* até as últimas pesquisas realizadas que compõe o estado da arte realizada ao longo da pesquisa. No Capitulo 4 estão descritos os pontos principais, esquematização desse projeto, os resultados obtidos na pesquisa com a detecção de postagens com informações falsas sobre a COVID-19 e também uma análise sobre os melhores modelos, os classificadores que apresentaram os melhores resultados. Por fim, o Capitulo 5 apresenta as conclusões desta pesquisa e uma lista de trabalhos futuros.

# Capítulo 2
# FUNDAMENTAÇÃO TEÓRICA

## 2.1 Considerações Iniciais

Um dos grandes objetivos de estudo na área de Ciência da Computação é a produção de máquinas que automatizem tarefas antes feitas apenas manualmente e exclusivamente por humanos, de forma que as máquinas consigam realizar as tarefas de forma automática ou semiautomática com o mínimo de intervenção humana. Uma dessas tarefas que está em processo de automatização é a leitura de textos. A escrita também é uma tarefa realizada por humanos e vem sendo a forma mais natural de coletar, armazenar e transmitir informações por vários séculos. Escrita e Leitura não são mais um meio exclusivo de comunicação entre pessoas, eles são também tarefas que podem ser realizadas entre máquinas e entre máquinas e pessoas. Nos dias atuais é comum digitalizar os textos impressos para algum formato eletrônico, seja este formato de uma imagem ou uma extensão de leitura, como *pdf*. Dependendo da forma que o arquivo foi digitalizado, ele pode ser editado, pesquisado, armazenado ou enviado pela internet para um ou mais usuários. Documentos em formatos digitais economizam espaço físico e são sustentáveis, pois não é necessário realizar cópias físicas do documento. Além disso, o documento pode ser replicado ou acessado simultaneamente entre diversos usuários. Os textos digitais são mais fáceis para se visualizar em uma tela de computador e poder executar processos, tais como tradução de textos, narração de textos ou mineração de textos (YAMASAKI, 1978; BUNKE; WANG, 1997).

O Processamento da Linguagem Natural (PLN) é um importante campo de pesquisa da área da Ciência da Computação e de suas subáreas da inteligência artificial e linguística computacional. PLN objetiva processar a linguagem natural humana por meio do uso de tecnologias computacionais e/ou da linguística computacional. Desta forma, PLN está preocupado com as interações entre computadores e linguagens humanas (naturais). Em outras palavras, PLN realiza o processamento automático (ou semiautomático) da linguagem humana.(HUTCHINSON, 2015)

De modo mais genérico, o Aprendizado de Máquina (AM) é o campo (e a arte) da computação que aprende com os dados. Para uma melhor definição podemos recorrer a definição cunhada pelo Dr. Yoshua Bengio, professor da Universidade de Montréal e um dos maiores pesquisadores da área: “A pesquisa em aprendizado de máquina é um campo de estudo dentro da

pesquisa em inteligência artificial, que busca fornecer conhecimento aos computadores por meio de dados, observações e interações com o mundo."(FENNER, 2019)

Na primeira seção 2.2 são apresentados os principais conceitos de *Opitical Caracter Recognition* (OCR). A seção 2.3 define os principais conceitos sobre Processamento de Linguagem Natural (PLN). A penúltima seção 2.4 de Aprendizado de Máquina traz os conceitos mais importantes desse campo de estudo dentro da Ciência da Computação e a última seção 2.6 apresenta as considerações finais desse capitulo.

## 2.2 Reconhecimento Óptico de Caracteres

Reconhecimento Óptico de Caracteres ou *Optical Character Recognition* (OCR) é uma importante área de pesquisa que está dentro do domínio de inteligência artificial e visão computacional (MORI et al., 1999). Vem sendo utilizada tanto por aplicações comerciais quanto por trabalhos acadêmicos, devido ao seu grande potencial. A intensificação das pesquisas nesse campo não foram apenas para simular a tarefa de leitura, mas também para prover aplicações automatizadas eficientes e que tenham a capacidade de processamento em grande escala de textos, convertendo-os em dados digitais que possam ser inseridos em páginas *web*, bancos de dados, simples arquivos de texto ou arquivos do tipo *pdf* editável. A técnica de OCR consiste em transcrever letras que foram escritas a mão, impressas ou digitalizadas para textos codificados em uma linguagem que a máquina compreenda. Ao longo do processo, os caracteres são lidos, reconhecidos e gravados. Essa técnica é utilizada como uma forma de carregar dados que foram impressos em papéis para bases de dados. OCR pode ser definido também como um processo de classificação de padrões ópticos contidos em uma imagem digital que correspondem a um número ou uma letra alfanumérica. Esse processo é alcançado por meio de passos importantes como segmentação, extração de recursos e classificação, tais etapas serão abordadas na Seção 2.2.2.

Alguns exemplos do uso de OCR são: uso de passaportes, criação de base de dados para tomadas de decisão que agregue os recibos e extratos de transações e organização e separação de correspondências (CHAUDHURI et al., 2017). A tecnologia do processo de OCR permite converter diferentes tipos de documentos sejam eles escaneados, arquivos do tipo *pdf* ou imagens (CHAUDHURI et al., 2017) captados por câmeras digitais em dados editáveis e pesquisáveis.

### 2.2.1 História

A origem do reconhecimento de caracteres é datada do ano de 1870 quando um inventor americano chamado Charles R. Carey desenvolveu um *scanner* de retina (MORI et al., 1999). Esta máquina de transmissão de imagem utilizava um mosaico de fotocélulas e nas suas primeiras versões havia a necessidade de treinar as imagens de cada fonte, de cada caractere, um por vez (BUNKE; WANG, 1997; CHAUDHURI et al., 2017). A história do OCR começou a ser delineada

próximo de 1900, quando um cientista russo chamado Tyurin tentou projetar um aparelho para deficientes visuais(SCHANTZ, 1982). Duas décadas mais tarde Paul Nipkow inventou um *scanner* sequencial e foi um grande avanço tanto para TVs quanto para máquinas de leitura (RICE et al., 1999). Os primeiros estudos que enfocam especificamente em algo parecido com o atual processo OCR datam a década de 30 e se originaram na Alemanha com uma patente de Gustav Tauschek, que desenvolveu o primeiro dispositivo com um sensor de luz que apontava as palavras as quais ele correspondia em um modelo salvo na memória do dispositivo (SCHANTZ, 1982; BUNKE; WANG, 1997; MORI et al., 1999).

A visão moderna do OCR não apareceu até meados de 1940. A motivação para o desenvolvimento de sistemas de OCR começou por meio da visão de possibilidades de uso e aplicações comerciais. Os primeiros reconhecedores de caracteres começaram a aparecer nessa época devido ao avanço das tecnologias de *hardware* dos computadores. O primeiro trabalho em reconhecimento automático de caracteres enfocou em caracteres impressos de máquinas (datilográficas ou impressoras matriciais) ou por pequenos grupos de letras escritas a mão que eram bem legíveis e distintas. Nesse período, os sistemas OCR geralmente usavam um modelo para comparar as imagens dos caracteres impressos e os que estavam salvos em uma biblioteca de imagens. Para textos que foram escritos a mão eram transformados em imagens binárias onde são extraídos os vetores de recursos que ao alimentar os classificadores estatísticos por meio de cálculos escolhem o caractere que mais se aproximava a aquele que estava sendo analisado (SCHANTZ, 1982; BUNKE; WANG, 1997).

Por volta de 1950, a revolução tecnológica estava avançando em grande velocidade e o processamento de dados estava começando a se tornar um campo importante para pesquisas. No meio de 1950, as máquinas OCR se tornaram comercialmente disponíveis. A primeira máquina de leitura OCR foi instalada em uma empresa chamada Reader's Digest. Este equipamento foi usado para converter letras impressas e relatórios de vendas em cartões de punção para dados que foram depois armazenados em computadores (YU; JUTAMULIA, 1998).

Os sistemas comerciais OCR começaram a aparecer entre 1960 a 1965, uma época frequentemente referenciada como primeira geração de OCR. As máquinas OCR desta geração foram caracterizadas pelo formato e padronização das letras. Os símbolos foram especialmente desenhados para leitura de máquina. Quando máquinas com múltiplas fontes começaram a surgir, eles puderam realizar uma leitura de diferentes fontes com bastante precisão de acerto. Era permitido um número bem limitado de fontes pelo padrão de reconhecimento, método aplicado e combinado ao modelo ao qual compara a imagem do caractere com as bibliotecas de imagens de cada caractere de cada tipo de fonte (BUNKE; WANG, 1997; RICE et al., 1999).

No meio de 1965 e no início de 1970, as máquinas de leitura da segunda geração apareceram. Estes sistemas foram capazes de reconhecer caracteres impressos e os caracteres que eram escritos a mão começaram a ser explorados também. Quando as letras escritas a mão começaram a ser consideradas, o grupo de caracteres eram limitados e agrupados em duas

categorias compreendidas em letras e números. O primeiro sistema e o mais famoso foi a IBM 1287 em 1965. Durante este período, a empresa Toshiba desenvolveu a primeira máquina de alto desempenho que separava códigos postais. A empresa Hitachi também fez sua primeira máquina OCR de alto desempenho e baixo custo. Nesse período, o trabalho estava focado na área de normalização. Em 1966, um estudo de requerimentos de OCR foi concluído e um padrão de caracteres OCR americano foi definido como OCR-A, sendo que esta fonte foi altamente estilizada e desenhada para facilitar o reconhecimento óptico como pode ser vista na Figura 1 (a). Um modelo de fonte Europeia foi também criado, o OCR-B, ao qual tem uma aparência mais natural com formas arredondadas do que possui o padrão americano e pode ser visto na Figura 1 (b). Algumas tentativas foram feitas para que ambos padrões fossem utilizados nas máquinas dessa geração, mas os desenvolvedores desta época não obtiveram sucesso (SCHANTZ, 1982).

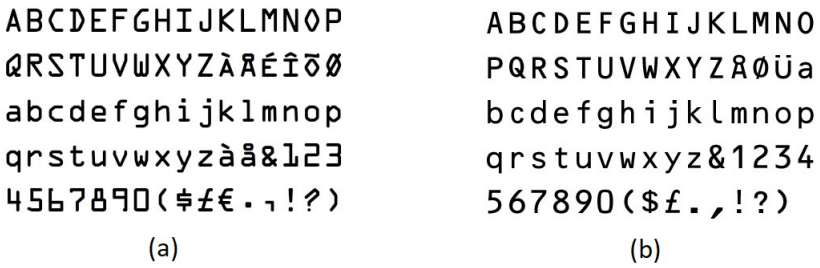


**Figura 1 – Padrões OCR**

**Fonte: (RICE et al., 1999)**

No meio de 1970 iniciou-se a terceira geração de sistemas OCR. O obstáculo era lidar com documentos de baixa qualidade e grandes grupos de palavras (escritas a mão ou impressas). O baixo custo e a desempenho foi um objetivo atendido nessa época com os avanços tecnológicos na parte de *hardware* resultando no desenvolvimento de máquinas OCR sofisticadas. No período anterior, os computadores pessoais e impressoras a *laser* começaram a dominar a área de produção de texto, a digitação era um nicho especial para o OCR. O espaçamento uniforme de impressão e o pequeno número de fontes fez os dispositivos OCR começarem a se tornarem mais simples e eficazes (MORI et al., 1999).

Nas próximas décadas e até hoje, o foco dos sistemas OCR deixou de ser em *hardware* e passou a ser em *software*, pois houve grandes avanços na capacidade de processamento dos computadores e a alta resolução para se digitalizar qualquer documento, seja por um *scanner* ou câmeras, ambos avanços são o ponto chave dos principais problemas para OCR. O principal tópico e questão que está tentando ser resolvida atualmente no domínio do OCR é o reconhecimento de caracteres feitos a mão, e em diferentes tipos de caractere latinos, orientais e outros idiomas. Existem diversos sistemas OCR *open source* e muitos outros proprietários e pagos, algumas distribuições serão comentadas na seção 2.2.3.

### 2.2.2 Metodologia

Para que seja possível a aplicação de OCR é necessário inicialmente treinar a máquina com as possíveis classes de padrões que possam vir a ocorrer e como os padrões podem vir a parecer. Em OCR, os padrões são letras, números e alguns símbolos especiais como pontuações. O aprendizado de máquina produz diferentes desempenhos de exemplos de caractere de todas as diferentes classes que possam ser atribuídas. Baseado nesses exemplos, a máquina constrói um modelo ou descrição de cada classe de caracteres. Durante o reconhecimento, caracteres que não foram reconhecidos são comparados com os previamente obtidos e são classificados com os que tiveram a melhor correspondência. Na maioria dos sistemas de OCR de caráter comercial para a fase de reconhecimento de caractere é necessário realizar uma fase de treinamento. Alguns sistemas, entretanto, incluem ferramentas de treinamento para o caso de inclusão de novas classes de caracteres.

Um típico sistema OCR consiste de várias etapas. A primeira etapa é a digitalização do documento usando um *scanner* óptico. Quando as regiões que contêm texto são localizadas cada símbolo é extraído por meio do processo de segmentação. Os símbolos que foram extraídos são pré-processados, elimina-se o ruído para facilitar a extração de recursos. A identidade de cada símbolo é descoberta por comparações dos recursos extraídos com as descrições das classes de símbolos conseguidas por intermédio de processos de aprendizados prévios. Finalmente a informação contextual é usada para reconstruir as palavras e números do texto original.

Na Figura 2 encontra-se as etapas de processamento OCR que serão abordados nos parágrafos a seguir de acordo com a ordem em que são trabalhadas na metodologia segundo Chaudhuri et al. (2017) e Bunke e Wang (1997).

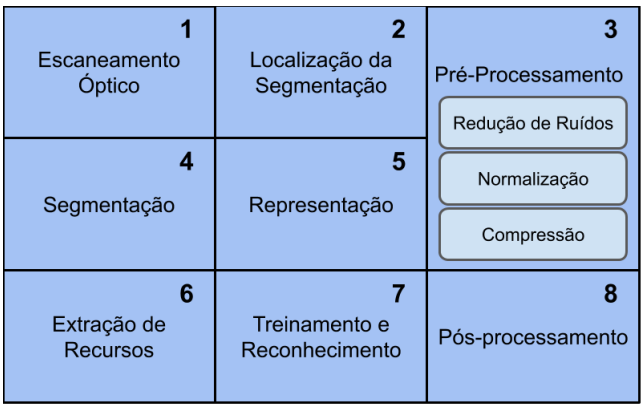


**Figura 2 – Etapas do OCR**

**Fonte: O próprio autor**

A primeira etapa do processo de OCR é o escaneamento óptico. Nesta etapa, a imagem é capturada do documento original por meio de escaneamento. Os *scanners* ópticos do OCR são mecanismos que fazem a transposição do documento com dispositivos sensíveis que convertem

a intensidade da luz aplicada no papel e transforma o resultado dessa captura em imagens em níveis de cinza. Geralmente os documentos impressos consistem em tinta preta em um fundo branco. Quando se opera OCR numa imagem multinível, esta imagem é convertida para o modo bi nível de preto e branco. Esse processo é conhecido como *thresholding*, sendo realizado no *scanner*. O objetivo é economizar espaços de memórias e melhorar o desempenho computacional. O processo de *thresholding* é importante para o resultado do processo geral e que depende totalmente da qualidade da imagem bi nível. Utiliza-se um *threshold* fixo a um nível de cinza, e os níveis abaixo desse ponto são considerados pretos e os níveis acima são considerados brancos. Para altos contrastes no documento com fundos uniformes é escolhido um *threshold* médio que possa ser o suficiente na classificação de ambas as cores (preto e branco). Nos casos de imagens que sejam multiníveis é possível que o ponto de *threshold* esteja em uma grande faixa de valores. Nesse caso é mais eficiente utilizar métodos mais modernos de *thresholding* que vão permitir melhores resultados. (CHAUDHURI et al., 2017; BUNKE; WANG, 1997).

Com relação a segunda etapa de OCR, a segmentação e localização determinam elementos da imagem que serão considerados como áreas candidatas e as áreas que serão descartadas. É necessário localizar regiões dos documentos as quais tenham texto e distingui-las de figuras ou gráficos. Quando se organiza automaticamente correspondências por endereços de envelopes deve ser localizada a linha com a transcrição do endereço e separá-la de outras imagens como selo ou logo de companhias. Quando aplicado a um texto, a segmentação isola os caracteres aos quais irá reconhecer individualmente. Geralmente a segmentação é executada com o isolamento de cada componente mesmo que estejam conectados. Esta técnica é fácil de ser implementada, mas problemas podem aparecer se o caractere estiver muito próximo ou fragmentado um do outro. Os principais problemas deste método é distinguir ruídos do texto ou má interpretação no processo de leitura dos textos. As principais técnicas utilizadas nessa etapa são algoritmos de segmentação, como algoritmos de bordas que utiliza os *pixels* de contorno das imagens para delimitar a área que será analisada (CHAUDHURI et al., 2017; BUNKE; WANG, 1997).

Já na terceira etapa de OCR, o componente de pré-processamento tem como objetivo produzir dados que sejam mais fáceis para os sistemas de OCR processar, além de ajudar na acurácia. Os objetivos principais do pré-processamento são: redução de ruídos, normalização dos dados e compressão de grupos de informação retidos. Os dados brutos dependendo de como foram adquiridos estão sujeitos a processos preliminares para que sejam utilizáveis nas demais etapas. O resultado da imagem escaneada pode conter alguns ruídos, dependendo da resolução do *scanner* e do *thresholding*, os caracteres podem estar embaçados ou até mesmo ilegíveis. Alguns desses defeitos podem causar uma baixa taxa de reconhecimento (CHAUDHURI et al., 2017; BUNKE; WANG, 1997).

Na quarta etapa de OCR a imagem do caractere é segmentada em subcomponentes (grupo de caracteres). A segmentação interna é usada para isolar linhas e curvas nas letras escritas em letras cursivas. Existem vários métodos que podem ser implementados e uma

variedade de técnicas para a segmentação. Porém, os três tipos são agrupados em segmentação explícita, implícita e mista. Na segmentação explícita, os segmentos são identificados baseados nas propriedades do caractere. O processo de cortar a imagem do caractere em algo que traga significado é possível por meio deste método de dissecção. O critério para uma boa segmentação é a compreensão de propriedades gerais de segmentos com os que são esperados para validá-los. Os métodos disponíveis baseados em dissecação do caractere utilizam espaço em branco, análise de projeção vertical e análise de componentes conectados, os quais podem ser utilizados para uma avaliação de contexto linguístico. Com a segmentação implícita, a estratégia é baseada no reconhecimento, feito por pesquisas nos componentes da imagem que combinam com as classes predefinidas, seu desempenho usa a taxa de confiança de reconhecimento incluindo sintaxes ou semânticas no resultado geral. A imagem é dividida sistematicamente em várias sobreposições em relação ao conteúdo. Estes métodos originalmente vêm de esquemas desenvolvidos para reconhecimento de palavras impressas por máquinas. Esta abordagem de reconhecimento é baseada no uso de probabilidade, o conceito de regularidade, singularidades e combinações *backward*. A mistura das duas segmentações anteriores dá origem a estratégias mistas que empenha um melhor resultado comparado com as anteriores. O erro de detecção e os mecanismos de correção são frequentemente embutidos dentro do sistema, O uso sensato do contexto e o classificador de confiança geralmente conduzem para uma melhor acurácia.

A quinta etapa de OCR é a representação da imagem e tem um papel muito importante em qualquer sistema de reconhecimento. No caso mais simples, o nível de cinza ou imagens binárias são alimentados pelo reconhecedor. Entretanto, na maioria dos sistemas de reconhecimento, com o intuito de se evitar complexidades extras e para aumentar a acurácia dos algoritmos, uma representação mais compacta é uma característica requerida. Por isso uma série de atributos são extraídos para cada classe de padrões de forma a ajudar a diferenciar uma classe de outra. A grande meta da representação é extrair e selecionar um grupo de atributos que maximizem a taxa de reconhecimento com o menor grupo de elementos. Algumas funções e técnicas utilizadas nessa etapa são transformadas de Fourier, transformada de Gabor e expansão de Karhunen Loeve (CHAUDHURI et al., 2017; BUNKE; WANG, 1997).

A sexta etapa é a extração de recursos e seu objetivo é capturar características essenciais dos símbolos. A extração de recursos é conhecida por ser um dos problemas mais difíceis para o padrão de reconhecimento. A maneira mais direta de descrever um caractere é por meio da rasterização da sua imagem. Uma outra abordagem é extrair certos recursos que caracterizem símbolos, mas que desconsiderem os atributos não tão importantes. Outra tarefa importante é a classificação que é um processo de identificar cada caractere e assinalar a classe correta. As duas abordagens de OCR para classificação são as de decisões teóricas e métodos estruturais.

O penúltimo componente de um sistema OCR utiliza metodologias de padrões de reconhecimento que atribui uma amostra desconhecida em uma classe predefinida. O OCR é investigado em quatro abordagens de padrões de conhecimento, eles são a combinação de uso de

modelos, técnicas estatísticas, técnicas estruturais e Redes Neurais Artificiais. Algumas técnicas estatísticas maximizam a probabilidade de o padrão observado pertencer a determinada classe, tais como as técnicas Gaussian, grupos fuzzys e hidden markov. As técnicas estruturais utilizam métodos gramaticais ou métodos gráficos para auxiliar, como árvores, diagramas ou grafos. As Redes Neurais Artificiais (RNA) possuem uma arquitetura paralela bem massiva tanto que executa computações em uma alta taxa comparada com técnicas clássicas. Adapta-se a mudanças nos dados e aprende as características do sinal de entrada. RNA contém vários nós, a saída deles alimentam uns aos outros na rede e a decisão final depende da complexidade dessa interação de todos os nós.

Na última etapa é comum usar atividades que incluem agrupamento e detecção de erros e correção de erros. No agrupamento, os símbolos no texto são associados com as *strings*. O resultado do reconhecimento de símbolos de caracteres especiais no texto é um grupo individual de símbolos. Entretanto esses símbolos não têm usualmente uma boa quantidade de informação e para superar isso são associadas outras palavras a estes símbolos. O agrupamento de símbolos em *strings* é baseado na localização do símbolo no documento e, assim, sinais próximos são agrupados juntos. Outras abordagens são o uso de dicionários, que são os mais eficientes meios de detecção e correção de erros. Dado uma palavra que pode estar errada e uma palavra que foi buscada no dicionário, se a palavra não estiver no dicionário, um erro é detectado e depois o erro é corrigido pela troca de palavras dentro de algo mais similar que conste no dicionário.

### 2.2.3 Tesseract

Como discutido em seções anteriores, a acurácia de um sistema OCR depende de uma boa segmentação e pré-processamento de textos. Algumas vezes é difícil recuperar o texto da imagem por causa do tamanho ser diferente, não ter o mesmo estilo ou orientação ou ainda ter um fundo colorido que dificulta a legibilidade do texto. Sistemas OCR permitem que a máquina reconheça o texto automaticamente, atuando como uma combinação dos olhos e da mente humana. Um olho pode ver um texto nas imagens, mas o cérebro que processa e interpreta o que é extraído pela leitura é feita pela visão. Alguns dos sistemas mais utilizados em OCR são Tesseract, Ocropus, GOCR, Abbyy Finereader e Omnipage. Sendo os três primeiros *open source* e os dois últimos *software* comercial (proprietário) pago. Além desses sistemas existem várias outras plataformas de OCR *online*, uma delas é o Google Doc. Basta adicionar um arquivo no formato *pdf* ou imagem com texto na conta do Google Drive e abrir com o Google Docs que a plataforma irá utilizar o OCR Tesseract para fazer a conversão. Na próxima subseção será apresentado o funcionamento do sistema Tesseract.

O Tesseract é um sistema de OCR *open source*, desenvolvido pela HP entre os anos de 1984 e 1994, e que foi modificado e melhorado em 1995 apresentando boa acurácia. Em 2005, o sistema foi alterado para uma distribuição *open source*, facilmente móvel, com foco em prover menos foco em rejeição que acurácia.

O Tesseract trabalha em 5 passos demonstrados na Figura 3 (SMITH, 2013). No primeiro bloco o sistema aplica o *thresholding* na imagem para poder convertê-la em imagem binária facilitando a segmentação e a busca por áreas textuais dentro da imagem. O próximo bloco é a Análise que é usada para extrair os esboços dos caracteres. Inicialmente utiliza-se a segmentação para dividir os espaços da imagem em blocos de texto, que posteriormente é dividido em linhas, e por último em palavras. Em casos de livro, onde se tem várias páginas, essas etapas são realizadas em cada uma das páginas de forma individual. Existem dois módulos de reconhecimento de caracteres: o Reconhecimento de Palavra e o Reconhecimento de Palavra 2. O primeiro permite que um classificador adaptável em tempo real seja treinado e utilizado na primeira passagem pela imagem e, caso exista um resultado insatisfatório, a palavra é novamente trabalhada em uma segunda passagem no modulo de Reconhecimento de Palavra 2. Os caracteres individuais de cada palavra reconhecida com bons resultados apresentados pelo classificador se tornam dados de treinamento utilizados no módulo de Reconhecimento de Palavra 2. O último módulo Ajuste trata de corrigir espaçamentos e o tamanho do caractere para identificar quando a letra é maiúscula ou minúscula em casos de fontes que não diferencie uma fonte da outra (SMITH, 2013).

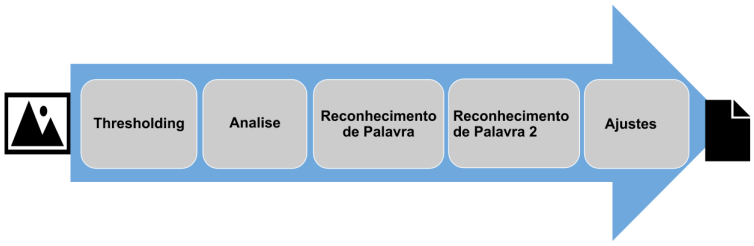


**Figura 3 – Fluxo do processamento do Tesseract.**

**Fonte: (SMITH, 2013)**

## 2.3 Processamento de Linguagem Natural

Há uma coleção de tarefas fundamentais que aparecem frequentemente em vários projetos de PLN. Devido à sua natureza repetitiva e fundamental, essas tarefas têm sido estudadas extensivamente. As tarefas mais conhecidas dentro de PLN são destacadas a seguir(THOMAS, 2020; MOLDEN, 2009):

- Modelagem de idiomas: Esta é a tarefa de prever qual será a próxima palavra em uma frase baseada no histórico de palavras anteriores. O objetivo desta tarefa é aprender a probabilidade de uma sequência de palavras aparecer em uma determinada língua. A modelagem de linguagem é útil para a construção de soluções para uma grande variedade de problemas, como reconhecimento de fala, reconhecimento óptico de caracteres, reconhecimento de caligrafia, tradução automática e correção ortográfica.

- Tradução automática: Esta é a tarefa de converter um texto de um idioma para outro. Ferramentas como o Google Translate são aplicações comuns dessa tarefa.
- Classificação de texto: Esta é a tarefa de colocar o texto em um conjunto conhecido de categorias com base em seu conteúdo. A classificação de texto é de longe a tarefa mais popular no PLN, sendo usada para solucionar uma ampla variedade de problemas, desde a identificação de spam de e-mail até a análise de sentimentos.
- Extração de informações: Como o nome indica, esta é a tarefa de extrair informações relevantes do texto, como eventos de calendário de e-mails ou nomes de pessoas mencionadas em uma postagem nas redes sociais.
- Recuperação de informações: Esta é a tarefa de encontrar documentos relevantes para uma consulta de usuário de uma grande coleção. Aplicativos como o Google Search são casos de uso bem conhecidos de recuperação de informações.
- Agente conversador: Esta é a tarefa de construir sistemas de diálogo que possam conversar em línguas humanas. Alexa e Siri são algumas aplicações comuns desta tarefa.
- Resposta de perguntas: Esta é a tarefa de construir um sistema que possa responder automaticamente às perguntas colocadas na linguagem natural.
- Resumo de texto: Esta tarefa visa criar resumos curtos de documentos mais longos, mantendo o conteúdo principal e preservando o significado geral do texto.
- Modelagem de tópicos: Esta é a tarefa de descobrir a estrutura de tópicos de uma grande coleção de documentos. A modelagem de tópicos é uma ferramenta comum de mineração de texto, a qual é amplamente usada em vários domínios, da literatura à bioinformática.

Aplicativos relacionados ao PLN são construídos usando uma enorme quantidade de dados. Em termos leigos, pode se dizer que uma grande coleção de dados é chamada de *corpus*. *Corpus* é uma coleção de material de língua natural escrito ou falado, armazenado no computador, e usado para descobrir como a linguagem é usada (HUTCHINSON, 2015). Assim, mais precisamente, um *corpus* é uma coleção informatizada sistemática de linguagem autêntica que é usada para análise linguística, bem como análise de *corpus*. Com a ajuda de um *corpus*, é possível realizar algumas análises estatísticas, como distribuição de frequências e co-ocorrências de palavras. O uso de *corpus* pode ajudar a definir e implementar regras linguísticas para várias aplicações de PLN, como exemplo um sistema de correção gramatical, onde se usará o *corpus* de texto e tentará descobrir as instâncias gramaticalmente incorretas, e então serão definidas as regras gramaticais que ajudaram a corrigir essas instâncias (MAJUMDER ANUJ GUPTA, 2020).

### 2.3.1 Aprendizado de Máquina para PLN

As técnicas de aprendizado de máquina são aplicadas aos dados textuais, assim como são usados em outras formas de dados, como imagens, fala e dados estruturados. Técnicas supervisionadas de aprendizagem de máquina, como métodos de classificação e regressão, são fortemente utilizadas para processar tarefas de PLN. Como exemplo, uma tarefa de classificação de PLN seria classificar artigos de notícias em um conjunto de tópicos de notícias, tais como esportes ou política. Por outro lado, as técnicas de regressão, que dão uma previsão numérica, podem ser usadas para estimar o preço de uma ação com base no processamento da discussão nas redes sociais sobre esse estoque. Da mesma forma, algoritmos de agrupamento (*clustering*) não supervisionados podem ser usados para juntar documentos de texto. Qualquer abordagem de aprendizado de máquina para PLN, supervisionada ou não supervisionada, pode ser descrita como sendo composta por três etapas comuns: extrair recursos, usar a representação do recurso para aprender e gerar um modelo e por fim avaliar e melhorar o modelo (JOSHI, 2017).

### 2.3.2 Técnicas de Pré-Processamento

Antes de usar o texto, para alguma atividade de classificação ou regressão em AM, é interessante realizar um processo de "limpeza"no texto. Existe diversas técnicas para isso, algumas delas são a segmentação de frases e a segmentação de palavras. O software PLN normalmente analisa o texto dividindo-o em palavras (*tokens*) e frases. Na Figura 4, possui uma pipeline de PLN com todos os possíveis pré-processamentos que serão discutidos nos próximos parágrafos.

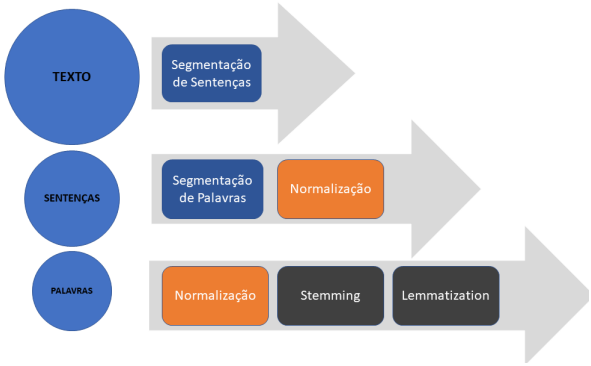


**Figura 4 – Fluxo de pré-processamento em PLN.**

**Fonte: Adaptado de (MAJUMDER ANUJ GUPTA, 2020)**

Qualquer pipeline PLN tem que começar com um sistema confiável para dividir o texto em frases ou sentenças (segmentação de sentenças) e dividir ainda mais a frase em palavras (segmentação de palavras). Em nível de sentença, algumas normalizações podem ser feitas como exemplo a exclusão de pontuação. Stemização refere-se ao processo de remoção de sufixos e redução de uma palavra para alguma forma de base de modo que todas as diferentes variantes

dessa palavra podem ser representadas pela mesma forma. Por exemplo, "carro"e "carros"são ambos reduzidos a "carro". Isso é feito aplicando um conjunto fixo de regras. Embora essas regras nem sempre possam acabar em uma forma base linguisticamente correta, a stemização é comumente usados em mecanismos de busca para combinar consultas de usuários a documentos relevantes e na classificação de texto para reduzir o espaço de recurso para treinar modelos de aprendizado de máquina.(JOSHI, 2017). Lematização é o processo de mapear todas as diferentes formas de uma palavra à sua palavra base. Embora isso pareça próximo à definição de stemização, eles são, de fato, diferentes. A lematização é o processo de reflexionar uma palavra para determinar o seu lema, sendo que as flexões se chamam lexemas. A lematização é útil para analisar os usos de palavras em contextos sem importância das flexões. A lematização requer mais conhecimento linguístico e modelar e desenvolver lematizadores eficientes continua sendo um problema aberto na pesquisa de PLN até agora (HUTCHINSON, 2015).

O texto das redes sociais é muito diferente da linguagem que são usadas nos jornais. Uma palavra pode ser escrita de diferentes maneiras, inclusive em formas encurtadas, um número de telefone pode ser escrito em diferentes formatos (por exemplo, com e sem hifens), nomes às vezes estão em minúsculas, e assim por diante. Quando se trabalha no desenvolvimento de ferramentas PLN para trabalhar com esses dados, é útil alcançar uma representação canônica do texto que captura todas essas variações em uma representação. Isso é conhecido como normalização de texto. Alguns passos comuns para a normalização do texto são converter todo o texto em minúsculas ou minúsculas, converter dígitos em texto (por exemplo, 9 a nove), expandir abreviaturas e assim por diante. Uma maneira simples de incorporar a normalização do texto pode ser encontrada no código-fonte de um dicionário que mostra diferentes ortografias de uma coleção predefinida de palavras mapeadas para uma única ortografia. (MAJUMDER ANUJ GUPTA, 2020)

### 2.3.3 Abordagens básicas de vetorização

O processo de vetorização consiste em mapear cada palavra no vocabulário (N) do corpo de um texto para um ID único (valor inteiro) e, em seguida, representar cada frase ou documento no *corpus* como um vetor N-dimensional.

Por exemplo, reduzindo o texto da lista abaixo e ignorando a pontuação, o vocabulário deste *corpus* é composto por seis palavras: [gato, morde, homem, come, peixe, comida]. Pode-se organizar o vocabulário em qualquer ordem. Neste exemplo, simplesmente é tomado a ordem em que as palavras aparecem no *corpus*. Cada documento neste *corpus* pode agora ser representado com um vetor de tamanho seis.

1. O gato morde o homem.
2. O homem morde o gato.

3. O gato come peixe.

4. O homem come comida.

#### 2.3.3.1 Codificação one-hot

Em uma codificação *one-hot*, cada palavra X no vocabulário do *corpus* recebe um ID inteiro único que está entre 1 e N, onde N é o conjunto do vocabulário do corpus. Cada palavra é então representada por um vetor binário N-dimensional de 0s e 1s. Isso é feito por meio de um vetor de dimensão preenchido com todos os 0s barrando o índice, onde o valor do índice é igual a idX. A representação de palavras individuais é então combinada para formar uma representação de sentença. Pelo lado positivo, a codificação *one-hot* é intuitiva de entender e é simples de implementar. No entanto, a codificação *one-hot* sofre de algumas deficiências. Em alguns casos os vetores podem ser da casa de milhões se tornando portanto grandes e esparsos, desde que cada vetor possui apenas um valor "1" e vários "0"s. Outro fator negativo ao empregar a codificação do *one-hot* está relacionado à similaridade entre as palavras. Nessa representação, a distância entre quaisquer duas palavras é a mesma, pois cada palavra é um vetor perpendicular a todos os outros, logo, o produto interno entre dois vetores é igual a zero. Um exemplo é mostrado na Tabela 1 utilizando as 4 frases da subseção 2.3.3.

**Tabela 1 – Tabela de exemplo do One-Hot**

| **palavra** | **ID** | **vetor** | | | | | |
|---|---|---|---|---|---|---|---|
| gato | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| morde | 2 | 0 | 1 | 0 | 0 | 0 | 0 |
| homem | 3 | 0 | 0 | 1 | 0 | 0 | 0 |
| come | 4 | 0 | 0 | 0 | 1 | 0 | 0 |
| peixe | 5 | 0 | 0 | 0 | 0 | 1 | 0 |
| comida | 6 | 0 | 0 | 0 | 0 | 0 | 1 |

#### 2.3.3.2 Saco de Palavras

Sacos de Palavras ou no inglês *Bag of Words*(BoW) é uma técnica clássica de representação de texto que tem sido usada comumente em PLN, especialmente em problemas de classificação de texto. A ideia-chave é a seguinte: representar o texto em consideração como um saco (coleção) de palavras, ignorando a ordem e o contexto. A intuição básica por trás disso é que ele assume que o texto pertencente a uma determinada classe no conjunto de dados é caracterizado por um conjunto único de palavras. Se duas peças de texto têm quase as mesmas palavras, então elas pertencem à mesma classe. Assim, analisando as palavras presentes em um texto, pode-se identificar a classe a que pertence, semelhante à codificação de *one-hot*. Cada documento no *corpus* é então convertido em um vetor de N dimensões, onde no 1º componente do vetor, i possui o valor NID (i = NID), é simplesmente o número de vezes que a palavra N

ocorre no documento, ou seja, simplesmente pontuamos cada palavra em N pela sua contagem de ocorrências no documento.

Na Tabela 2 temos a representação de um saco de palavras das 4 frases usadas como exemplo na subseção 2.3.3. Nota-se que a representação do BoW que as frases não possuem palavras repetidas por isso cada um possui o número 1. Porém algumas palavras como "gato"e "homem"possuem uma frequência maior de ocorrência dentro do *corpus* (conjunto das 4 frases).

**Tabela 2 – Tabela de exemplo do BoW**

| ID | gato | morde | homem | come | peixe | comida |
|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 0 | 0 | 0 |
| 2 | 1 | 1 | 1 | 0 | 0 | 0 |
| 3 | 1 | 0 | 0 | 1 | 1 | 0 |
| 4 | 0 | 0 | 1 | 1 | 0 | 1 |

Pesquisadores têm mostrado que tal representação sem considerar a frequência é útil para a análise de sentimentos, outros pontos são destacados a seguir:

- BoW é bastante simples de entender e implementar.
- Com essa representação, documentos com as mesmas palavras terão suas representações vetoriais mais próximas umas das outras no espaço euclidiano em comparação com documentos com palavras completamente diferentes. Então, se dois documentos têm vocabulário semelhante, eles estarão mais próximos um do outro no espaço vetorial e vice-versa.
- Codificação fixa para qualquer sentença de comprimento arbitrário.

#### 2.3.3.3 Saco de N-Grams

Todos os esquemas de representação que vimos anteriormente tratam as palavras como unidades independentes. Não há noção de frases ou pedidos de palavras. A abordagem saco-de-n-gramas (BoN) tenta remediar isso. Ele faz isso quebrando texto em pedaços de n palavras contíguas (ou tokens). Isso pode nos ajudar a capturar algum contexto, o que abordagens anteriores não conseguem fazer. Cada pedaço é chamado de n-grama. O vocabulário V do *corpus* corresponde a uma coleção de todos os n-gramas únicos que ocorrem no corpo do texto. Em seguida, cada documento no *corpus* é representado por um vetor de comprimento N. Este vetor contém as contagens de frequência de n-gramas presentes no documento e zero para os n-gramas que não estão presentes.

Na Tabela 3 é mostrado o agrupamento das palavras caso fosse utilizado um Saco de N-grams. Nessa abordagem, pode-se analisar a frequência de vezes que alguma dupla de palavra se repete. Na Tabela 4 o processo de Trigram é realizado.

**Tabela 3 – Tabela de exemplo de BiGrams**

| ID | BiGram | | | |
|---|---|---|---|---|
| 1 | O gato | gato morde | morde o | o homem |
| 2 | O homem | homem morde | morde o | o gato |
| 3 | O gato | gato come | come peixe | |
| 4 | O homem | homem come | come comida | |

**Tabela 4 – Tabela de exemplo de TriGrams**

| ID | TriGram | | |
|---|---|---|---|
| 1 | O gato morde | gato morde o | morde o homem |
| 2 | O homem morde | homem morde o | morde o gato |
| 3 | O gato come | gato come peixe | |
| 4 | O homem come | homem come comida | |

#### 2.3.3.4 TF-IDF

Nas abordagens anteriores todas as palavras no texto são tratadas como igualmente importantes, não há noção de que algumas palavras no documento sejam mais importantes do que outras. A frequência de termo inversa ou TF-IDF, aborda esse problema. Tem como objetivo quantificar a importância de uma determinada palavra em relação às outras palavras no documento e no *corpus*. É um esquema de representação comumente usados para sistemas de recuperação de informações, para extrair documentos relevantes de um *corpus* para uma determinada consulta de texto.

A intuição por trás do TF-IDF é a seguinte: se uma palavra N aparece muitas vezes em um documento Doc[i], mas não ocorre muitas vezes no resto dos documentos Doc[j] no *corpus*, então a palavra N deve ser de grande importância para o documento Doc[i]. A importância do N deve aumentar proporcionalmente à sua frequência em Doc[i], mas ao mesmo tempo, sua importância deve diminuir em proporção à frequência da palavra em outros documentos Doc[j] no *corpus*. Matematicamente isso é capturado usando duas medidas: TF e IDF. Os dois são então combinados para chegar à pontuação TF-IDF.

TF (frequência de termo) mede com que frequência um termo ou palavra ocorre em um determinado documento. Uma vez que diferentes documentos no *corpus* podem ser de diferentes comprimentos, um termo pode ocorrer mais frequentemente em um documento mais longo em comparação com um documento mais curto. Para normalizar essas contagens, dividimos o número de ocorrências pelo comprimento do documento.

O IDF (frequência de documentos inversos) mede a importância do termo em um corpus. Na computação TF, todos os termos possuem igual importância (ponderação). No entanto, é um fato bem conhecido que certas palavras como "é", "o", "de", etc., não são importantes, mesmo que ocorram com frequência. Para explicar esses casos, o IDF pondera os termos que são muito

comuns em um *corpus* e pesa os termos raros.

A Tabela 5 exemplifica o uso de TD-IDF. As palavras "peixe"e "comida"são pouco usadas no *corpus*, ao contrário de "gato"e "homem". Por isso, "gato"e "homem possuem pontuação mais baixa que as anteriores.

**Tabela 5 – Tabela de exemplo de TD-IDF**

| ID | gato | morde | homem | come | peixe | comida |
|---|---|---|---|---|---|---|
| 1 | 0.14 | 0.05 | 0.1 | 0 | 0 | 0 |
| 2 | 0.21 | 0.12 | 0.01 | 0 | 0 | 0 |
| 3 | 0.1 | 0 | 0 | 0.3 | 0.75 | 0 |
| 4 | 0 | 0 | 0.2 | 0.1 | 0 | 0.8 |

## 2.4 Aprendizado de Máquina

O conhecimento que é adquirido com o uso de AM permite que computadores generalizem corretamente novos eventos e configurações. O uso deste campo de estudo existe há décadas, inclusive aplicado para algumas aplicações especializadas, como o OCR. Entretanto o primeiro aplicativo AM que realmente se tornou *mainstream*, melhorando a vida de centenas de milhões de pessoas, foi o filtro de *spam*, que tecnicamente se qualifica como Aprendizado de Máquina (ele realmente aprendeu tão bem que você raramente precisa sinalizar um e-mail como *spam* de forma manual) (NEWNHAM, 2018). Depois disso, centenas de aplicativos AM surgiram, os quais alimentam silenciosamente centenas de produtos e recursos que se usa regularmente, desde sistemas de recomendação até pesquisa por voz.

Outros exemplos de Aplicações incluem:

- Análise de imagens de produtos em uma linha de produção para classificá-los automaticamente;
- Detecção de tumores em imagens de tomografias cerebrais;
- Classificação automática de artigos de notícias;
- Sinalização automática de comentários ofensivos em fóruns de discussão;
- Resumo de documentos longos;
- Criação de um *chatbot* e de assistente pessoal;
- Construção de um robô inteligente para um jogo;
- Previsão de receita de uma empresa, com base em métricas de desempenho;

- Uso de comandos de voz para acionar funcionalidades de um aplicativo;
- Detecção de fraudes de cartão de crédito;
- Segmentação de clientes com base em suas compras com o intuito de se projetar uma estratégia de marketing diferente para cada segmento;
- Representação de conjuntos de dados complexos e de alta dimensão em um diagrama claro e perspicaz;
- Recomendação de produtos que um cliente possa estar interessado, com base em compras passadas;

O objetivo dessa sessão é descrever brevemente os principais métodos de aprendizado de máquina. com enfoque na aprendizagem supervisionada que será utilizada como método de aprendizado nesse trabalho. Existem basicamente 4 modos de aprendizado, a saber: supervisionado, semi-supervisionado, não supervisionado e reforço (ESPOSITO, 2020).

No aprendizado não supervisionado, os dados de treinamento não possuem uma classe alvo pré-determinada (chamado de rótulos em aprendizado supervisionado) e o sistema tenta aprender sozinho, por exemplo para gerar agrupamentos de dados conforme a similaridade entre os dados, de forma que os dados sejam bem similares dentro do grupo e bem dissimilares com dados de outros grupos. Alguns dos algoritmos que auxiliam neste método são agrupamento ou *Clustering* (K-Means, DBSCAN e Análise de *cluster* hierárquico), detecção de anomalias e detecção de novidades (One-class SVM e Floresta de Isolamento), visualização e redução de dimensionalidade (Análise de componentes principais (ACP), *Kernel* ACP, *Locally Linear Embedding* e *t-Distributed Stochastic Neighbor Embedding* (t-SNE)) e adoção de regras de associação (Apriori e Eclat). Um exemplo clássico do método não supervisionado é um grupo de dados sobre visitantes de um *blog*. Por meio de um algoritmo de agrupamento tenta-se detectar grupos de visitantes semelhantes. Em nenhum momento o usuário interage com o algoritmo e indica a qual grupo um visitante pertence. O algoritmo encontra essas conexões sem este tipo de ajuda. Pode ser que 40% dos visitantes são homens que amam histórias sobre a segunda guerra mundial e geralmente leem este *blog* à noite, enquanto 20% são jovens amantes do seriado Jornada nas Estrelas e que frequentemente visitam o *blog* durante os fins de semana. Se usar um algoritmo hierárquico de agrupamento, este algoritmo também pode subdividir cada grupo em grupos menores, isso pode ajudar a direcionar as postagens para cada grupo especifico de acordo com os gostos da maioria (FENNER, 2019; NEWNHAM, 2018).

No aprendizado supervisionado, existe um conjunto de dados de treinamento que alimenta o algoritmo que, por meio de dados já classificados ou também chamados de rotulados, pode criar um modelo de classificação que será depois usado para classificar novos conjuntos de dados. Uma tarefa típica de aprendizagem supervisionada é a classificação. Na classificação, o objetivo é prever um rótulo de classe, que é uma escolha a partir de uma lista predefinida de

possibilidades. A classificação às vezes é separada em classificação binária, que é o caso especial de distinguir entre exatamente duas classes, e classificação multi-classe, que é a classificação entre mais de duas classes (NEWNHAM, 2018). Outra tarefa típica é prever um valor numérico de destino, como exemplo o preço de um carro, dado um conjunto de recursos (quilometragem, idade, marca, etc.) chamados preditores. Esse tipo de tarefa é chamado de regressão. Para treinar o sistema, você precisa dar-lhe muitos exemplos de carros, incluindo tanto seus preditores quanto suas etiquetas (ou seja, seus preços). Para tarefas de regressão, o objetivo é prever um número contínuo (ou um número de ponto flutuante em termos de programação ou ainda um número real em termos matemáticos). Prever a renda anual de uma pessoa a partir de sua educação, sua idade e onde ela vive é um exemplo de uma tarefa de regressão. Ao prever a renda, o valor previsto é um valor numérico real e pode ser qualquer número em uma determinada faixa. Outro exemplo de uma tarefa de regressão é prever a produtividade de uma fazenda de milho, dado atributos como produtividades anteriores, clima e número de funcionários trabalhando na fazenda. A produtividade novamente pode ser um número arbitrário. Uma maneira fácil de distinguir entre tarefas de classificação e regressão é perguntar se há algum tipo de continuidade na saída. Se houver continuidade entre os possíveis desfechos, então o problema é um problema de regressão (FENNER, 2019; NEWNHAM, 2018).

Uma vez que rotular dados geralmente é demorado e custoso, muitas vezes haverá muitos dados sem rótulo, e poucas instâncias rotuladas. Alguns algoritmos podem lidar com dados que são parcialmente rotulados. Isso é chamado de aprendizado semi-supervisionado. A maioria dos algoritmos de aprendizagem semi-supervisionados são combinações de algoritmos não supervisionados e supervisionados. Por exemplo, redes de crenças profundas, ou em inglês *Deep Belief Network* (DBNs) são baseadas em componentes não supervisionados chamados Máquinas Boltzmann Restritas (MBR) empilhadas umas sobre as outras. Os MBRs são treinados sequencialmente de forma não supervisionada, e então todo o sistema é afinado usando técnicas de aprendizagem supervisionada (NEWNHAM, 2018).

O Aprendizado de Reforço é um o método mais diferente do grupo. O sistema de aprendizagem, chamado de agente nesse contexto, pode observar o ambiente, selecionar e executar ações e receber recompensas em troca (ou penalidades sob a forma de recompensas negativas). Ele deve então aprender por si mesmo qual é a melhor estratégia (chamada de política) para obter a maior recompensa ao longo do tempo. Uma política define qual ação o agente deve escolher quando está em uma determinada situação (FENNER, 2019).

### 2.4.1 Algorítimos

Nesta subseção são destacados os principais algoritmos de aprendizado de máquina utilizados em tarefas de classificação em modelos de aprendizado supervisionado.

#### 2.4.1.1 Máquina de Vetores de Suporte

Uma Máquina vetorial de suporte, ou comumente chamada pelo seu acrônimo em inglês *Support Vector Machine* (SVM) é um modelo poderoso e versátil de Aprendizado de Máquina, capaz de realizar classificação linear e não linear, regressão e até mesmo detecção de *outliers*. É um dos modelos mais populares em Aprendizado de Máquina. Embora os classificadores lineares SVM sejam eficientes e funcionem surpreendentemente bem em muitos casos, muitos conjuntos de dados não estão nem perto de serem linearmente separáveis. Uma abordagem para lidar com conjuntos de dados não lineares é adicionar mais recursos, como recursos polinomiais. O algoritmo SVM é versátil: permite ambas classificação linear e não linear, como também provê suporte para regressão linear e não linear (LEE, 2019).

#### 2.4.1.2 Árvores de Decisão

As Árvores de Decisão são algoritmos versáteis de Aprendizado de Máquina que podem realizar tarefas de classificação e regressão, e até mesmo tarefas de multi-produção. São algoritmos poderosos, capazes de encaixar conjuntos de dados complexos. São também componentes fundamentais das Florestas Aleatórias.

Um exemplo utilizado para explicar o seu funcionamento é a classificação de uma flor. Começa-se no nó raiz. Este nó pergunta se o comprimento da pétala da flor é menor que 2,45 cm. Se for, então move-se para baixo para o nó filho esquerdo da raiz. Neste caso, é um nó de folha, por isso não faz perguntas: basta olhar para a classe prevista para esse nó, e a Árvore de Decisão prevê qual a flor em questão. Agora suponha que encontre outra flor, e desta vez o comprimento da pétala é maior que 2,45 cm. Deve mover-se para baixo para o nó filho direito da raiz, que não é um nó de folha, então o nó faz outra pergunta: a largura pétala é menor que 1,75 cm? Se for, então a flor é provavelmente de uma outra classificação. De modo geral, esse algoritmo trabalha exatamente como uma arvore, onde no topo existe o ponto de partida e cada galho possui uma, duas ou mais possibilidades que correspondem com o dado analisado (RASCHKA, 2019).

#### 2.4.1.3 *Extremely Randomized Trees*

Este algoritmo é formado por um grupo de algoritmos de Floresta Aleatória, em cada nó apenas um subconjunto aleatório das características é considerado para a divisão. É possível tornar as árvores ainda mais aleatórias usando também limiares aleatórios para cada recurso em vez de procurar os melhores limiares possíveis (como árvores de decisão regulares fazem). Este algoritmo também torna as Extra-Árvores muito mais rápidas de treinar do que florestas aleatórias normais, porque encontrar o melhor limiar possível para cada recurso em cada nó é uma das tarefas mais demoradas de cultivar uma árvore.(LEE, 2019)

#### 2.4.1.4 Florestas Aleatórias

Os métodos *Ensemble* se trata de uma agregação das previsões de um grupo de preditores (classificadores ou regressores), muitas vezes a combinação de todas essas previsões terão resultados melhores do que individuais. No caso esse grupo se chama *Ensemble* e assim esta técnica é chamada de método *Ensemble*. Um exemplo de método *Ensemble* consiste em treinar um grupo de classificadores da Árvore de Decisão, cada um usando um subconjunto aleatório diferente do conjunto de treinamento. Para fazer previsões, você obtém as previsões de todas as árvores individuais, depois prevê a classe que obtém mais votos. Tal conjunto de Árvores de Decisão é chamado de Floresta Aleatória, e apesar de sua simplicidade, este é um dos algoritmos de Aprendizado de Máquina mais poderosos disponíveis atualmente (LEE, 2019).

#### 2.4.1.5 *AdaBoost*

Uma maneira de um novo preditor corrigir seu antecessor é prestar um pouco mais de atenção às instâncias de treinamento que o antecessor usou. Isso resulta em novos preditores focando cada vez mais nos casos difíceis. Esta é a técnica usada pela *AdaBoost*. Por exemplo, ao treinar um classificador *AdaBoost*, o algoritmo primeiro treina um classificador base (como uma Árvore de Decisão) e o usa para fazer previsões sobre o conjunto de treinamento. O algoritmo então aumenta o peso relativo de instâncias de treinamento mal classificadas. Em seguida, ele treina um segundo classificador, usando os pesos atualizados, e novamente faz previsões sobre o conjunto de treinamento, atualiza os pesos da instância, e assim por diante.(RASCHKA, 2019)

#### 2.4.1.6 *Gradient Boosting*

Outro algoritmo de impulsionamento muito popular é o *Gradient Boosting*. Assim como *AdaBoost*, *Gradient Boosting* funciona adicionando sequencialmente preditores a um conjunto, cada um corrigindo seu antecessor. No entanto, em vez de ajustar os pesos de instância em cada iteração como o *AdaBoost* faz, este método tenta encaixar o novo preditor aos erros residuais cometidos pelo preditor anterior (RASCHKA, 2019).

### 2.4.2 Treinamento e Teste

A única maneira de saber o quão preciso um modelo de classificação irá generalizar para novos casos é realmente experimentá-lo em novos casos. Uma maneira de fazer isso é colocar o modelo em produção e monitorar o desempenho dele. Uma opção melhor é dividir os dados em dois conjuntos: o conjunto de treinamento e o conjunto de testes. Como os próprios nomes implicam, o primeiro conjunto treina o modelo, e o segundo testa este modelo. A taxa de erro em novos casos é chamada de erro de generalização (ou erro fora da amostra) e ao avaliar o modelo no conjunto de testes é possível estimar a taxa desse erro. Este valor diz o quão preciso o modelo irá funcionar em instâncias que nunca foram treinadas anteriormente. Se o erro de treinamento for baixo (ou seja, o modelo produz poucos erros no conjunto de treinamento), mas o erro de

generalização é alto, isso significa que o modelo está sobreajustado para os dados de treinamento (AGRAWAL, 2020). Na etapa de treinamento é comum particionar os dados em subconjuntos e usar a técnica de *k-fold cross validation*, a qual consiste em dividir o conjunto total de dados em k subconjuntos mutuamente exclusivos do mesmo tamanho e, a partir daí, um subconjunto é utilizado para teste e os k-1 restantes são utilizados para estimação dos parâmetros, fazendo-se o cálculo da acurácia do modelo.

### 2.4.3 Medidas de desempenho

Existem muitas medidas de desempenho disponíveis nesta subseção serão descritas as principais medidas.

#### 2.4.3.1 Matriz de Confusão

No campo da análise de modelos de aprendizado de máquina, uma matriz de confusão é uma tabela que permite a visualização do desempenho de um algoritmo, cujo objetivo é prever as classes de uma variável. O desempenho descrito é qualitativo. O nome "confusão" vem do fato de tornar mais fácil ver se o sistema está classificando incorretamente uma classe como outra. A matriz de confusão na Tabela 6 mostra duas linhas e duas colunas que relatam o número de falsos positivos, falsos negativos, verdadeiros positivos e verdadeiros negativos.

O verdadeiro positivo é o número de dados previstos corretamente como um membro da classe e o verdadeiro negativo, o número de dados rejeitados corretamente como membro da classe. Já o falso positivo é o número de dados identificados incorretamente como um membro da classe e o falso negativo é o número de dados rejeitados incorretamente como membro da classe.

Isso permite uma análise mais detalhada do que, por exemplo, a mera proporção de suposições corretas (precisão). A matriz de confusão é uma métrica mais confiável para o desempenho real de um classificador porque não produzirá resultados enganosos se o conjunto de dados for desequilibrado (ou seja, quando o número de amostras em classes diferentes varia muito). Esta é a principal vantagem de representar os resultados usando esta estrutura: evitamos estatísticas que retornam uma alta taxa de sucesso enganosa.

**Tabela 6 – Tabela de Matriz de Confusão (Binária)**

| | | **Predição** | |
|---|---|---|---|
| | | CLASSE A | CLASSE B |
| **Gabarito** | CLASSE A | Verdadeiro Positivo | Falso Negativo |
| | CLASSE B | Falso Positivo | Verdadeiro Negativo |

#### 2.4.3.2 Acurácia

A acurácia é o número de previsões corretas sobre o tamanho da saída. É uma medição incrivelmente direta e, graças à sua simplicidade, é amplamente útil. A acurácia é uma das primeiras métricas que se calcula ao avaliar os resultados. A fórmula para se calcular acurácia segue abaixo:

$$Acuracia = \frac{VerdadeiroPositivo + VerdadeiroNegativo}{VerdadeiroPositivo + VerdadeiroNegativo + FalsoPositivo + FalsoNegativo}$$

#### 2.4.3.3 Precisão

A precisão é uma métrica similar à Acurácia, mas mede apenas a taxa de verdadeiros positivos com relação ao que foi recuperado corretamente. Em certos domínios, como a detecção de *spam*, um falso positivo é um erro pior do que um falso negativo (geralmente, perder um *e-mail* importante é pior do que a inconveniência de excluir um pedaço de *spam* que escapou do filtro).

$$Precisao = \frac{VerdadeiroPositivo}{VerdadeiroPositivo + FalsoPositivo}$$

#### 2.4.3.4 Revocação

Revocação é o oposto de precisão, ele mede falsos negativos contra verdadeiros positivos. Os falsos negativos são especialmente importantes para prevenção na detecção de doenças e outras previsões envolvendo segurança.

$$Revocacao = \frac{VerdadeiroPositivo}{VerdadeiroPositivo + FalsoNegativo}$$

#### 2.4.3.5 F1

Calcula-se a pontuação F1 como a média harmônica de precisão e revocação para conseguir exatamente isso (harmonia). Embora possa ser calculado de forma simples por meio de média das duas pontuações, as médias harmônicas são mais resistentes a *outliers*. Assim, a pontuação F1 é uma métrica balanceada que quantifica apropriadamente a exatidão dos modelos em muitos domínios.

$$F1 = 2 * \frac{Revocacao * Precisao}{Revocacao + Precisao}$$

#### 2.4.3.6 Curva ROC e AUC

A curva característica de operação do receptor, ou do inglês *Receiver Operating Characteristic* (ROC) é outra ferramenta comum usada com classificadores binários. É muito semelhante à curva de precisão/revocação, mas em vez de traçar precisão versus revocação, a curva ROC

traça a taxa real positiva (outro nome para revocação) contra a taxa falsa positiva (TFP). O TFP é a razão de instâncias negativas que são incorretamente classificadas como positivas. Uma maneira de comparar classificadores é medir a área sob a curva, ou do inglês *Area Under Curve* (AUC). Um classificador perfeito terá um ROC AUC igual a 1, enquanto um classificador puramente aleatório terá um ROC AUC igual a 0,5.

### 2.4.4 Generalização, Sobreajuste, Sub-ajuste e Validação Cruzada

No aprendizado supervisionado é construído um modelo de classificação sobre os dados de treinamento e, em seguida, o modelo é usado para fazer previsões sobre novos dados que ainda não foram rotulados, mas possuem as mesmas características do conjunto de treinamento que foi usado nos testes. Se um modelo é capaz de fazer previsões precisas sobre dados que ainda não foram rotulados, dizemos que é capaz de generalizar desde o conjunto de treinamento até o conjunto de testes. Sendo assim o objetivo final é construir um modelo capaz de generalizar de forma mais precisa possível (AGRAWAL, 2020).

Normalmente é construído um modelo que possa fazer previsões precisas sobre o conjunto de treinamento. Se os conjuntos de treinamento e teste tiverem bastante em comum, esperamos que o modelo também seja preciso no conjunto de testes. No entanto, há alguns casos em que isso pode dar errado. Por exemplo, modelos muito complexos, podem sempre ser tão precisos quanto no conjunto de treinamento. Construir um modelo muito complexo para a quantidade de dados é o que chamamos de sobreajuste. O excesso de adequação ocorre quando se encaixa um modelo muito próximo às particularidades do conjunto de treinamento e obtém um modelo que funciona bem no conjunto de treinamento, mas não é capaz de generalizar para novos dados. Por outro lado, se o modelo for muito simples, então pode não ser capaz de capturar todos os aspectos e a variabilidade nos dados, e o modelo vai se ter baixo desempenho para o conjunto de treinamento. Escolher um modelo muito simples é chamado de sub-ajustado.

Generalizar demais é algo que humanos fazem com muita frequência, e infelizmente as máquinas podem cair na mesma armadilha se não houver cuidado na implementação delas. Ocorre quando modelo é muito simples para aprender a estrutura subjacente dos dados. Além disso, se o conjunto de validação for muito pequeno, então as avaliações do modelo serão imprecisas: podendo então o modelo acabar selecionando um modelo sub-ótimo por engano. Por outro lado, se o conjunto de validação for muito grande, então o conjunto de treinamento restante será muito menor do que o conjunto completo de treinamento. como o modelo final será treinado no conjunto completo de treinamento, não é ideal comparar modelos candidatos treinados em um conjunto de treinamento muito menor. Uma maneira de resolver esse problema é realizar repetida validação cruzada, usando muitos pequenos conjuntos de validação. Cada modelo é avaliado uma vez por conjunto de validação após ser treinado no restante dos dados. Ao fazer uma média de todas as avaliações de um modelo, recebe-se uma medida muito mais precisa de seu desempenho. Há uma desvantagem, no entanto: o tempo de treinamento é multiplicado pelo

número de conjuntos de validação (AGRAWAL, 2020; RASCHKA, 2019).

## 2.5 Informações Falsas

Existem diversas correntes e pensamentos na literatura de como classificar as informações falsas, alguns dos mais citados são: Notícias falsas, Rumores, Farsas, *Clickbait*, Sátira e Propaganda.

Alguns pesquisadores definem notícias falsas (Fake News) como artigos de notícias que são potencialmente ou intencionalmente desinformativos para os leitores, pois os fatos são verificáveis e deliberadamente falsos. Além disso, *Fake News* consistem em informações não verídicas transmitidas por meio de mensagem, áudio, imagem ou vídeos editados para atrair a atenção do leitor no intuito de desinformá-lo e obter tipo algum de vantagem sobre ele, sem que haja fonte verídica determinada, mas apresentando uma maquiagem que transparece uma aparente ingenuidade para quem recebe como recebe.

Da mesma forma, para a desinformação, boatos são concebidas para enganar os leitores; qualitativamente, eles são descritos como humorístico ou malicioso. Duas grandes teorias psicológicas explicam essa dificuldade, respectivamente chamada de realismo ingênuo e viés de confirmação. O primeiro refere-se à tendência dos usuários de acreditar que sua visão é a única precisa, enquanto aqueles que discordam são tendenciosos ou desinformados. Esta última, também chamada de exposição seletiva, é a inclinação para preferir (e receber) informações que confirma visões já existentes. Todos esses fatores estão relacionados, até certo ponto, ao conhecido efeito de câmara de eco (ou bolha de filtro), que dá origem à formação de aglomerados homogêneos onde os indivíduos são semelhantes, e que compartilham e discutem ideias semelhantes.(PIERRI; CERI, 2019)

Existem duas terminologias americanas que possuem conceitos semelhantes são eles *Hoax* (Farsas) e Rumores. Enquanto o primeiro trata de informações que contêm fatos que são falsos ou imprecisos, entretanto quando apresentados textualmente são tratados como fatos legítimos. Esta categoria também é conhecida na comunidade de pesquisa tanto como histórias com meio verdade ou factoides. Os rumores se referem a histórias cuja veracidade é ambígua ou nunca confirmada. Este tipo de informação falsa é amplamente propagada em redes sociais (ZANNETTOU et al., 2019).

Propagandas são informações ou ideias difundidas com o objetivo de influenciar as opiniões de um grupo ou de uma pessoa, de forma intencional, por omissão de fatos ou enfatizando apenas uma narrativa dos fatos. Para que a propaganda seja eficaz, a fonte ou fornecedor precisa entender completamente os valores de seu público-alvo, tornando assim o intelecto do alvo ineficaz. PLN mostrou que as notícias podiam ser separadas com precisão em notícias verdadeiras, notícias falsas e sátira por meio da análise de características linguísticas. Isso sugere que, por meio do uso de ferramentas automatizadas, é possível uma medida de distância da

propaganda ao fato real. Algumas das características linguísticas como o tamanho da história em relação ao título, o uso de palavras descritivas e outros recursos auxiliam na comparação entre as notícias. Considerando que a missão original da rede social é de reunir pessoas com gostos em comum que compartilham informações no espírito de amizade ou companheirismo. No entanto, esses grupos também forneceram um canal para distribuir notícias falsas, pois um usuário pode compartilhar uma notícia sem verificar. Alguns usuários são incapazes de determinar a veracidade de uma notícia quando não dispõem de tempo e recursos para procurar a história em questão, eles se baseiam em atalhos mentais, como preconceitos e reputação da fonte, para determinar a confiabilidade (SAMPLE et al., 2018).

Notícias satíricas são escritas com o objetivo principal de entreter ou criticar os leitores, mas da mesma forma que as notícias falsas, elas podem ser prejudiciais quando compartilhados sem nenhum contexto. Algumas dessas notícias são caracterizadas pelo humor, ironia, absurdo e podem imitar notícias genuínas também. Já a propaganda é definida como informação que tenta influenciar as emoções, como opiniões e as ações do público-alvo por meio de mensagens enganosas, o propósito pode ser político, ideológico ou religioso. *Clickbait* é definido como jornalismo de qualidade baixa que visa atrair tráfego e monetizar em cima de títulos de notícias (BOVET; MAKSE, 2019; BARRETT, 2019).

Por fim, devido à grande difusão de postagens falsas sobre COVID-19, eleições e outros assuntos, foi cunhado o termo "infodemia"que está relacionado à disseminação em massa de notícias falsas e boatos que comprometem a credibilidade de explicações oficiais baseadas em evidências científicas. A desinformação sobre assuntos de saúde, muitas vezes definida como informações que contrariam as melhores evidências disponíveis de especialistas médicos na época, tem sido documentada em quase todas as plataformas de mídia social, incluindo Facebook, Twitter, YouTube, Pinterest e Instagram.

## 2.6 Considerações Finais

Os sistemas de OCR são importantes não somente por serem uma ferramenta que ajuda a dar acessibilidade a documentos ou imagens, eles são também ferramentas que vem integrando e produzindo mais dados úteis para análises complexas dentro de setores da indústria ou de pesquisa. Não existe um consenso sobre qual é o melhor sistema de OCR *open source* disponível atualmente. Há divergências quanto a isso. Por exemplo, enquanto o trabalho de Gabasio (2013) demostra que o sistema Ocropus tem uma taxa de erro de 31.42% e o sistema Tesseract tem um taxa superior de 33.9%, no trabalho de Correia e Rivera (2018) o resultado foi o inverso, ou seja, o sistema Tesseract teve uma taxa de erro de 24.13% e o sistema Ocropus teve uma taxa de 44.26%. O uso de PLN é bem vasto na computação, com ele é possível executar várias tarefas de forma automatizada como exemplo criar de *Chat boots*, auxiliar o OCR usando técnicas de processamento que reduzem um texto de uma página em sentenças ou palavras que

posteriormente com dicionários podem ser corrigidas. O AM auxilia tanto o OCR como o PLN em termos de modelos que analisam estatisticamente uma pilha de dados com diversos padrões que alimentam um modelo, que mais tarde pode por meio de tarefas de classificação, classificar um texto ou até ajudar a identificar possíveis caracteres que não estão legíveis em um texto, mas por meio do padrão ficam fáceis de se deduzir. O próximo capítulo irá tratar dos principais trabalhos relacionados ao tema dessa pesquisa de Mestrado.

# Capítulo 3

# TRABALHOS RELACIONADOS

## 3.1 Considerações Iniciais

Este capítulo é um compilado dos principais trabalhos relacionados ao assunto investigado nessa pesquisa de Mestrado. Para uma melhor organização, os trabalhos foram classificados em três grupos. A primeira seção 3.2 trata especificamente sobre pesquisas na área de redes sociais e conteúdos falsos com ênfase em boatos e notícias falsas. O segundo grupo descrito na seção 3.3 destaca trabalhos sobre COVID-19 e a "infodemia"de informações. Já a última seção 3.4 apresenta o terceiro grupo de trabalhos relacionados sobre conjuntos de dados (*datasets*) relacionados a conteúdos falsos de vários outros domínios.

## 3.2 Redes Sociais e Conteúdos Falsos

A Tabela 7 apresenta a lista dos trabalhos relacionados que discutidos nessa seção.

**Tabela 7 – Tabela com os principais trabalhos de Redes Sociais e Conteúdo Falsos**

| Referencia | Titulo | Foco | Destaque |
|---|---|---|---|
| Singhal et al. (2019) | SpotFake: A multi-modal framework for fake news detection | Notícia | BRET / VGG-19 |
| Shu et al. (2019a) | Studying Fake News via Network Analysis: Detection and Mitigation | Notícia | Redes de grafos e redes neurais profundas |
| Zhou e Zafarani (2019) | Network-based Fake News Detection: A Pattern-driven Approach | Notícia | Redes neurais homogêneas e heterogêneas |
| Monti et al. (2019) | Fake News Detection on Social Media using Geometric Deep Learning | Notícia | Rede Neural Convolucional (RNC) e |
| Recuero et al. (2019) | Cascatas de Fake News Políticas: um estudo de caso no Twitter | Notícia | Gephi |

**Table 7 continuação**

| Referencia | Titulo | Foco | Destaque |
|---|---|---|---|
| Jain e Kasbe (2018) | Fake News Detection | Notícia | *Naive Bayes* |
| Reis et al. (2019a) | Explainable machine learning for fake news detection | Notícia | XGBoost , *K-Means* |
| Sample et al. (2018) | Fake News: A Method to Measure Distance from Fact | Notícia | PLN |
| Nyow e Chua (2019) | Detecting Fake News with Tweets' Properties | Tweets, *Sites* | Árvore de decisão e floresta randômica. |
| Reis et al. (2019b) | Supervised Learning for Fake News Detection | Notícia | k-Vizinhos mais próximos, floresta randômica, algoritmo XG Boost. |
| Faustini e Covões (2019) | Fake news detection using one-class classification | Notícia, Tweet, WhatsApp | *One-Class* SVM |
| Traylor et al. (2019) | Classifying Fake News Articles Using Natural Language Processing to Identify In-Article Attribution as a Supervised Learning Estimator | Documentos | *Bag-of-words* |
| Najar et al. (2019) | Fake news detection using bayesian inference | *Sites* | *Bag-of-words*, EDCM |
| Atodiresei et al. (2018) | Identifying Fake News and Fake Users on Twitter | Tweet | *Named Entity Recognition* (NER), *Part of Speach* (POS), *Stopword*, Análise de Sentimentos |
| Bagade et al. (2020) | The Kauwa-Kaate fake news detection system: DemO | Tweet | Manipulação de imagens usando função *hash* |
| Dong et al. (2020) | Two-path Deep Semi-supervised Learning for Timely Fake News Detection | Notícia | Aprendizado semi-supervisionado, RNC, TDIDF |
| Tschiatschek et al. (2018) | Fake News Detection in Social Networks via Crowd Signals | Notícia | Métodos híbridos de inteligência |
| Bovet e Makse (2019) | Influence of fake news in Twitter during the 2016 US presidential election | Tweet | Análise de rede |

**Table 7 continuação**

| Referencia | Titulo | Foco | Destaque |
|---|---|---|---|
| Previti et al. (2020) | Fake news detection using time series and user features classification. | Tweet | Floresta randômica com 500 árvores |
| Ahmed et al. (2019) | Combining machine learning with knowledge engineering to detect fake news in social networks - A survey | Notícia | Classificadores de Texto |
| Conroy et al. (2015) | Automatic deception detection: Methods for finding fake news | Notícia | Gramáticas Livres de Contexto de Probabilidade. Análise Semântica |
| Pérez-Rosas et al. (2017) | Automatic detection of fake news. | Notícia | SVM linear |
| Rubin et al. (2015) | Automatic deception detection: Methods for finding fake news | Notícia | LIS, CMC, PLN |
| Tandoc et al. (2018) | Defining "Fake News": A typology of scholarly definitions | Notícia | Tipos de conteúdo falso |
| Vosoughi et al. (2018) | The spread of true and false news online | Twitter | Difusão na rede |
| Silva et al. (2019) | Can Machines Learn to Detect Fake News? A Survey Focused on Social Media | Facebook, Twitter e Google+ | Léxicos de sentimento, Modelos de Análise de Topologia e Redes Neurais Artificiais, Algoritmo de Formiga |
| Patwa et al. (2020) | Media-Rich Fake News Detection: A Survey | Notícia | Ngrams, Pontuação, Características psicolinguísticas, Legibilidade e Sintaxe. |
| Pierri e Ceri (2019) | False news on social media: A data-driven survey | Notícia | Aprendizado da máquina tradicional e aprendizado de máquina profundo |
| Mahid et al. (2018a) | Fake news on social media: Brief review on detection techniques | Notícia | N-gram, POS, *bag-of-words* |
| Elhadad et al. (2019) | Fake News Detection on Social Media: A Systematic Survey | Twitter | Aprendizado profundo. |

O trabalho de Krishnan e Chen (2018) desenvolveu um *framework* que possui dois módulos principais que são o *crowd sourcing* e o *crowd sourcing*. O módulo *crowd sourcing* registra e mostra os possíveis usuários para determinados *tweets*, enquanto o módulo *storage manager* é responsável por controlar as interações entre o módulo do *crowd sourcing* e o banco

de dados da Amazon chamado DynamoDB. Um dos recursos extraídos são a *tag* de autenticação do usuário se ele for verificado. Apenas um grupo seletivo recebe esse tipo de classificação como escritores e jornalistas. Outro recurso é a credibilidade da imagem se vier acompanhada, por meio de pesquisa reversa de imagens do Google[1] ajudando a encontrar imagens similares na *Internet* e as fontes de onde foi primeiramente publicada. A pontuação de sentimento é extraída por processo de análise que determina computacionalmente se um texto é positivo, negativo ou neutro, atribuindo uma pontuação que varia entre um negativo e um positivo. Para essa tarefa é utilizado a API TextBlob e o *Natural Language Toolkit* (NLTK). A classificação de *tweets* como falsos ou reais (verdadeiros), com base no extrator do conjunto de recursos é feita pelo algoritmo de árvore de decisão e por SVM usadas para treinar modelos e executar a classificação. Além de classificar como falso ou verdadeiro, outro recurso desenvolvido no *framework* é gerar relatórios de avaliação da credibilidade do usuário.

Outro *framework* criado é o de Singhal et al. (2019) que consiste em uma estrutura multimodal para detecção de notícias falsas, que explora os recursos textuais e visuais de um texto. As observações obtidas na pesquisa confirmam que os recursos multimodais são mais benéficos na detecção de notícias falsas em comparação aos recursos unimodais que só possuem texto como fonte de extração de características. O *framework SpotFake* é dividido em três submódulos: um extrator de recurso textual que usa modelos de linguagens *Bidirectional Encoder Representations from Transformers* (BRET), um extrator de características visuais de postagem tratado com rede convulacional *Visual Geometry Group with 19 Weight Layer* (VGG-19) e um submódulo de fusão multimodal que combina as representações obtidas dos dois módulos anteriores, formando assim um vetor de notícias para utilizar um modelo de detecção.

Shu et al. (2019a) afirmam que as notícias falsas podem mudar a maneira como as pessoas respondem a notícias legítimas, devido ao potencial que possuem para quebrar a confiabilidade dos usuários. O ecossistema de disseminação de notícias nas mídias sociais envolve três dimensões: uma dimensão de conteúdo, uma dimensão social e uma dimensão temporal. A dimensão do conteúdo descreve a correlação entre notícias, publicações nas mídias sociais, comentários etc. A dimensão social envolve as relações entre editores, divulgadores de notícias e consumidores. A dimensão temporal ilustra a evolução dos comportamentos de publicação e publicação dos usuários ao longo do tempo. Essas relações podem ser usadas para detectar e atenuar os efeitos de notícias falsas. A detecção de notícias falsas pode ser formalizada como uma tarefa de classificação que exige extração de recursos e construção de modelos, que podem ser realizadas com a incorporação de redes de grafos e redes neurais profundas. Com os modelos de difusão de rede podem ser aplicados os rastreamentos da origem dos nós e os caminhos de proveniência de notícias falsas. Essas redes podem ser classificadas como homogêneas ou como redes heterogêneas. Redes de interação descrevem os relacionamentos entre diferentes entidades, como editores, notícias e usuários. Por fim, Shu et al. (2019a) conclui que as redes de interação

[1] https://www.labnol.org/reverse/

podem representar as correlações entre diferentes tipos de entidades, como editor, notícias e publicação em mídia social, durante o processo de divulgação das notícias. As características dos editores e usuários e as interações editor-notícias e usuários-notícias têm potencial para dar suporte a detecção de notícias falsas.

Na mesma linha, a abordagem de Zhou e Zafarani (2019) enfatiza que os seres humanos podem se tornar irracionais e vulneráveis na diferenciação entre o que é verdade e o que é falso quando sobrecarregados por informações enganosas. Um estudo empírico aponta que a capacidade humana de detectar engano é de uma precisão média de 54% em 1.000 participantes em mais de 100 experimentos. Um dos padrões encontrados nesse estudo são os relacionamentos dos propagadores de notícia com todos os usuários de sua rede, tornando-se importante o armazenamento dessa estrutura para estudos que expliquem alguns fenômenos. Estudos como esses baseados no comportamento da detecção de notícias falsas em rede utiliza características de contexto social extraídas por meio das redes sociais em sua propagação de notícias. De forma geral, são considerados dois tipos de rede: homogêneas e heterogêneas. Redes homogêneas contêm tipos únicos de nós e arestas. Um exemplo é a rede de apoio, que representa a semelhança entre notícias e postagens de notícias em que apresentam apenas uma postura (opinião) seja ela contra ou a favor. Redes heterogêneas têm vários tipos de nós ou arestas. Essas redes, quando se explora as relações entre entidades como artigos de notícias, editores, usuários, propagam as postagens e as interações entre os usuários de redes homogêneas distintas. Para análise de redes heterogêneas foi utilizado Algoritmo PageRank-like, fatoração de matriz/tensor e Redes Neurais Recorrentes (RNN). Já as redes homogêneas utilizam *frameworks* de mineração para otimização de grafos ou análise da propagação do grafo. O autor conclui que as redes de notícias falsas são mais populares, pois possuem um padrão de maior difusão, percorrem uma distância maior dentro da plataforma por causa de sua rede densa e com forte engajamento. Com os efeitos filtro bolha e câmara de eco, os usuários compartilham entre si atributos, interesses e comportamentos semelhantes, aumentando a coesão do grupo que pode ser de usuários reais ou robôs que aplicam um peso maior na potência do compartilhamento.

Em um período de 2013 a 2018 Monti et al. (2019) extraíram recursos de postagens feitas no Twitter. As características encontradas descrevem as notícias, os usuários e suas atividades na rede social. Para a análise foram agrupados em categorias, uma delas é o perfil do usuário e possui informações de geolocalização, configurações do perfil, idioma, a autodescrição, data de criação da conta e se é uma conta verificada. Outra categoria é de atividade do usuário, com o número de favoritos, listas e *status* (se a conta está em uso). Sobre a estrutura da rede, foi possível obter informações de conexões sociais entre os usuários, número de seguidores, compartilhamentos em cascata, *timestamps* de *retweet*, dispositivo de origem, se é do aplicativo da plataforma ou de um navegador de internet, número de respostas da postagem, quantidade de citações, número de favoritos e *retweets* do *tweet* de origem da cascata. A última categoria envolve o conteúdo, e incorpora as palavras do conteúdo textual do *tweet* incluindo as *hashtags*. O modelo é construído por um grafo de 4 camadas Redes Neurais Convolucional (RNC) com

outras duas camadas convolucionais e duas camadas conectadas para predizer as probabilidades de as notícias serem verdadeiras ou falsas com dois grupos diferentes de análise, um orientado ao endereço da notícia e outro orientado a cascata de compartilhamento. O método proposto permite a integração de dados heterogêneos pertencente a um perfil de usuário, as estruturas de uma rede social, a difusão da notícia e o conteúdo. A importância de uso dessa proposta é a habilidade para automaticamente aprender tarefas atribuídas específicas por meio dos dados. E o aprendizado profundo é importante por se tratar de uma estrutura orientada a grafo. Esse modelo também é geograficamente e linguisticamente independente.

Recuero et al. (2019) coletaram dados, utilizando um API do Twitter a partir de *tweets* com a palavra-chave "lula", constituindo um conjunto de dados de 2.430.468 *tweets*, no período de fevereiro a abril de 2018. Período de julgamento e prisão do ex-presidente Lula, considerado um período com muita difusão de notícias falsas sobre o julgamento do *Habeas Corpus* e o ataque a caravana do presidente. O objetivo, desse artigo foi examinar a estrutura das cascatas de notícias falsas, a sua difusão entre as diferentes identidades ideológicas e o papel dos agentes responsáveis por grandes impactos de difusão dentro da rede social. Para a análise e visualização da rede foi utilizado o software Gephi e as métricas de nós que nesse caso são as postagens do Twitter e as métricas de rede que descrevem a estrutura da rede. As métricas de nó computam a quantidade de conexões realizadas de entrada ou saída de um nó com o outro, e a centralidade de intermediação é o número de nós centrais que fazem a interligação de um nó a outro. As medidas utilizadas para analisar a rede são o coeficiente de agrupamento médio para apontar quanto um nó tende a conectar com um grupo. O caminho médio entre os nós representa a distância média entre eles, quanto menor a medida, mais conectado a rede está. O diâmetro da rede verifica a interconexão da rede, quanto maior ela for mais difusa dentro da plataforma será a postagem.

Jain e Kasbe (2018) utiliza o *Naive Bayes* para classificar as notícias em falsas ou verdadeiras. O Teorema de Bayes trabalha com probabilidade condicional, segundo o qual um evento que deve acontecer, com relação a um certo evento que já ocorreu. A base de dados utilizada são 11 mil artigos com índice, título, texto e classificação de real ou falso. Os artigos costumam usar um mesmo conjunto de palavras para notícias falsas e outro conjunto específico de palavras para as notícias verdadeiras. Foi observado que poucos conjuntos de palavras têm maior frequência de aparecer em notícias falsas do que em notícias verdadeiras e um certo conjunto de palavras pode ser encontrado em alta frequência em notícias verdadeiras. O grupo de palavras foi gerado pelo modelo *bag-of-words*, e a acurácia de detecção de título com informações falsas foi de 80% e para textos foi de 93%.

Reis et al. (2019a) entende que os fenômenos das notícias falsas mudaram dramaticamente a maneira como as notícias são produzidas, disseminadas e consumidas na sociedade. As mudanças iniciaram uma guerra de informações nos últimos anos favorecendo campanhas de desinformação e reduzindo a credibilidade dos veículos de comunicação nesses ambientes. A literatura é bastante ampla se considerado os esforços relacionados à credibilidade da informação,

detecção de boatos e divulgação de notícias. Normalmente são abordados os conteúdos das notícias por técnicas de PLN e outra abordagem é sobre a fonte das notícias e sua confiabilidade e credibilidade por meio da popularidade e características como domínio, endereço IP e outros. Para a avaliação dessas características é utilizado o algoritmo derivado de árvore de decisão chamado de XGBoost que discrimina notícias reais de notícias falsas. O trabalho considerou a possibilidade de utilizar até 172 características extraídas da combinação de características das notícias e redes sociais. A maioria dessas características se concentrou no conteúdo textual e visual do artigo e do título. Algumas delas são estruturas de linguagem, características lexicais, padrões psicolinguísticos, estrutura semântica e subjetividade. A fonte da notícia e o ambiente de interação da rede social são outros grupos de características que podem colaborar na detecção de uma notícia falsa. Para uma geração de modelo imparcial, foram realizados diversos experimentos com diversas combinações dessas características, totalizando um total de 294.292 modelos, cada um desse modelos tinham até 20 características diferentes. Em cada modelo, foi garantido que cada recurso seja incluído o mesmo número de vezes e que nenhum recurso apareça duas vezes no mesmo modelo. Para agrupar esses modelos, foi utilizado o algoritmo *K-Means* baseado em distâncias euclidianas. Para encontrar o valor ideal de K, foi usado o índice de Silhueta, que mede, em média, quão agrupados estão todos os membros em diferentes grupos e seleciona o valor de K mais alto do índice de Silhueta. Além disso foi avaliado as características que agregam melhores valores e as que possuem um melhor desempenho combinados a outras. Embora os recursos das mídias sociais como exemplo quantidade de compartilhamento, o IP do domínio e a toxicidade do conteúdo das notícias, sejam muito frequentes em modelos com baixa variabilidade, recursos do envolvimento do usuário. como exemplo o número de comentários em períodos aproximados de 7200 segundos, o viés político e página do Facebook, ocorreram com mais frequência em modelos com alta variabilidade. Os recursos de nível de presença e os recursos psicolinguísticos são muito frequentes tanto em modelos com alta e baixa variabilidade.

Para combater o problema de propagandas enganosas que são um tipo de informação falsa, os autores Sample et al. (2018) utilizam alguns recursos ainda não estudados como o tamanho de um texto de mesmo conteúdo em sites diversos. Usando como base a quantidade de caracteres de uma notícia no site de verificação de fatos e totalmente independente o *Associated Press* (AP)[2] e comparando com outras plataformas de notícia, consideraram que textos com uma quantidade muito inferior de palavras indicam que a notícia não estaria completa. Outra análise foi feita de acordo com o número de advérbios utilizados no texto. Os advérbios são utilizados para modificar os verbos e aumentando ou diminuindo a intensidade ou a causa da frase em que está presente, tais recursos sensacionalistas são utilizados em notícias falsas como forma de convencer o leitor por meio de suas emoções.

O modelo de Nyow e Chua (2019) é construído com a introdução de um grupo de dados treinados com classificações pré-categorizadas. Dessa forma, o algoritmo pode aprender no

[2] https://apnews.com/

processamento dos dados. Depois o modelo utiliza uma outra fonte de dados e as classes são retiradas, deixando que o modelo identifique os valores das classes. Algumas das características do twitter extraídas para implementar esse modelo são os números totais de *tweets*, *tweets* favoritos e re*tweets* e a porcentagem dos *tweets* disponíveis. São extraídos dos endereços eletrônicos localizados nos *posts* o protocolo utilizado, número de diretórios e se o *site* possui o www no endereço. Pelo título da notícia é computado a quantidade de palavras, e por último é adicionado um *status* da notícia se é real ou falsa. Essa classificação é feita por sites de verificações de fatos como PolitictFacto e GossipCop. Os métodos utilizados para a classificação são árvore de decisão e floresta randômica. A floresta randômica conseguiu alcançar uma acurácia de 98.6% com base nos aspectos dos atributos do *tweet* e da notícia. Um outro estudo Reis et al. (2019b) usa o aprendizado supervisionado, que extrai características linguísticas e semânticas do corpo da notícia e também informações da fonte da notícia e atributos das redes sociais são calculados por meio de algoritmos e classificadores, tais como os k-vizinhos mais próximos, floresta randômica e o algoritmo XG Boost.

O trabalho de Faustini e Covões (2019) faz uma abordagem de classificação, na qual os dados possuem apenas uma classe. Durante a fase de teste, se um objeto não se encaixa nas características da classe de destino, é considerado um ponto externo, o objetivo é distinguir entre objetos da classe de destino interna e objetos de destinos externos. Foram usados dados de notícias falsas para o aprendizado do modelo. Algoritmos baseados em SVM estão entre as técnicas mais populares nas classificações de uma classe. Um deles, o *One-Class* SVM propõe uma função binária para identificar regiões nas quais a maioria dos dados se encontra. A redução de dimensionalidade funciona com a extração dos recursos do texto bruto. Em seguida, um vetor é criado a partir da soma de todos os vetores de recursos dos objetos. Esse vetor é chamado vetor de classe. O valor exclusivo que representa cada objeto nos dados de treinamento é a distância entre si e o vetor de classe. Os resultados mostram que os valores para SVM de classe única são os mais bem alcançados após a pesquisa em grade sobre a função polinomial e radial dos núcleos.

Traylor et al. (2019) usa *Grounded Theory* que é uma técnica de pesquisa em ciências sociais de base indutiva usada para construir teorias e estruturas a partir de dados existentes observando os dados e procurando padrões, tendências e diferenças. Os padrões linguísticos foram utilizados para desenvolver uma hipótese de aprendizado de máquina. A base utilizada para este trabalho foi um *corpus* com 421 citações de documentos que a equipe de pesquisa classificou entre verdadeiros e falsos. A classificação binária resultante é baseada na presença de informações descobertas com um modelo do tipo *bag-of-words* na origem do texto. A identificação da sugestão é baseada na aprendizagem de verbos associados ou informações de sugestão contidas no conjunto de treinamento. Outro estudo que também utiliza *bag-of-words* é o de Najar et al. (2019), no qual um framework de detecção de notícias falsas foi aplicado em um conjunto de dados chamado *Bs Detector*, que incluiu 244 sites com vários textos que possuem num total 172.817 palavras. Algumas tarefas executadas antes foram a remoção de todas as

palavras curtas com menos de 2 caracteres, palavras muito longas com mais de 15 caracteres, palavras raras com menos de 10 ocorrências do vocabulário associado ao conjunto de dados. Com base no processamento e nas métricas de avaliação é comprovada a robustez do aprendizado *Bayesiano* para estimar os parâmetros do *Exponential Dirichlet Compound Multinomial* (EDCM) no caso de detecção de notícias falsas. Na teoria e estatística das probabilidades, EDCM é uma família de distribuições de probabilidades multivariadas discretas em um suporte finito de números inteiros não negativos.

As técnicas utilizadas em Atodiresei et al. (2018) envolvem aprendizado automático, uso de dicionários com termos específicos de um domínio monitorado, classificadores *Naive Bayes*, modelos baseados em Entropia Máxima e SVM, combinados com ferramentas específicas para processamento de linguagem natural comøo *Part of Speach* (POS), remoção de *stop-word*, identificação de sentimentos, dentre outras. O *crawler* desse trabalho captura *tweets* e os adiciona em um banco de dados. O primeiro módulo é possível inserir o *link* do *tweet* na interface para o processamento da credibilidade de um *tweet* novo. O algoritmo criado para o cálculo da credibilidade usa um componente *Named Entity Recognition* (NER), que divide o texto em partes compostas traz as entidades que são geralmente substantivos com a importância relativa no contexto, os tópicos, as *tags* sociais, o sentimento geral do *tweet* e da *hashtag*. Para a análise NER, utilizam uma API pública chamada OpenCalais, enquanto para o cálculo de sentimentos, usam o Sentiment140. No módulo de análise combinam a verificação do *tweet* atual com a base de dados armazenados localmente em um banco de dados. Outra forma é a procura de *tweets* semelhantes em fontes confiáveis e de perfis verificados, como exemplo a BBC que é um perfil jornalístico. Por fim, os *tweets* são armazenados e há um aumento da pontuação de credibilidade quando o *tweet* é semelhante ao *tweet* de um perfil verificado.

O estudo de Bagade et al. (2020) é similar ao de Atodiresei et al. (2018), porém eles fornecem várias interfaces mais amigáveis que também verificam a autenticidade de um *site* de um determinado *tweet*. Outro diferencial do trabalho de BAGADE et al. é a manipulação de imagens usando a função *hash* no conteúdo da imagem para oferecer suporte à correspondência exata, extraindo e indexando assinaturas de imagem para oferecer suporte à indexação aproximada. Quando um artigo que está sendo verificado tem exatamente a mesma imagem que em um site de verificação de fatos, basta uma *hash* na imagem para encontrar a sua correspondência. Outro problema que pode ser encontrado são as capturas de tela compartilhadas com informações textuais. Para lidar com esse problema, foi utilizado a biblioteca de OCR do Tesseract nas imagens em artigos a serem verificados. O fluxo do modelo inicia com um usuário enviando um artigo ou *tweet* do aplicativo ao principal gerenciador de conteúdo. O gerenciador recebe também as informações dos *crawlers* e do módulo de verificação. As análises e classificações são realizadas em outro módulo que pode envolver interação humana se necessário, caso contrário após toda a classificação o artigo é enviado para o banco de dados se ainda não tiver uma notícia parecida.

Neste artigo, Dong et al. (2020) propõem uma estrutura profunda de aprendizado semi-supervisionado por meio da construção de redes neurais convolucionais de dois caminhos que realiza a detecção de notícias falsas no caso de dados que foram limitadamente rotulados. Devido à rápida propagação de notícias falsas, a detecção oportuna é essencial para mitigar seus efeitos. No entanto, geralmente muito poucas amostras de dados podem ser rotuladas em pouco tempo, o que, por sua vez, inviabiliza os modelos de aprendizado supervisionado. Portanto, um modelo de aprendizado semi-supervisionado profundo é implementado o modelo que consiste em três componentes, uma RNC que compartilha outra RNC supervisionada e uma RNC não supervisionada. Aplica *Term Frequency–Inverse Document Frequency* (TF-IDF) para extrair recursos e empregar técnicas de incorporação de palavras, como word2vec. Uma tarefa é aprender como extrair padrões de notícias falsas que já foram classificadas, enquanto a outra tarefa é otimizar as representações das notícias que não estão classificadas. Por isso existe uma RNC compartilhada para extrair recursos de baixo nível para alimentar as duas RNCs posteriores. A estrutura proposta envolve tanto aprendizado supervisionado quanto aprendizado não supervisionado, enquanto todas as tarefas do aprendizado profundo sobre várias tarefas são baseadas apenas no aprendizado supervisionado. Esses dois caminhos podem ter RNCs independentes com configurações diferentes ou idênticas para aprendizado supervisionado e aprendizado não supervisionado, respectivamente.

Existem dois tipos tradicionais de detecção de notícias, a verificação de um especialista e os métodos computacionais. Tschiatschek et al. (2018) percebe algumas limitações dos métodos computacionais atuais e constrói uma abordagem alternativa com métodos híbridos de inteligência artificial e inteligência humana. Neste trabalho foi desenvolvido um algoritmo chamado DETECTIVE que implementa uma abordagem de rede *Bayesiana* para efetivamente utilizar a sinalização de notícias feitas por uma multidão de usuários e detectar notícias falsas. Dado um grupo de notícias, o objetivo do algoritmo é selecionar um pequeno subconjunto de todas essas notícias, e enviar para um especialista de verificação e depois bloquear todas as notícias que retornarem como falsas depois da avaliação.

Bovet e Makse (2019) compilam uma base de notícias do Twitter, analisando todos os *tweets* em um conjunto de dados que continham pelo menos uma endereço *Web* vinculado a um *site* fora do Twitter. Como primeira tarefa separam-se em duas categorias os sites vinculados no *tweet*: *sites* que contêm informações incorretas e meios de comunicação tradicionais baseados em fatos. A categoria de notícias incorretas é separada em duas subcategorias de notícias falsas e as de notícias extremamente tendenciosas. Sites de notícias falsas são sites que foram sinalizados por disseminar consistentemente notícias fabricadas ou teorias de conspiração por várias agências de verificação de fatos. Sites extremamente tendenciosos incluem sites mais controversos que não necessariamente publicam informações fabricadas, mas distorcem fatos e podem fazer o uso excessivo de *clickbaits*, informações descontextualizadas ou opiniões distorcidas como fatos. Para a categoria de notícias tradicionais, foram separadas as orientações políticas dos sites. Os *tweets* foram coletados pelo Twitter Search API entre junho de 2016 a novembro de 2016,

totalizando 171 milhões de *tweets* no idioma inglês, mencionando os dois principais candidatos concorrendo nas eleições gerais para a presidência dos Estados Unidos. Foram filtrados 30,7 milhões de *tweets* com um URL direcionado para um site de notícias, enviados por 2,3 milhões de usuários. Para descobrir os usuários mais influentes de cada rede de *retweet*, foi usado o algoritmo de Influência Coletiva baseado na solução da melhor vazão de uma rede. Para que um usuário do Twitter seja altamente classificado pelo algoritmo do Influência Coletiva, ele não precisa necessariamente realizar o *retweet* diretamente, mas ele precisa estar cercado por usuários que foram *retweet*ados. Na execução das análises foi utilizado a ferramenta de graph-tool e python, com módulos de análise de rede, e o software TIGRAMITE usado para analisar séries temporais.

Previti et al. (2020) investigam a difusão diferencial de notícias verdadeiras e falsas devidamente verificadas no Twitter, entre os anos de 2006 e 2017, selecionando 126 mil histórias compartilhadas e publicadas por 3 milhões de usuários, para devidamente encontrar os pontos com maiores difusão tanto de uma categoria quanto da outra, por meio de marcas temporais, quantidades de *tweets*, cálculos de média e quantidade de seguidores envolvidos em todas as cascatas de um boato. Para o treinamento dos dados é utilizado o classificador de Floresta randômica com 500 árvores, após o treino é aplicado a técnica de Gini-importance para calcular a importância de cada propriedade extraída dos *tweets*. Explorando as características temporais e a informação sobre os usuários (como número de seguidores e quantidades de *tweets*) há uma acurácia de 84,61% na pré-categorização de notícias por um período de 24hs com amostras de 1h.

A integração da máquina de aprendizagem e da engenharia do conhecimento pode ser útil na detecção de notícias falsas. combinação de dados orientados e conhecimento projetado para combater notícias falsas. Nos estudos de (AHMED et al., 2019) foram comparados três tipos diferentes de classificadores de texto, aplicativos de detecção de postura e verificação de fatos que podem ajudar a detectar notícias falsas. O primeiro passo na detecção de notícias falsas é classificar o texto imediatamente após a notícia publicada *online*. A classificação do texto é uma das questões importantes da pesquisa no campo da mineração de texto.

Uma das tarefas mais difíceis nessa nova geração de jornalismo é conseguir categorizar a quantidade de verdade que pode ser associada a uma notícia publicada. A natureza da publicação de notícias *online* mudou, de tal forma que a verificação de fatos de possíveis enganos é impossível contra a inundação decorrente de geradores de conteúdo, bem como vários formatos e gêneros. Algumas variedades ou métodos de avaliação da veracidade são de dois grupos: abordagens linguísticas e abordagens de análise de rede. O objetivo na abordagem linguística é procurar tais casos de vazamento ou, os chamados "sinais de engano preditivo"encontrados no conteúdo de uma mensagem. O método mais simples de representar textos é a abordagem do saco de palavras. Na abordagem do saco de palavras, as palavras individuais ou as frequências n-Grams são agregadas e analisadas para revelar sinais de engano. Além de depender exclu-

sivamente da linguagem, o método conta com n-Grams isoladas, muitas vezes divorciados de informações de contexto úteis. Estruturas linguísticas mais profundas foram analisadas para prever casos de engano. Também foi realiza análise profunda da sintaxe para prever casos de engano, implementada usando Gramáticas Livres de Contexto de Probabilidade.

Análise Semântica caracteriza o grau de compatibilidade entre uma experiência pessoal em comparação com um "perfil"de conteúdo derivado de uma coleta de dados análogos. O conteúdo extraído de palavras-chave consiste em atributo: par descritor. Existem duas limitações potenciais neste método: a capacidade de determinar o alinhamento entre atributos e descrições depende de uma quantidade suficiente de conteúdo extraído de perfis, e o desafio de associar corretamente descritores com atributos extraídos. Estrutura retórica e Análise do Discurso: No nível do discurso, as sugestões de engano se apresentam tanto nas comunicações da CMC quanto no conteúdo de notícias. Uma descrição do discurso pode ser alcançada por meio do quadro analítico da Teoria da Estrutura Retórica (RST), que identifica instâncias de relações retóricas entre elementos linguísticos. O uso de relações retóricas pode ser um indicativo de engano. Ferramentas para automatizar a classificação retórica estão se tornando disponíveis, embora ainda não empregadas no contexto de avaliação da veracidade. (CONROY et al., 2015)

Neste trabalho de Pérez-Rosas et al. (2017) são introduzidos dois novos conjuntos de dados para a tarefa de detecção de notícias falsas, abrangendo sete diferentes domínios de notícias. O trabalho descreve o processo de coleta, anotação e validação em detalhes e apresenta diversas análises exploratórias sobre a identificação de diferenças linguísticas em conteúdo de notícias falsas e verdadeiras. Além disso, o trabalho descreve um conjunto de experimentos de aprendizagem para construir detectores de notícias falsas precisos e fornece análises comparativas da identificação automática e manual de notícias falsas. Algumas características são linguísticas como N-Grams, extrato de uni-Grams e bi-Grams derivados do saco de palavras para representação de cada artigo de notícias. Para explicar diferenças ocasionais no comprimento do conteúdo, esses recursos são codificados como valores de TF-IDF. Outras medidas usadas são Pontuação, Características Psicolinguísticas (emoções positivas), Legibilidade e Sintaxe. O trabalho usou um classificador SVM linear e cinco vezes validação cruzada, com as medidas de precisão, revocação e medida F1 médias ao longo das cinco iterações.

Para análise de texto e modelagem preditiva, continua sendo essencial filtrar, vetar e verificar informações *online* tanto em bibliotecas quanto na área da Ciência da Informação (LIS). Isso decorre do fato que as linhas entre notícias tradicionais e informações *online* estão confusas e difíceis de identificar. Embora a questão da verificação de notícias e verificação de fatos tenha sido tradicionalmente uma questão de esforço jornalístico, a LIS oferece *insights* significativos sobre avaliação de credibilidade e tecnologias de informação de ponta. Como o LIS está se redefinindo na era do *fast-streaming* de *Big Data* e análise de texto, ele está se voltando para métodos automatizados para filtrar, vetar e verificar informações *online*. Notícias enganosas, como notícias falsas, falsos comunicados de imprensa e farsas, podem ser enganosas

ou até mesmo prejudiciais, especialmente quando são desconectadas de suas fontes e contextos originais. Na psicologia interpessoal, questionários, entrevistas, discussões de cenários de caso e dados de observação (muitas vezes em formato verbal) tendem a formar a base para análises. Confiabilidade e consistência das pistas são contestadas, mas um fato é regularmente citado: ao detectar mentiras os humanos não fazem muito melhor do que o acaso, e as máquinas podem ligeiramente superar os humanos em tarefas restritas. Na comunicação mediada por computador (CMC), *corpora* (de *e-mails*, *posts* em fóruns, etc.) são coletados por meio de eleições ou submissões de mensagens pré-existentes. Dados como *e-mail* interpessoal enganoso (Keila Skillicorn, 2005) podem ser reaproveitados para análise de texto. Os dados de mídia social disponíveis publicamente são uma fonte frutífera para a análise de sentimentos. Os tweets são especialmente adequados para detectar tais comportamentos irregulares de comunicação no compartilhamento de informações como fraudes. Em PLN (ou análise de texto), outros tipos de dados falsos foram coletados rastreando a *Web* ou *crowdsourcing*: revisões de produtos falsos, currículos*online* falsificados, *spam* de opinião, perfis falsos de redes sociais, perfis falsos de namoro, *spam* e *phishing* e trabalho científico forjado. Requisitos para Detecção de Notícias Falsas, *Corpus* são disponibilizados com casos verdadeiros e falsos; acessibilidade do formato textual digital e verificabilidade da verdade. A questão é o que constitui a verificação e como é realizada: homogeneidade em comprimentos; homogeneidade na matéria escrita; prazo predefinido; a forma de entrega de notícias (por exemplo, humor, credibilidade, absurdo, sensacionalismo); preocupações pragmáticas; língua e cultura.

Notícias enganosas podem ser colhidas, *crowdsourced* ou imitadas por participantes qualificados do estudo e constituem sub-tarefas na detecção de notícias falsas. A reportagem fraudulenta não é inédita tanto em mídias antigas quanto em novas mídias. A imprensa amarela e os tabloides apresentam um amplo espectro de novidades não verificadas e usa manchetes atraentes ('clickbait's'). O jornalismo amarelo é uma fonte adequada para o corpus de notícias falsas em casos de falsificação, fabricação ou exagero óbvio ou exposto, e pode exigir investigação. O boato é outro tipo de fabricação deliberada ou falsificação nas mídias tradicionais ou sociais. Tentativas de enganar audiências mascaradas como notícia, e pode ser escolhida para cima e equivocadamente validado por veículos de notícias tradicionais. Se os leitores estão cientes da intenção humorística, eles podem não estar mais predispostos a levar a informação em um formato típico do jornalismo *mainstream*, mas dependem fortemente da ironia e do humor morto para imitar uma fonte de notícias genuína, imitando fontes de notícias confiáveis e histórias, e muitas vezes alcançando ampla distribuição.(RUBIN et al., 2015)

O caso do *Pizzagate* que é considerado uma dentre muitas teorias da conspiração, é apenas um dos numerosos casos de notícias falsas que inundam as mídias sociais. Do Papa Francisco endossando o então candidato presidencial republicano das eleições de 2016, para uma mulher presa por defecar na mesa de seu chefe depois que ela ganhou na loteria. O termo *Fake News* não é novo. O discurso contemporâneo, particularmente a cobertura da mídia, parece definir notícias falsas como se referir a postagens virais baseadas em contas fictícias feito para

parecer notícias verdadeiras. Um estudo recente definiu notícias falsas "para serem artigos de notícias que são intencionalmente e verificáveis falsas e podem enganar os leitores". Duas principais motivações estão por trás da produção de notícias falsas: financeira e ideológica. Uma revisão de estudos anteriores que usaram o termo *Fake News* revela seis tipos de definição: (1) sátira de notícias, (2) paródia de notícias, (3) fabricação, (4) manipulação, (5) publicidade, e (6) propaganda. O que é comum nessas definições é como as notícias falsas se apropriam da aparência e da sensação de notícias reais; de como os sites se parecem; à forma como os artigos são escritos; de como as fotos incluem atribuições. As notícias falsas se escondem sob uma face de legitimidade à medida que assume alguma forma de credibilidade, tentando parecer uma notícia verdadeira baseada em fatos reais. Além disso, indo além do simples aparecimento de um item de notícias, por meio do uso de *boots* de notícias, as notícias falsas imitam a onipresença das notícias construindo uma rede de sites falsos. A primeira dimensão, refere-se ao grau em que as notícias falsas se baseiam em fatos. Por exemplo, a sátira se baseia em fatos, mas a apresenta em um formato de desvio, enquanto paródias e notícias fabricadas tomam um amplo contexto social sobre o qual ela forma contas fictícias. A publicidade nativa usa fatos, enquanto as fabricações são sem base factual. A segunda dimensão, que é a intenção imediata do autor, refere-se ao grau em que o criador das *Fake News* pretende enganar. Sátiras de notícias e paródias usam algum nível de suspensão mutuamente compreendida da realidade para funcionar — a intenção imediata é divertir os leitores por meio de algum nível de fatos de dobra. Os autores de fabricação e manipulação, por sua vez, pretendem no ponto de partida enganar, sem qualquer aviso. Embora, em última análise, o objetivo da fabricação e manipulação seja desinformar as pessoas ou apenas atrair cliques para publicidade de dinheiro, tais metas são alcançadas por intermédio da intenção imediata de enganar as pessoas de que as notícias falsas que leem são reais.(TANDOC et al., 2018)

(VOSOUGHI et al., 2018) realizaram o levantamento de notícias falsas e verdadeiras do Twitter de 2006 a 2017, compreendendo 126.000 histórias tuitadas por 3 milhões de pessoas mais de 4,5 milhões de vezes. A classificação foi feita por 6 verificadores independentes de fatos. O estudo descobriu que notícias falsas eram mais novas do que notícias verdadeiras, o que sugere que as pessoas eram mais propensas a compartilhar novas informações. Ao contrário da sabedoria convencional, os robôs aceleraram a disseminação de notícias verdadeiras e falsas na mesma taxa, implicando que notícias falsas se espalham mais do que a verdade porque os humanos, não os robôs são mais propensos a espalhá-la. Estudos sobre a disseminação da desinformação estão atualmente limitados a análises de pequenas amostras *ad hoc* que ignoram duas das questões científicas mais importantes: como a verdade e a falsidade difundem de forma diferente, e quais fatores do julgamento humano explicam essas diferenças. Eles quantificaram a profundidade das cascatas (o número de saltos retweet do tweet de origem ao longo do tempo, onde um lúpulo é um retweet por um novo usuário único), tamanho (o número de usuários envolvidos na cascata ao longo do tempo), amplitude máxima (o número máximo de usuários envolvidos na cascata em qualquer profundidade) e viralidade estrutural (uma medida que interpola entre o conteúdo

espalhado por várias gerações, com qualquer indivíduo diretamente responsável por apenas uma fração do espalhamento total). Uma maior fração de boatos falsos experimentados entre 1 e 1000 cascatas, enquanto uma fração maior de rumores reais experimentou mais de 1000 cascatas. Os autores descobriram que a falsidade difundiu significativamente mais longe, mais rápido, mais profundo e mais amplamente do que a verdade em todas as categorias de informação.

Por meio de leituras, (SILVA et al., 2019) constatou que a maioria das obras utiliza as mídias sociais e micro *blogs* como sua principal fonte de análise. Isso se deve ao uso crescente de redes sociais por todos, como Facebook, Twitter e Google+. Principalmente Twitter e Sina Weibo. Além disso, as plataformas de *microblogging* geralmente fornecem uma API (*Application Programming Interface*) para consultar e consumir seus dados. A API geralmente fornece o conteúdo da plataforma em dados estruturados ou texto simples, reduzindo assim a etapa de pré-processamento que é comumente usada com rastreadores webs usados para filtrar as informações de interesse das páginas da Web. Redes sociais que expressam opiniões individuais de muitos usuários diferentes com diferentes crenças, contextos e origens culturais. Muitos artigos usaram a análise de sentimentos para classificar a polaridade de uma notícia. Alguns usaram léxicos de sentimento, que exigem muito esforço humano para construir e manter e construir um classificador supervisionado baseado em aprendizagem. O uso de outras técnicas baseadas na sintaxe é relativamente baixo. Os artigos usam principalmente *parsing*, *pos-tagging* e tipos de entidades nomeadas. Por outro lado, o uso da semântica é mais comum. Muitos artigos usavam léxicos como conhecimento externo sobre palavras, criando listas de palavras baseadas em propriedades de interesse. Outro uso da semântica na detecção de notícias falsas é o uso de modelagem de linguagem. Alguns papéis usaram *n-Grams* como linhas de base para comparações com suas características artesanais. Outros usaram *n-Grams* como características para seus classificadores. Trabalhos mais recentes usaram incorporações de palavras para modelagem de idiomas, principalmente aqueles que estão construindo um classificador usando aprendizado não supervisionado. Incorporações de palavras são uma família de modelos linguísticos, onde um vocabulário é mapeado para um vetor de alta dimensão. Esses modelos de linguagem atribuem um vetor real a cada palavra no vocabulário. O agrupamento de recursos encontrados nos classificadores é definido com base na fonte do recurso. O primeiro conjunto de grupos apresenta recursos baseados em atributos de mídia social (curtidas, retweets, amigos), o segundo conjunto tem recursos baseados no conteúdo das notícias (pontuações, incorporações de palavras, polaridade de sentimento das palavras). Algoritmos de classificação que se concentram fortemente nos aspectos linguísticos. Modelos de Análise de Topologia e Redes Neurais Artificiais que exploram o vínculo entre usuários e outras metas informações fornecidas pela estrutura de dados predefinidas das mídias sociais, alguns autores propõem classificar as entradas das mídias sociais como falsas, analisando sua interação entre os usuários. O Algoritmo de Formiga funciona muito como uma colônia de formigas, as notícias são pulverizadas com feromônios, enquanto existe tal nas proximidades dos dados adquiridos, o algoritmo opera até que o feromônio evapora, cada vez mais prevendo e atualizando sua razão de erro, até que o fio de feromônios totais seja totalmente evaporado.

O problema de informações falsas e verdadeiras tem sido tradicionalmente formulado como um problema de classificação binária supervisionada, começando por conjuntos de dados compostos por artigos de notícias rotulados, tweets relacionados e postagens no Facebook que permitem capturar diferentes características, desde as bases em conteúdo (texto, imagem, vídeo) até relativas ao contexto social (redes de difusão, perfil dos usuários, metadados) e, em alguns casos, a bases de conhecimento externo (Wikipedia , Google News). Para o que diz respeito ao método de classificação, são utilizados uma ampla gama de técnicas, desde o aprendizado da máquina tradicional (Regressão Logística, Máquinas vetoriais de suporte, Floresta Aleatória) até o aprendizado profundo (Redes Neurais Convolucionais e Recorrentes) e outros modelos (Factorização Matricial, Inferência Bayesiana) (PIERRI; CERI, 2019).

As abordagens baseadas no conhecimento por especialistas é um processo muito exigente e demorado, pois depende de especialistas humanos para validar a reivindicação. A abordagem baseada no conhecimento por *crowdsourcing* envolve a participação de um público de pessoas não especialistas para usar sua sabedoria, conhecimento e bom senso para interpretar o conteúdo das notícias, depois os dados são sintetizados para se alcançar um consenso. Trabalhos recentes propuseram um método que usa recursos visuais e estatísticos para detectar notícias falsas em *microblogs*. A detecção de fraude baseada em estilo é um método que pode ser utilizado para detectar pistas de fraude no texto, considerando o estilo de escrita do conteúdo da notícia. A estrutura analítica da teoria da estrutura retórica avalia a coerência e a estrutura histórica para diferenciar conteúdo falso do verdadeiro, identificando informações deliberadamente enganosas. Este método de detecção de fraude baseado em estilo atinge taxas de precisão relativamente altas de 90% de precisão, 84% de revocação e 87% de medida F1. A detecção de notícias falsas é um mecanismo ou sistema que auxilia os usuários com as ferramentas e funções na predição de notícias falsas. A classificação das técnicas varia desde aquelas estritamente baseadas em conteúdo a abordagens baseadas em contexto social e híbridas. Alguns elementos são extraídos para auxiliar na classificação, a estrutura linguística extrai o conteúdo do texto por meio de caracteres, palavras e frases. Outras técnicas que podem ser utilizadas é o *bag-of-words*. Outro tipo de representação de texto são métodos que utilizam os recursos lexicais e sintáticos, como análise superficial, pontuação e POS (MAHID et al., 2018b).

O avanço da tecnologia móvel permitiu que as mídias sociais desempenhassem um papel vital na organização de atividades a favor e ou contra países ou governos. Os tipos de dados e as categorias de recursos utilizados no modelo de detecção, bem como os conjuntos de dados de referência são discutidos no trabalho deElhadad et al. (2019). Eles definem a desinformação como uma informação falsa criada e compartilhada por pessoas com intenção de prejudicar. Alguns recursos podem ser usados para ajudar na identificação, tai recursos são baseados em conteúdo, baseados em contexto e baseados em domínio. No conteúdo pode-se extrair características textuais, já no de contexto pode-se extrair informações sobre os indivíduos que criaram e compartilharam, grupos que consomem determinada informação, recursos de base social, como número de seguidores, contagem de amigos, idade de registro, número de

*posts*/tweets de autoria, grupos relacionados sociais, informações demográficas, postura do usuário, pontuação média de crédito, etc. Além disso, pode-se extrair recursos sobre o domínio a que a notícia pertence, extraindo recursos de propagação que consideram características relacionadas à árvore de propagação, que pode ser construídos a partir dos retweets de uma mensagem em um determinado domínio (domínios aqui podem ser governamentais ou não, e também o nome do site em questão). O autor usou os vetores de características ponderadas para melhorar os resultados de classificação por meio de aprendizado profundo.

## 3.3 COVID-19 e Conteúdos Falsos

A Tabela 8 ilustra com os trabalhos relacionados a detecção de conteúdos falsos sobre a pandemia de COVID-19, destacando o enfoque de cada um.

**Tabela 8 – Tabela com os trabalhos de COVID19**

| Referencia | Titulo | Foco |
|---|---|---|
| Bang et al. (2021) | Model Generalization on COVID-19 Fake News Detection | Twitter, Facebook e Instagram |
| Gundapu e Mamidi (2021) | Transformer based Automatic COVID-19 Fake News Detection System | Twitter, Facebook e Instagram |
| Chen et al. (2021) | Transformer-based Language Model Fine-tuning Methods for COVID-19 Fake News Detection | Twitter, Facebook e Instagram |
| Wani et al. (2021) | Evaluating Deep Learning Approaches for Covid19 Fake News Detection | Twitter, Facebook e Instagram |
| Koloski et al. (2021) | Identification of COVID-19 related Fake News via Neural Stacking | Twitter, Facebook e Instagram |
| Das et al. (2021) | A Heuristic-driven Ensemble Framework for COVID-19 Fake News Detection | Twitter, Facebook e Instagram |
| Li e Li (2021) | Exploring Text-transformers in AAAI 2021 Shared Task: COVID-19 Fake News Detection in English | Twitter, Facebook e Instagram |
| Gupta et al. (2021) | Hostility Detection and Covid-19 Fake News Detection in Social Media | Notícia |
| Galhardi et al. (2020) | Fact or fake? an analysis of disinformation regarding the covid-19 pandemic in brazil. | WhatsApp e Facebook |
| Linden et al. (2020) | Inoculating against fake news about covid-19. | Notícia |

**Table 8 continuação**

| Referencia | Titulo | Enfoque |
|---|---|---|
| Apuke e Omar (2021) | Fake news and covid-19: modelling the predictors of fake news sharing among social media users. | Notícia |
| Júnior et al. (2020) | Da desinformação ao caos: uma análise das fake news frente à pandemia do coronavírus (covid-19) no brasil. | Notícia |
| Neto et al. (2020) | Fake news no cenário da pandemia de covid-19. | Notícia |
| Ferreira et al. (2020) | Desinformação, infodemia e caos social: impactos negativos das fake news no cenário da covid-19. | Teórico |
| Singh et al. (2020) | A first look at covid-19 information and misinformation sharing on twitter. | Twitter |
| Rovetta e Bhagavathula (2020) | Global infodemiology of covid-19: Analysis of google web searches and instagram *hashtags*. | Google Trends e Instagram |
| Inuwa-Dutse (2021) | Towards combating pandemic-related misinformation in social media. | Twitter |

Os trabalhos de Bang et al. (2021), Gundapu e Mamidi (2021), Chen et al. (2021), Wani et al. (2021), Koloski et al. (2021), Das et al. (2021), Li e Li (2021) utilizaram um conjunto de dados divulgado pelo *Workshop* de tarefas compartilhadas CONSTRAINT 2021. A meta desse evento foi combater a divulgação de notícias falsas sobre o COVID-19 em plataformas de mídia social, tais como Twitter, Facebook, Instagram e qualquer outra rede social popular. O conjunto de dados consiste em 10.700 *posts* que foram realizados em redes sociais e artigos de notícias falsas, todos esses conteúdos foram escritos na língua inglesa.

*Bidirectional Encoder Representations from Transformers* (BERT) é uma técnica de aprendizado de máquina baseada em *Transformers* para pré-treinamento de PLN desenvolvida pelo Google. XLNet é uma versão aprimorada do BERT. Para entender o contexto da linguagem mais profundamente, a *XLNet* usa o *Transformer-XL* como um modelo de engenharia de recursos, que por si só é uma adoção sobre o Transformador nativo. Este modelo *Transformer-XL* integra dois componentes: Mecanismo de Recorrência e Codificação Posicional Relativa (RPE). Isso é feito por meio de integração ao Transformador usado em BERT para lidar com as dependências de longo prazo para textos que são mais longos do que o comprimento máximo de entrada permitido. O Mecanismo de Recorrência descreve o contexto entre duas sequências em segmentos específicos e o RPE, que carrega informações de similaridade entre dois *tokens*.

De acordo com (BANG et al., 2021). ao lidar com texto, os modelos de PLN baseados em *Transformers* (modelos de aprendizado profundo) são comumente usados como extratores de características textuais, por contarem com uma larga escala de dados pré-treinados. As funções de perda que são calculadas e comparadas com a tradicional função de perda de entropia cruzada não ajudaram muito na melhoria do valor de F1 em modelos que usam o conjunto de dados FakeNews-19 (*dataset* disponibilizado pelo evento). Entretanto, mostraram melhor capacidade de generalização no grupo de dados de Tweets-19, uma troca justa se comparado com os resultados entre RoBERTa-large (*Robustly Optimized BERT Pretraining Approach*) com perda por entropia cruzada e *curriculum loss*. Ao realizar a limpeza de dados de influência com alta porcentagem de limpeza, a porcentagem de limpeza pode alcançar o melhor desempenho de avaliação no Tweets-19 com 61,10% de precisão e 54,33% de medida F1, mantendo o alto desempenho de teste suficiente no FakeNews-19.

Com relção ao trabalho de (GUNDAPU; MAMIDI, 2021), inicialmente foram desenvolvidos algoritmos de Aprendizado de Máquina usando Frequência do Termo–Inverso da Frequência nos Documentos, que também é conhecido pelo seu acrônimo em inglês TF-IDF, o qual é usado para detectar desinformação sobre o conjunto de dados fornecido. Esses métodos supervisionados de TF-IDF ainda são relevantes para muitas tarefas de classificação e tiveram um desempenho muito bom para detecção de notícias falsas. Além disso, foi desenvolvido um modelo de conjunto eficaz integrado com três modelos de transformadores (BERT, ALBERT e XLNet) para detectar notícias falsas nas plataformas de mídia social. Isso resultou em maior precisão e um modelo mais generalizado.

No modelo de Chen et al. (2021) também é proposto uma abordagem baseada em *Transformers*. Como primeiro passo, o vocabulário simbólico do modelo individual é expandido para a semântica de frases reais e profissionais. Em segundo, é adaptado a perda por SoftMax para distinguir as amostras de mineração bruta, que são comuns a notícias falsas por causa da desambiguização que um texto curto pode ter. O treinamento contraditório projetado para aumentar a robustez do modelo por meio da adição pequenas perturbações aos dados de treinamento, mas também serve parar aumentar a generalização em última instância. Modelos tradicionais de aprendizado de máquina provaram que o aprendizado de conjuntos (*ensembles*) desempenha um papel importante na melhoria do efeito do modelo, como *Bagging* e *Boosting*. A principal razão está na característica de complementação entre os modelos que ajuda o modelo a fazer um julgamento correto sobre os resultados.

No trabalho de (WANI et al., 2021), os algoritmos de classificação são baseados em Redes Neurais Convolucionais (CNN), Memória de Longo Curto Prazo (LSTM) e Representações de Codificadores Bidirecionais de Transformers (BERT). Também foi avaliado a importância da aprendizagem não supervisionada na forma de pré-treinamento de modelo de linguagem e representações de palavras distribuídas usando *corpus* de tweets sobre COVID-19 não rotulados. As técnicas incluem modelos pré-treinados baseados em BERT e modelos brutos baseados na

CNN e LSTM. Os *Transformers* superaram modelos sequenciais anteriores em várias tarefas de PNL. O principal componente dos *Transformers* é a auto atenção, que é uma variante do mecanismo de atenção. Além disso, foram efetuados pré-processamento do conjunto de dados, remoção de *tags* HTML, conversão de caracteres acentuados em caracteres ASCII, expansão de contrações, remoção de caracteres especiais, remoção de ruído, normalização e remoção de *stopwords*.

A solução proposta em Koloski et al. (2021) emprega um conjunto de representação heterogênea, *ensemble* para a tarefa de classificação composta por múltiplas camadas ocultas. Houve um pré-processamento dos dados reduzindo o conteúdo do tweet e removendo as *hashtags*, pontuação e *stopwords*. A partir do texto limpo, foram gerados as *tags* POS usando a biblioteca NLTK, utilizando BoW e bigramas para palavras e BoW, bi-grama e tri-grama pata caracteres. O método proposto consiste em múltiplos sub métodos que visam abordar diferentes aspectos do problema. Por um lado, esse trabalho enfoca no aprendizado das características artesanais dos autores e, por outro lado enfoca em aprender a representação do espaço problemático com diferentes métodos. Algumas das características utilizadas são o comprimento máximo e mínimo de palavras baseados em palavras em um tweet, comprimento médio da palavra, desvio padrão do comprimento da palavra no tweet, número de palavras começando com letras maiúsculas ou minusculas. As características baseadas em caracteres consistiram nas contagens de dígitos, letras, espaços, pontuação, *hashtags* e cada vogal, respectivamente. Nesta abordagem é adotado uma representação densa com rede neural profunda de 5 camadas.

Utilizando o aprendizado de transferência na abordagem de Das et al. (2021), se mostrou extremamente eficaz em tarefas de classificação de texto, com um tempo de treinamento reduzido, pois não necessita de treinar cada modelo do início. Os principais passos dessa abordagem incluem inicialmente pré-processamento de texto, tokenização, previsão de modelos e criação de conjuntos usando um esquema de votação suave. Cada item de notícias precisa ser classificado em duas categorias distintas: "real"(verdadeiro) ou "falso". As principais partes desse trabalho consiste em: (a) Pré-processamento de texto, (b) Tokenização, (c) Arquiteturas de Modelos de *Backbone*, (d) *Ensemble* e (e) Processamento de Postes Heurísticos. Alguns itens de mídia social, como tweets, são escritos principalmente em linguagem coloquial. Também eles contêm várias outras informações como nomes de usuário, URLs, emojis, etc. É necessário filtrar tais atributos a partir dos dados dado como uma etapa básica de pré-processamento, antes alimentando-o no modelo de conjunto. Durante a tokenização, cada frase é dividida em tokens antes de ser alimentada em um modelo. Usa-se uma variedade de modelos de linguagem pré-treinados como modelos *backbone* para classificação de texto, com vetores de previsão de modelos dos diferentes modelos para obter o resultado de classificação, ou seja, "real"ou "falso". Para equilibrar as limitações de um modelo individual, um método de conjunto (*ensemble*) foi útil para uma coleção de modelos de bom desempenho. Nesta abordagem, aumenta-se a estrutura original com uma abordagem heurística que pode levar em consideração o efeito das alças de nome de usuário e domínios de URL presentes em alguns dados, como tweets.

No trabalho de Li e Li (2021) foram propostos dois modelos de detecção de notícias falsas: um é o modelo Text-RNN baseado no LSTM bidirecional, e o outro é *Text-Transformers*. A descrição dos dois modelos é a seguinte. No modelo TextRNN, usa-se o vetor de palavras GloVe como camada de incorporação com a dimensão de 200. Depois que o vetor de palavras codificadas passa pelo LSTM bidirecional, pega-se o estado oculto da última camada e o resultado final é produzido pela camada totalmente conectada. Foram usados cinco modelos de idiomas diferentes, incluindo Bert, Ernie, Roberta, XL-net e Electra treinados com a validação cruzada de cinco vezes.

Usando técnicas de PNL, o trabalho de Gupta et al. (2021) criou um modelo que faz uso de um detector de idiomas, juntamente com características extraídas por meio dos modelos hindi BERT e Hindi FastText e metadados. O conjunto de dados continha uma divisão de quase 50-50 tweets reais (verdadeiros) e falsos em treinamento, validação e teste. Com o uso de modelos baseados em regras com base na análise de dados, observou-se que o comprimento dos tweets, número de *hashtags*, número de menções etc. tiveram um impacto estatisticamente significativo nos rótulos das classes. Usando incorporações de palavras distribucionais como Word2Vec, FastText etc. teve enorme sucesso na classificação de texto. Além disso, a adição de informações baseadas em entidades usando BoW fez com que os dados tivessem presença significativa de entidades como nomes de órgãos governamentais, revistas científicas, locais, organizações, etc. e a distribuição variou de acordo com a classe. Usando *rankings* baseados em TF-IDF foi possível construir um vetor de recursos do BoW.

Segundo (GALHARDI et al., 2020), na época da escrita do artigo 301 casos da doença tinham sido confirmados no Brasil. Apesar das medidas recomendadas, a curva de casos e óbitos aumentou exponencialmente em níveis assustadores. Como resultado, uma circulação crescente de rumores sobre infecção produziu uma segunda pandemia, ou seja, produziu a disseminação de notícias falsas relacionadas ao COVID-19, cujas principais fontes eram as redes sociais. A análise de conteúdo de notícias falsas coletadas envolveu o período de 17 de março a 10 de abril de 2020. Esse tipo de análise pode ser aplicado a partir de um ponto de vista hermenêutico e quantitativo. As conclusões mostram que o WhatsApp é o principal canal de compartilhamento de notícias falsas, seguido pelo Instagram e Facebook. A OMS e o Fundo das Nações Unidas para a Infância (UNICEF) representam juntos 2% das instituições citadas como fonte de informações sobre cuidados e medidas contra o Novo Corona vírus em mensagens de WhatsApp. Algumas das informações compartilhadas são: A água fervida com alho serve como tratamento para corona vírus; O Corona vírus é mais extenso do que o normal, por isso qualquer máscara impede que ele entre no corpo; Quando cai sobre uma superfície metálica, o vírus permanece vivo por 12 horas e, no tecido, por nove horas. Portanto, lavar roupas ou colocá-las ao sol por duas horas elimina o vírus; O vírus vive nas mãos por 10 minutos; Morre o vírus exposto a uma temperatura acima de 26 graus; Como o vírus não resiste à temperatura acima de 26 graus, a água exposta ao sol pode ser consumida sem qualquer perigo; Evite comer sorvete ou pratos frios; Gargarear com água quente ou salgada impede que o vírus chegue aos pulmões; O álcool gel pode ser feito em casa

com apenas dois ingredientes. As notícias falsas também são identificadas nas mensagens de certas pessoas influentes para disseminar conteúdo que não é necessariamente falso, mas que visa capitalizar politicamente e economicamente. Alguns criminosos se aproveitam do medo e pânico da sociedade para espalhar *links* e apreender dados confidenciais das vítimas. O aplicativo "Eu Fiscalizo"notificou um golpe bancário. A mensagem SMS solicitou ao cliente que clicasse em um *link* para atualizar seus dados; caso contrário, sua conta seria bloqueada, que se traduz na tendência de muitos indivíduos em buscar informações que reafirmem suas próprias crenças, seja por meio de memórias seletivas ou leituras de fontes que estão ao seu lado. O psiquiatra brasileiro Cláudio Martins afirmou que pessoas que compartilham notícias falsas experimentam uma sensação de bem-estar de usuários de drogas. Martins argumenta que mecanismos imediatos de recompensa e prazer são desencadeadas pelo cérebro quando alguém recebe notícias que lhe agradam, o que leva as pessoas a transmitir compulsivamente as mesmas informações para que seu círculo de amigos sinta o mesmo.

Segundo (LINDEN et al., 2020), a desinformação sobre a pandemia de COVID-19 tem proliferado amplamente nas mídias sociais, desde o venda de falsas "curas", como gargarejar com limão ou água salgada e injetar-se com alvejante. Teorias conspiratórias de que o vírus foi feito em um laboratório na cidade de Wuhan ou que a rede celular 5G está causando ou exacerbando sintomas de COVID-19. As notícias falsas sobre o vírus também têm sido ativamente promovidas por elites políticas. Como a desinformação se espalha pelas redes como um vírus real "infectando seu hospedeiro"e transmitindo rapidamente falsidades de uma mente para outra, o antídoto natural é uma vacina psicológica contra notícias falsas desencadeiam respostas protetoras, como anticorpos. Em uma inoculação de persuasão, um forte desafio (por exemplo, uma teoria da conspiração) é enfraquecido a ponto de não mudar a posição da pessoa, o estado saudável da pessoa, mas desencadeará respostas protetoras, como o pensamento crítico aprimorado. Esta é a abordagem clássica da teoria da inoculação. No contexto do corona vírus, isso implicaria proteger as atitudes das pessoas que já estão seguindo as diretrizes de saúde pública. O fortalecimento de suas defesas atitudinais diminuirá a potência dos ataques de desinformação. No entanto, uma abordagem mais recente dentro da teoria da inoculação expande sua eficácia para incluir também uma aplicação terapêutica — tratamentos de inoculação que visam um estado insalubre.

Segundo (APUKE; OMAR, 2021), pesquisas sobre a proliferação de notícias falsas estão surgindo na era da pandemia COVID-19. Alguns estudos têm tentado perceber a conexão entre as mídias sociais e a desinformação nesta era da pandemia. Pesquisas recentes mostraram que, nos últimos meses, o compartilhamento mais notável de notícias falsas para a saúde tem sido no COVID-19.

Já para (JúNIOR et al., 2020), como mensagens falsas são espalhadas em diversos formatos, geralmente possuem um texto afirmativo, o que leva as pessoas, que não verificam a veracidade das informações, a acreditarem e a compartilharem a notícia falsa. Como mensagens falsas relacionadas ao COVID-19 estão espalhando desinformação e medo, o que acaba

atrapalhando o trabalho dos órgãos envolvidos na contenção novo vírus. Na página criada pelo Ministério da Saúde, dedicada ao esclarecimento de conteúdos falsos compartilhados nas redes sociais, no dia 20 de março de 2020 já se contabilizava 58 títulos de informações não verídicas, entre elas "utilizar álcool em gel nas mãos para prevenir corona vírus altera bafômetro nas blitz", "chá de abacate com hortelã previne corona vírus"e "uísque e mel contra corona". Para auxiliar no levantamento das notícias sobre a saúde que circulam nas mídias sociais, o Ministério da Saúde disponibilizou um número de WhatsApp para que a população informe as mensagens para conferência da veracidade pelo órgão. A preocupação do ministério vai além de dizer que se trata de uma notícia falsa, prezando também pela disseminação de informações corretas, pois a finalidade é proporcionar mais conhecimento para a população, incentivando a educação para a saúde.

Em 2018, o Ministério da Saúde brasileira criou um espaço em um sítio eletrônico e nas redes sociais visando combater *Fake News*, e se propôs a esclarecer os fatos com base nas evidências científicas e suas fontes (NETO et al., 2020). Isto foi necessário em virtude de um parecer que apontou que aplicativos de trocas de mensagens dificultavam a população de se proteger de doenças, tais como febre amarela, gripe e sarampo. A busca das *Fake News* ocorreu no banco de dados do Ministério da Saúde, no cenário da pandemia de COVID-19, no período de 29 de janeiro a 31 de março de 2020, quando foram identificados 70 registros. Foram originadas cinco categorias: informações relacionadas aos discursos de autoridades na saúde (40), terapêutica (17), medidas de prevenção (nove), prognósticos da doença (dois) e vacinação (dois). Os achados apontam para quatro tipificações de registros com a temática da COVID-19, trazendo informações relacionadas aos discursos de autoridades na saúde, medidas de prevenção, prognósticos da doença, terapêutica e sobre a vacinação. Ademais, destaca-se que a literatura brasileira é escassa sobre a pandemia de COVID-19 e a velocidade desta produção do conhecimento vai de encontro com a produção das *Fake News*.

(FERREIRA et al., 2020) afirma que atualmente as notícias falsas apresentam-se como instrumento de manipulação em massa, muitas missões para obter vantagens em conflitos sociais, políticos e econômicos, principalmente, em situações de fragilidade a capacidade humana de discernimento. "Embora não seja um fenômeno novo, a desinformação e a manipulação pelo meio de notícias falsas são bem grandes hoje em dia, sendo um aspecto emergente da revolução da mídia e da informática em que vivemos."Ao analisar a disseminação de informações falsas sobre COVID-19 em plataformas de mídia social, reconhecem que esses acessos direto a uma quantidade sem precedentes de conteúdos que podem ampliar rumores e informações questionáveis. É importante, pois, que se entender a dinâmica social por trás do consumo de conteúdo e da mídia social, dede que isso pode ajudar os modelos mais eficientes e responsáveis de comportamento social, e a estratégia para implementar um meio de comunicação mais eficaz em tempos de crise. Como mudanças nos processos info-comunicacionais promovidos pelo avanço tecnológico contribuíram também para a propagação de notícias falsas em massa, que promovem a infodemia, principalmente, nos canais eletrônicos de comunicação, revelando-se como um

problema social, uma vez que a informação falsa ou manipulada interfere no comportamento direto dos sujeitos. Também foi destacado que as pessoas que deixam compartilhar, por não ter certeza da veracidade da informação e para não contribuir com a disseminação de *Fake News*, pode ocultar algo de grande utilidade pública. Assim, diante do caos social, parte da sociedade pode deparar com tal dilema, onde a decisão de compartilhar ou não uma informação se torna tão difícil quanto acreditar ou não nela.

No entanto, as mídias sociais também estão repletas de desinformação sobre saúde. A desinformação sobre a saúde foi documentada em quase todas as plataformas de mídia social, incluindo Facebook, Twitter, YouTube, Pinterest e Instagram. Além disso, a desinformação em saúde não se limita a nenhuma questão, mas inclui desinformação de vacinação (geralmente e casos específicos, como HPV ou gripe), e desinformação sobre crises globais de saúde, como o surto de Ebola em 2014, e a disseminação do Zika em 2016. (SINGH et al., 2020) analisa a quantidade de conversas que ocorrem nas mídias sociais, especificamente no Twitter, no que diz respeito ao COVID-19, aos temas de discussão, de onde a discussão está emergindo e quanto dela está conectada a outras informações de alta e baixa qualidade na *Internet* por meio de *links* compartilhados de URL. Usando a API de *streaming* do Twitter foram coletados tweets relacionados ao COVID-19 em 16 de janeiro de 2020. Os dados apresentados no estudo são de 16 de janeiro de 2020 a 15 de março de 2020. Aproximadamente 80% dos tweets foram rotulados com um ou mais temas. Inicialmente foram identificados dez mitos que vão desde remédios caseiros, teorias conspiratórias sobre a origem do vírus e desinformação sobre o clima quente matando o vírus. Enfocando-se no número de tweets originalmente contendo um ou mais desses mitos citados, foi descoberto que aproximadamente 16.000 tweets (pouco menos de 0,6% dos tweets) estão discutindo um ou mais dos mitos considerados. Na análise de tweets relacionados á pandemia de COVID-19 mostra que 40,5% do conteúdo original do tweet (5,1% do conteúdo de retweet e 9,6% do conteúdo geral) inclui uma URL.

Segundo (ROVETTA; BHAGAVATHULA, 2020), a ferramenta *Google Trends* fornece *insights* em tempo real sobre comportamentos de pesquisa na *Internet* sobre vários tópicos, incluindo COVID-19. Plataformas de mídia social, tais como Facebook, Twitter e Instagram, permitem que os usuários comuniquem seus pensamentos, sentimentos e opiniões compartilhando mensagens curtas. Um aspecto único dos dados de mídia social do Instagram é que as postagens baseadas em imagens são acessíveis, e o uso de *hashtags* relacionadas a tópicos permite o acesso a informações relacionadas a tópicos para os usuários. Os seis principais termos relacionados ao COVID-19 pesquisados no Google foram "coronavírus", "corona", "COVID", "vírus", "coronavírus"e "COVID-19". Países com maior número de casos de COVID-19 apresentaram maior número de consultas COVID-19 no Google. Os apelidos "coronavirus ozone", "coronavirus laboratory", "coronavirus 5G", "coronavirus conspiracy"e "coronavirus bill gates"foram amplamente divulgados na *Internet*. As pesquisas sobre "dicas e curas"para o COVID-19 aumentaram em relação ao presidente dos EUA especulando sobre uma "cura milagrosa"e sugerindo uma injeção de desinfetante para tratar o vírus. Cerca de dois terços dos usuários do Instagram usaram as

*hashtags* "COVID-19"e "coronavirus"para dispersar informações relacionadas ao vírus. Os dados coletados incluíam conteúdos publicados no Instagram e informações demográficas de usuários. Não foram coletadas informações pessoais, como *e-mails*, números de telefone ou endereços. Os dados das *hashtags* do Instagram foram coletadas manualmente por meio de *tags* sugeridas pelo Instagram associadas a países específicos. Usando as *hashtags* do *Google Trends* e do Instagram, o estudo identificou que houve um crescente interesse no assunto COVID-19 globalmente e em países com maior incidência do vírus. Pesquisas relacionadas a "notícias COVID-19"são bastante frequentes e dois terços dos usuários do Instagram usaram "COVID-19"e "coronavirus"como *hashtags* para dispersar informações relacionadas ao vírus. Vários apelidos infodinâmicos estão circulando na *Internet*, com "conspiração coronavírus"e "laboratório coronavírus"identificados como os mais perigosos.

(INUWA-DUTSE, 2021) destaca que uma quantidade maciça de dados pode ser facilmente obtida de plataformas como o Twitter. Tweets, geralmente trechos de texto curtos, referem-se ao fluxo de postagens que os usuários compartilham no Twitter e permitem estudos longitudinais. Um objeto de tweet é uma estrutura de dados complexa, expressa no formato JSON, que descrevem informações específicas sobre o tweet e o titular da conta (o usuário). Como um texto marcado, os diferentes campos no objeto do tweet definem características importantes do tweet. A complexidade de um tweet e sua natureza não estruturada dificulta o processamento diretamente em uma forma utilizável, o que requer uma série de pré processamento antes que uma análise efetiva possa ser conduzida. O fluxo de tweets difere dos textos convencionais de fluxo em termos de taxa de postagem, dinamismo e flexibilidade; eles são gerados a uma velocidade rápida e tendem a ser altamente dinâmicos. Uma das razões pelas quais as plataformas de mídia social *online* são muito populares entre o público tem a ver com a capacidade dos usuários de gerar e consumir conteúdo simultaneamente levando a várias formas de informação - modismos, opiniões, notícias de última hora. Essa razão também contribui para o aumento do número de postagens sem censura sobre diversos fenômenos sociais, em parte devido ao seu curto tamanho e à velocidade de comunicação. A demanda por serviços *online* está no seu auge durante o bloqueio, expondo assim a população a várias vulnerabilidades. Entre as repercussões do aumento do volume de informações (relevante e irrelevante) sobre a pandemia é a tendência de criar um sentimento de perplexidade por parte do público sobre quais medidas preventivas tomar e qual informação acreditar. Como tal, é crucial entender como o conteúdo enganoso *online* se propaga e estudar como otimizar métodos que favoreçam o domínio de conteúdo relevante sobre os irrelevantes.

Ainda segundo (INUWA-DUTSE, 2021), existem várias fontes de desinformação e conspiração capazes de enganar o público sobre a pandemia COVID-19. Apesar das medidas tomadas pelas plataformas de mídia social para reduzir conteúdo irrelevante, muitas fontes de informações enganosas e rumores ainda existem. Um repositório abrangente de conjuntos de dados validados e espúrios sobre pandemia facilitará a autenticação da veracidade de uma determinada informação sobre o assunto. A análise de redes sociais é útil para revelar o dinamismo de muitas formas de

relações sociais em vários níveis. No domínio da ciência social, a sociometria é um meio de medir ou estudar as relações sociais entre as pessoas. Geralmente, as redes são caracterizadas por um certo grau de organizações em que grupos de nós formam unidades fortemente conectadas como comunidades. As comunidades representam entidades funcionais que refletem as relações topológicas entre elementos da rede subjacente. Observando o nível de resistência e aceitação em relação ao COVID-19, um agrupamento de alto nível de usuários pode potencialmente revelar a distribuição dos usuários para razões relacionadas à gestão, logística e contenção do surto. Outro problema útil a ser enfrentado é entender a percepção dos usuários sobre medidas tomada na gestão da pandemia. Por exemplo, será possível avaliar a política de bloqueio para entender a disposição dos usuários em cumprir e a mudança atitudinal ao longo do tempo.

## 3.4 Conjuntos de Dados para Conteúdos Falsos

Nesta seção são sumarizados os trabalhos listados na Tabela 9.

(PATWA et al., 2020) afirma que normalmente os itens que mais chamam a atenção em notícias são o título da reportagem e a imagem de capa da reportagem. Vale ressaltar que uma notícia é geralmente feita por meio de três elementos: manchete, multimídia (imagem, vídeo, áudio etc.) e corpo (história real - conteúdo de texto), sendo os dois primeiros elementos sendo mais proeminentes e eficazes do que o terceiro. Um estudo mostra que 70% dos usuários do Facebook só leem a manchete de histórias científicas antes de comentar ou compartilhar. A detecção automática de notícias falsas não é um problema fácil de resolver, uma vez que hoje em dia um artigo de notícias geralmente compreende imagens e vídeos (em comparação apenas com texto), o que é fácil de falsificar. Outros elementos que dificultam o rastreamento de notícias falsas é a fonte da notícia que pode ser de sites anônimos em formatos de publicação como sites de notícias populares, ou *blogs*, ou até mesmo mídias sociais (redes sociais, *e-mails* e podcasts). Alguns elementos e técnicas de processamento de texto podem ser usadas para extrair características, por exemplo: *N-Grams*, Pontuação, Características psicolinguísticas, Legibilidade e Sintaxe.

Tacchini et al. (2017) trabalharam com um conjunto de postagens públicas selecionadas do Facebook durante o segundo semestre de 2016 com ajuda do API da própria rede social. Foram criados posteriormente dois grupos, um grupo com fontes de notícias científicas e outro grupo com fontes de notícias conspiratórias. O Objetivo foi fazer uma classificação binária classificando as postagens como boatos(falsas) ou verdadeiras. O efeito da câmara de eco causou influência porque, pela análise de compartilhamento de mídia social, os usuários tendem a se agrupar em comunidades de interesse, o que causa reforço e promove viés de confirmação, segregação e polarização. Duas estratégias foram usadas para a classificação de cada postagem. A primeira estratégia foi usar a classificação por regressão logística e a outra estratégia foi usar classificação *booleana* de *crowdsourcing* harmônica. A primeira classificação é bem adequada a

**Tabela 9 – Trabalhos sobre conjunto de dados**

| Referência | Título | Foco | Idioma | Tamanho |
|---|---|---|---|---|
| Patwa et al. (2020) | Fighting an Infodemic: COVID-19 Fake News Dataset | Twitter | Inglês | 5.600 verdadeiros, 5.100 falsos |
| Tacchini et al. (2017) | Some Like it Hoax:Automated Fake News Detection in Social Networks | Facebook | Inglês | 15.500 postagens |
| Wang (2017) | Liar & Liar Pants on Fire: A New Benchmark Dataset for Fake News Detection | Declarações | Inglês | 12.846 declarações |
| Santia e Williams (2018) | BuzzFace: A News Veracity Dataset with Facebook User Commentary and Egos | Facebook | Inglês | 2,282 artigos de notícias compartilhados |
| Shu et al. (2019b) | FakeNewsNet: A Data Repository with News Content, Social Context and Spatialtemporal Information for Studying Fake News on Social Media | Noticias | Inglês | 1.056 (PolitiFact), 22.140 (GossipCop) |
| Monteiro et al. (2018) | Contributions to the study of fake news in portuguese: New corpus and automatic detection results | Notícia | Português | 3.600 verdadeiras, 3600 falsas |
| Zarei et al. (2020) | A First Instagram Dataset on COVID-19 | Instagram | Inglês e outros idiomas | 5.300 postagens |

problemas com um conjunto muito grande e uniforme de recursos que é o caso da base do estudo, no entanto há problemas entre usuários que gostaram de postagens que constam em categorias diferentes. Na segunda classificação os usuários fornecem os rótulos verdadeiro ou falso para as postagens, indicando se uma postagem é perigosa ou se viola as diretrizes da comunidade, o problema desta abordagem é que ela consiste em calcular os rótulos de consenso de entrada do usuário e poucas vezes há um consenso total entre todos os avaliadores. Após testagens e experimentos a segunda abordagem se mostrou melhor porque é computacionalmente mais eficiente, lidou com grandes conjuntos de dados e ofereceu boa precisão na prática, ambos os algoritmos forneceram bom desempenho, sendo que o algoritmo harmônico apresentou precisão de 99%. Os algoritmos podem ser dimensionados para o tamanho de redes sociais inteiras, exigindo apenas uma quantidade modesta de classificação manual.

O conjunto de dados de Wang (2017) inclui 12,8 mil declarações curtas de políticos e personalidades americanas manualmente extraídas da API do site Politifact, e cada declaração é avaliada por um editor dessa agência de verificação, atestando o nível de veracidade. As declarações são rotuladas como extremamente enganoso, falso, pouco verdadeiro, meio verdadeiro, quase verdadeiro e verdadeiro. A distribuição das classificações no conjunto de dados é relativamente bem equilibrada são 1.050 casos para declarações extremamente enganosas, e um volume variante de 2.063 a 2.638 para as demais classificações. Foram incluídos um rico conjunto de metadados para cada uma das declarações como filiações partidárias, trabalho atual, estado de origem e o histórico de credibilidade da pessoa em análise. As declarações foram produzidas em vários contextos e as principais categorias incluem comunicados à imprensa, entrevistas na TV ou rádio, discursos de campanha, anúncios na TV, tweets, debates e postagens no Facebook.

O conjunto de dados fornecido por Santia e Williams (2018) consiste em 2.282 artigos de notícias, juntamente com vários recursos do Facebook e a classificação de veracidade verificada e atribuída para cada artigo. Os artigos incluem todas as postagens num prazo de 7 dias durante o mês de setembro de 2016, tais postagens foram feitas nas páginas de notícias do Facebook da ABC News Politics, Addicting Info, CNN Politics, Eagle Rising, Freedom Rising, Freedom Daily, Occupy Democrats, Politico, Right Wing News e The Other 98%. Este período de tempo foi o auge das eleições presidenciais de 2016, uma das mais ricas fontes de produção de notícias falsas. As classificações dos artigos foram de quase verdadeiro, quase falso, misturado, verdadeiro e falso e sem conteúdo fatual. Outras informações foram acrescentadas posteriormente de outras redes sociais como o Twitter e o Reddit, as ligações foram feitas por meio de postagens que coincidem com a mesma mensagem do Facebook.

O Fake.Br é um corpus criado por Monteiro et al. (2018) com uma coleção de 7.200 notícias divididas homogeneamente em notícias reais e notícias falsas em um intervalo de janeiro de 2016 a janeiro 2018. As fontes de notícias falsas foram os sites de notícias A Folha do Brasil[3], Diário do Brasil[4], The Jornal Brasil e Top Five TV. As fontes de notícias verdadeiras foram G1 [5], Folha de São Paulo[6] e Estadão[7]. As pesquisas foram feitas por palavras chaves encontradas nas notícias falsas. Certificando-se da relação que a notícia verdadeira negou a notícia falsa de mesmo tópico. As técnicas utilizadas são de *stopwords*, remoção de pontuações, *bag of words* e o número normalizado de ocorrências de classes semânticas como indicado pelo *Linguistic Inquiry and Word Count* (LIWC) para língua portuguesa e como método se utiliza *Support Vector Machine* (SVM) para reconhecimento de padrões. Algumas características encontradas foram erros gramaticais em notícias falsas possui uma frequência maior, existe pausa frequentes em orações que foram computada pelo número de pontuações sobre o número de sentenças, a

[3] https://folhadobrasil.com/
[4] https://www.diariodobrasil.org/
[5] https://g1.globo.com/
[6] https://www.folha.uol.com.br/
[7] https://www.estadao.com.br/

emoção que pode ser indicada pela linguagem expressa na mensagem, a incerteza, mensuradas por número de ocorrências verbais ocorridas na voz passiva, e a quantidade de pronomes de primeiro e do segundo grau.

FakeNewsNet é um conjunto de dados que compila as informações de conteúdo das notícias e o contexto social majoritariamente abordados em outros estudos, porém este conjunto de Shu et al. (2019b) contribui adicionando um novo grupo de informações espaço-temporal. A informação espaço-temporal inclui informação espacial, no caso a localização e o *timestamp* de todas as informações coletadas dos sites PolitiFact e GossipCop[8]. fontes para características de conteúdo das notícias e a fonte para características de contexto social foi a rede social Twitter. Essa é uma das poucas bases de dados de notícias falsas que possui mais de um domínio que trabalham em uma mesma base, os domínios das notícias são política e celebridades.

Os dados do *dataset* descrito em (ZAREI et al., 2020) foram reunidos entre 5 de janeiro e 30 de março de 2020. O conjunto de dados abrange 18.500 comentários e curtidas de 329.000 em postagens de 5.300. Esses *posts* foram distribuídos por 2.500 publicadores (usuários que publicaram a postagem). As publicações coletadas são públicas do Instagram, rastreando todas as postagens associadas a um conjunto de *hashtags* COVID-19 apresentadas. Para poder coletar conteúdo público do Instagram (em forma de *post*), foi usada as API oficial do Instagram. A primeira versão desse processo de coleta de dados teve início em 5 de janeiro de 2020 e terminou em 30 de março de 2020. Para manter os dados organizados, o conjunto de dados foi dividido em quatro partes: conteúdo da postagem, informações do editor, métricas de comentários e recursos. As postagens contêm atributos-chave, como legenda, lista de *hashtags*, imagem/vídeo, número de curtidas, número de comentários, localização e data.

## 3.5 Considerações Finais

Com a conclusão desse capítulo, foi possível identificar os trabalhos mais relevantes ao assunto desta pesquisa de Mestrado sobre detecção de notícias falsas em todas as suas vertentes, sejam elas a produção de conjunto de dados para realizar *benchmarks*, produção de modelos ou análise de redes sociais. Foi possível observar as limitações de cada trabalho e das linhas de pesquisas realizadas até hoje, uma delas são pesquisas voltadas para a rede social Instagram.

Os trabalhos com maior impacto e que abordam também as redes sociais são as pesquisas de Bagade et al. (2020) e Atodiresei et al. (2018), entretanto a plataforma que esses artigos é o Twitter. O trabalho de ATODIRESEI et al. aborda também características das imagens que são postadas nos *tweets*, para fins de detecção de imagens semelhantes e não características das imagens. Sobre estudos de trabalhos com aprendizado supervisionado Reis et al. (2019b) mostra como resultados de pesquisa que os classificadores Random Forest e XG Boosting possuem um bom desempenho em tarefas de classificação de texto. Shu et al. (2019b) é um dos poucos autores

[8] https://www.gossipcop.com/

que não construiu apenas uma fonte de dados rotulados, algo que é comum apenas se tratar sobre política em pesquisas de informações falsas, na verdade o trabalho aborda dois domínios o de política e o de celebridade o que rompe a barreira de apenas se concentrar em um único assunto. Outro conjunto de dado importante para esta pesquisa foi Zarei et al. (2020), por extrair postagens da rede social Instagram a respeito de COVID-19, dois tópicos que também são o foco desse trabalho, porém o trabalho não foca na classificação de texto, não tem também a apuração dos dados textuais que são compartilhados nas imagens, e neste trabalho de mestrado são ambos analisados e implementados.

Não existem trabalhos até a publicação desse projeto que contam com a mesma linha de pesquisa, como já mencionado no primeiro capítulo deste trabalho. A plataforma Instagram não apresenta as mesmas características que outra plataforma de rede social como o Facebook, logo não se torna viável uma comparação com trabalhos dessa plataforma, pois mecanismos de compartilhamento não existe na plataforma do Instagram, o engajamento em ambas as plataformas também funciona de formas diferentes. No Facebook o sistema para engajamento é de círculo de amigos, uma vez que uma pessoa adiciona a outra, ambos podem compartilhar e acompanhar as postagens um do outro, já no Instagram ele trabalha com um sistema de seguidores, onde cada pessoa individualmente tem acesso apenas ao que segue ou ao que está disponível em arquivos públicos. Essas são apenas algumas das características que diferenciam as plataformas. Também não existem pesquisas que trabalhem com compartilhamento de informações falsas ou verdadeiras no Instagram para analises comparativas dos modelos de aprendizado de máquina utilizados e a diferença do desempenho dos mesmos.

O próximo capítulo descreve as investigações dessa pesquisa de Mestrado sobre postagens falsas relacionadas à COVID-19 na rede social Instagram.

# Capítulo 4

# DETECÇÃO DE POSTAGENS FALSAS SOBRE COVID-19

## 4.1 Considerações Iniciais

O objetivo desta pesquisa de Mestrado foi investigar a detecção de postagens com informações falsas no idioma português sobre a pandemia do Covid-19 especificamente aplicada na rede social Instagram.

Nesse sentido, este Capítulo primeiramente detalha a proposta do projeto de Mestrado. Na seção 4.2 é detalhada a contextualização em frente aos objetivos propostos na seção 1.4. Na seção 4.3 é descrita a metodologia que foi utilizada no desenvolvimento das atividades deste projeto de pesquisa. Além disso, na seção 4.4.1 deste capítulo são explicados os modelos de aprendizado de máquina produzidos para detectar as postagens falsas. As postagens usadas nos experimentos abrangem o período de janeiro a agosto de 2020, período intenso da Pandemia no Brasil em que ainda existiam muitas informações incertas e desinformações sendo compartilhadas.

Foram extraídos das redes sociais um total de 17.974 postagens em português com a *hashtag* vírus chinês entre janeiro a setembro de 2020, e separadamente foram extraídos e filtrados as seguintes quantidades de postagens das páginas jornalísticas e de verificadores de conteúdo que continham informações a respeito do corona vírus : Agencia Lupa (@agencia_lupa) 665 postagens, Aos Fatos (@aosfatos) 245postagens, CNN Brasil (@cnnbrasil) 339 postagens, Covid Verificado (@covidverificado) 97 postagens, Estadão (@estadao) 785 postagens, Folha de São Paulo (@folhadespaulo) 1246 postagens, Portal G1 (@portalg1) 403 postagens, Projeto Comprova (@comprova) 130 postagens, Istoé (@revistaistoe) 174 postagens, Uol Notícias (@uolnoticias) 633 postagens e revista Veja (@vejanoinsta) 540 postagens.

Para os experimentos foram extraídas 2.190 imagens que continham textos, metade se tratam de informações verdadeiras de páginas de jornais populares como Uol, Estadão, Veja, Folha , CNN, G1, Istoé e a outra metade contém informações falsas e se tratam de postagens feitas por usuários e em ambos, sendo que os dados foram retirados da plataforma Instagram,

que por meio de uma checagem manual em páginas de jornais e verificadores de conteúdo foram classificadas como falsas.As informações verificadas nas postagens foram o texto contido na imagem e na legenda, foi considerado como falso qualquer uma das postagens que tinham pelo menos uma frase contendo informação falsa sendo compartilhada.

## 4.2 Contextualização

O processo de comunicação vem acelerando faz muitos anos e com o advento das redes sociais esse processo começou a se tornar muito rápido e maciço. Se houver a necessidade de enviar um recado ou compartilhar uma informação com familiares, amigos e colegas de trabalho por meio das redes sociais, essa tarefa pode ser realizada com poucos cliques. Com a facilidade de compartilhar a informação e atingir de forma rápida um grupo específico, as redes sociais começaram a ser utilizadas como fontes de informação. Os jornais estão migrando todas as suas mídias sociais para a *Internet* e investindo massivamente em redes sociais para acompanhar esta evolução tecnológica. Com a alta capacidade de dispersão dentro de uma rede social existe o problema de compartilhar notícias falsas e mesmo que com muitos esforços as notícias verdadeiras que divergem com as falsas, as notícias verdadeiras acabam não tendo muito alcance.

Há muitos estudos sobre o tema de notícias falsas em diversas áreas de pesquisa acadêmica. Normalmente as áreas que trabalham e buscam entender como são feitas, as motivações e como combatê-las são as Ciências Sociais, Psicologia, Ciência da Informação e Ciência da Computação. Existen outras áreas devem começar a se debruçar sobre o tema como diz Lazer et al. (2018), pois há notícias falsas em vários tópicos como a eficácia das vacinas, nutrição ou mercado de valores, e não apenas sobre política. Segundo Acerbi (2019), as redes sociais vêm desempenhando um péssimo trabalho no combate as notícias falsas, seja pela alta complexidade de se filtrar os dados, ou pela simples falta de controle sobre a grande quantidade de dados que são produzidas nessas redes todos os dias.

Nos trabalhos relacionados que foram sintetizados no Capitulo 3 foram abordados diversos pontos de importância para se estudar sobre o fenômeno das notícias falsas dentro das redes sociais. Outro ponto importante é a falta de estudos sobre o comportamento e a difusão de notícias falsas na rede social Instagram, uma rede que de acordo com Barrett (2019) é uma das redes mais negligenciadas e que pode talvez ter impactado muito mais que as demais redes sociais tradicionais, como Facebook, Twitter e WhatsApp.

## 4.3 Metodologia

Nesta seção é discutida a proposta da arquitetura para se detectar postagens falsas sobre Covid-19, além da forma que será trabalhado o reconhecimento de caracteres e as principais

técnicas de PLN que serão utilizadas na implementação do projeto.

### 4.3.1 Proposta de Arquitetura

Na Figura 5 está desenhado a configuração da arquitetura do *framework* que foi implementado. Cada módulo é discutido a seguir:

- Agência de Verificação: Este módulo corresponde as postagens das páginas do Instagram de sites de verificação Aos fatos e Lupa. Todas as postagens, a respeito do corona vírus foram buscadas e selecionadas por palavras chaves como COVID-19 e corona vírus dentro da legenda dessas postagens. A busca foi limitada para artigos de dezembro de 2019 até julho de 2020.
- Base de Conhecimento: para poder utilizar as notícias em forma computacional, por isso os dados foram modulados (tratados e estruturados) e inseridos em uma base de dados unificada que possui apenas notícias verdadeiras.
- PLN: Este módulo tem como objetivo extrair as informações textuais de notícias verdadeiras e as notícias que serão classificadas, com análises morfológicas, sintáticas e semânticas por meio de técnicas como a classificação de reconhecimento de entidades, etiquetagem e tokenização.
- Banco de Dados *NoSQL* Orientado a Documentos: A parte textual que foi extraída pelo módulo do Instagram Scraper foi inserida nesse banco para permitir a realização das análises e detecção das notícias postadas na rede social.
- Sistema OCR: As imagens adquiridas e inseridas no Banco de Dados *NoSQL* Orientado a Chaves-Valor foram processadas neste módulo para a extração de possíveis textos contidos nas imagens. Após a detecção das partes do texto dentro da imagem, as partes são enviadas e associadas ao texto que pertence a imagem dentro do Banco de Dados *NoSQL* Orientado a Documentos.
- VirusChines: Por meio de uma API foram extraídas postagens que possuem a *hashtag viruschines* e armazenadas separadamente para futuros processamentos. O uso da *hashtags* foi proposital para medir o nível de informações falsas compartilhadas dentro de uma *hashtag* que foi usada de forma pejorativa por um especifico grupo de usuários.

### 4.3.2 Sistema de Reconhecimento de Caracteres

As atividades desenvolvidas compreenderam o uso de um sistema de reconhecimento de caracteres e de detecção de notícias falsas na rede social Instagram. Para a utilização de um sistema de reconhecimento de caractere voltado para redes sociais existem poucos trabalhos que

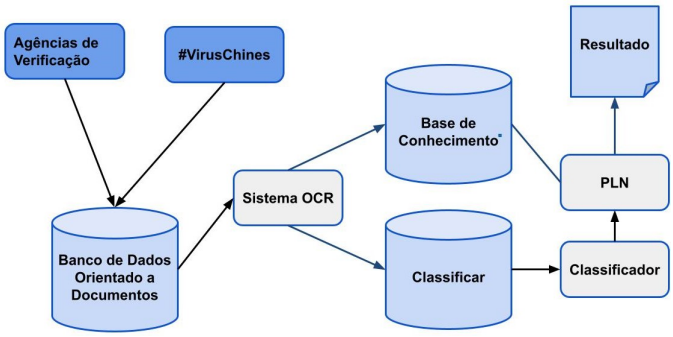


**Figura 5 – Arquitetura de Sistema para a Detecção de Notícias Falsas.**

**Fonte: O próprio autor**

implementem uma solução *open source* com resultados estatísticos sobre o desempenho. Um trabalho mais próximo a proposta de reconhecimento de caractere é o Rosetta (BORISYUK et al., 2018), um sistema desenvolvido pelo Facebook para a detecção de texto em imagens que são transferidas para a plataforma. Porém, por se tratar de um sistema proprietário não existem estudos independentes sobre o desempenho do sistema e não está disponível o detalhamento do processo de detecção de notícias falsas no final do processo, por enquanto a plataforma usa o sistema de sinalização e a verificação é realizada por empresas terceirizadas.

Para esta pesquisa foi escolhido um sistema *open source* de OCR para o experimento de reconhecimento de imagens postadas no Instagram chamado Tesseract e que foi descrito no Capítulo de Fundamentação Teórica.

### 4.3.3 Processamento de Linguagem Natural e Aprendizado de Máquina

Para a etapa de PLN foi investigada a possibilidade de uso do NTLK, uma biblioteca em Python que ajuda em técnicas de tokenizaçao, etiquetagem e uso de *bag of words*. Mesmo sendo técnicas muito utilizadas, não foram usadas para análises de padrões na rede social Instagram. A plataforma além de possuir imagens que podem conter texto, tem um limite de caracteres por legenda de 2.200 caracteres, podendo ser incrementado com uso de mais 30 *hasgtags*. Os comentários e as *hashtags* são muito utilizados para expressar mensagens positivas ou negativas em interações com a imagem. O compartilhamento funciona diferente nessa rede social, para tal função é necessário um aplicativo externo que realiza este processo. Portanto não existe um problema com cascatas, mas existe o problema do algoritmo de filtro bolha que aumenta a capacidade de alcance de uma imagem de acordo com o gosto do usuário.

## 4.4 Características das Postagens

Os modelos de detecção de postagens falsas usando aprendizado de máquina foram divididos em 3 grupos: a) textos encontrados nas imagens e nas legendas das postagens, b) características das imagens e C) os metadados da postagem. Cada subseção a seguir detalha quais dados foram usados, como foram usados e os resultados dos modelos para cada grupo, mostrando quais características retornaram os melhores resultados junto com os classificadores. Nem sempre as postagens contavam apenas com uma única foto, em alguns casos se tratava de um álbum contendo apenas fotos ou fotos com vídeos. Como explicado na seção 1.4, esse trabalho descartou os vídeos, pois não faz parte do escopo deste trabalho. A Tabela 10 contém todas as 31 características que foram extraídas para serem utilizadas nos modelos. As características com códigos de F1 à F8 são características do grupo de imagem, os códigos do F9 até o F20 são características do grupo de metadados e as demais características são dos grupos de texto (legenda e texto na imagem). Todas as características das postagens foram divididas em três grupos: os textos que contém a informação sendo compartilhada e está por sua vez foi classificada como falsa e verdadeira , as características da imagem e dos metadados da postagem que seguem a classificação que foi dada pelos textos, coincidentemente todas as postagens com informações verdadeiras contavam com legendas e com texto contido na imagem com a mesma classificação e o mesmo ocorreu com o outro grupo. As características mencionadas na tabela XX foram combinados entre seus próprios grupos, por exemplo as 8 características da imagem foram combinadas de todas as formas possíveis entre elas: com todas as características, com características individuais e assim por diante. Os demais grupos de características também seguiram o mesmo processo.

Nas Figuras18 e 19 são compartilhadas nos Apêndices A e B respectivamente. São mostrados 4 imagens compartilhadas por jornais e 4 compartilhadas por usuários comuns e por checagem manual as imagens contidas na Figura 19 no Apêndice B as informações são falsas. Mesmo se tratando de imagens publicadas em postagens públicas, por questões éticas as informações pessoais disponibilizadas nas imagens foram censuradas.

### 4.4.1 Características das Imagens

Para extrair algumas características das imagens para serem utilizadas nos algoritmos de aprendizado de máquina, todas essas extrações foram feitas por bibliotecas de Python que computavam em formato de vetor de números binários, sendo que as características são explicadas a seguir.

- Formatos: Existem várias formas de se extrair formatos de imagens capturando momentos da imagem por meio de uma média ponderada das intensidades dos *pixels* da imagem. Em outras palavras, todas as intensidades de *pixel* são ponderadas apenas com base em sua intensidade, mas não com base em sua localização na imagem. Para uma imagem binária,

**Tabela 10 – Códigos para as tabelas de experimento**

| Código | Descrição | Código | Descrição |
|---|---|---|---|
| F1 | Hu moments | F17 | Data (Minuto da postagem de acordo com o fuso horario de Brasilia) |
| F2 | Haralick | F18 | Data (Segundo da postagem de acordo com o fuso horario de Brasilia) |
| F3 | Histogram | F19 | Data (Dia da semana, exemplo: Segunda-Feira) |
| F4 | Threshold Adjacency Statistics (TAS) | F20 | Hashtags |
| F5 | Histogram of Oriented Gradients (HOG) | F21 | Vector TfidfVectorizer |
| F6 | Zenirke | F22 | Vector CountVectorizer |
| F7 | Local Binary Patterns (LBP) | F23 | Vector HashingVectorizer |
| F8 | Tamanho (pixels: altura x largura) | F24 | n-grams em caracteres |
| F9 | ID do criador da postagem na plataforma | F25 | n-grams em palavras |
| F10 | Quantidade de likes | F26 | n-grams em caracteres apenas nos limites da palavra |
| F11 | Quantidade de comentários | F27 | normal |
| F12 | Data (Ano da postagem) | F28 | Steming |
| F13 | Data (Mês da postagem) | F29 | Normalização (palavras em minúsculo) |
| F14 | Data (Semana da Postagem de acordo com o mês) | F30 | Ngram |
| F15 | Data (Dia do Mês) | F31 | Stopword |
| F16 | Data (Hora da postagem de acordo com o fuso horário de Brasília) | | |

o momento pode ser interpretado de algumas maneiras diferentes por meio de número de *pixels* brancos (ou seja, intensidade = 1). Um algoritmo que realiza esses cálculos é o algoritmo de Hu ou o de Zernike.(HU, 1962; TEAGUE, 1980)

- Cores: Em geral, um histograma preserva mais informações das estatísticas de primeira ordem dos dados originais do que a média simples dos valores de dados brutos de uma imagem em formato binário. No caso mais simples, um modelo de histograma pode ser especificado para um tipo de recurso de imagem específico, independentemente de qualquer conteúdo de imagem real. No extremo oposto, os histogramas podem ser tornados dependentes não apenas do conteúdo real da imagem, mas também das classes semânticas conhecidas das imagens e os canais de cores que são utilizados nas imagens. *Threshold adjacency statistics* (TAS) são estatísticas de limite para células de imagens. Uma vez que uma célula de imagem é limitada, as estatísticas são calculadas a partir da imagem limite. Para cada *pixel* branco, o número de *pixels* adjacentes que também são brancos é contado. *Histogram of Oriented Gradients* (HOG) é uma representação simplificada da imagem que contém apenas as informações mais importantes sobre a imagem. Histograma de gradientes orientados é um descritor de recurso frequentemente usado para extrair recursos de dados de imagem, sendo amplamente utilizado em tarefas de visão computacional para detecção de objetos.(DALAL; TRIGGS, 2005)

- Textura: Outros Recursos que podem ser extraídos são de textura, um algoritmo é o de

Haralick, que é calculado a partir de uma matriz de concorrência de nível de cinza, sendo um método comum para representar a textura da imagem, pois é simples de implementar e resulta em um conjunto de descritores de textura interpretáveis facilmente por computadores. Padrões Binários ou *Local Binary Pattern* (LBP), ao contrário dos recursos de textura Haralick que calculam uma representação global da textura com base na Matriz de concorrência de Nível de Cinza, os LBPs, em vez disso, calculam uma representação local da textura. Essa representação local é construída comparando cada *pixel* com sua vizinhança de *pixels*. A primeira etapa na construção do descritor de textura LBP é converter a imagem em tons de cinza. Para cada *pixel* na imagem em tons de cinza, selecionamos uma vizinhança de tamanho r em torno do *pixel* central. Um valor LBP é então calculado para este *pixel* central e armazenado na matriz 2D de saída com a mesma largura e altura da imagem de entrada.(WALT et al., 2014)

- Tamanho: Toda imagem impressa ou vista em um monitor de computador é composta por pontos extremamente pequenos. Esses pontos, os menores elementos que formam uma imagem, são chamados de *pixels*. O tamanho de uma imagem pode ser representado por centímetros, polegadas ou *pixels* e representada por dois números: altura e largura.

Das características anteriormente citadas, 5 características tiveram uma frequência superior aos 70% nos 38 modelos que tiveram os melhores resultados de acurácia de acordo com o classificador conforme descrito na Tabela 11. Essas características são *Hu moments*, TAS, HOG, Zernike e o tamanho da imagem respectivamente. Ambos algoritmos que extraem características de formatos das imagens tiveram desempenhos quase semelhantes, já no aspecto textura, o algoritmo Haralick teve um desempenho bem melhor do que o LBP. Os classificadores que tiveram um desempenho superior aos 95% de acurácia foram os classificadores baseados em *Ensemble*: AdaBoost, Gradiente *Boosting*, XGBoosting e o classificador baseado em redes neurais o MLP Classifier. Os classificadores baseados em *Naive Bayes* tiveram um desempenho regular de 70 à 80% de acurácia para as classificações, uma diferença de 25% com relação aos outros classificadores já citados.

### 4.4.2 Características de Metadados

O grupo da Tabela 12 refere-se aos campos de data da postagem, *hashtag*, criador (para manter o anonimato foi colocado o id do usuário, algo que não é público na plataforma), quantidade de comentários e curtidas registradas na postagem. As datas das postagens pode ser dívida em 8 características individuais como: dia, mês, ano, dia da semana, semana (número da semana de acordo com o mês), segundo minuto e hora. Dessa forma existem 12 características que podem ser combinadas e totalizar 4.095 possibilidades de agrupamento e aplicados nos 17 classificadores. Os metadados é um grupo particular, pois utilizam tanto dados em formato de texto (*hashtags*) que precisam ser convertidos para números por meio de vetorizados, e

**Tabela 11 – Tabela com os melhores modelos do grupo de características das imagens postadas**

| Classificador | F1 | F2 | F3 | F4 | F5 | F6 | F7 | F8 | Qtd | Acurácia |
|---|---|---|---|---|---|---|---|---|---|---|
| AdaBoost | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 0.9580 |
| Bagging | 4 | 2 | 4 | 0 | 0 | 4 | 2 | 4 | 4 | 0.9493 |
| Bernoulli (Naive Bayes) | 8 | 8 | 8 | 16 | 16 | 16 | 0 | 8 | 16 | 0.7050 |
| Complement (Naive Bayes) | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 0.8046 |
| DecisionTree | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 0.8986 |
| ExtraTree | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0.8881 |
| ExtraTrees | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 0.9429 |
| Gradient Boosting | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.9639 |
| K Neighbors | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 0.9187 |
| Logistic Regression | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 0.9443 |
| MLP[1] Classifier | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 1 | 0.9516 |
| Multinomial (Naive Bayes) | 1 | 2 | 0 | 2 | 0 | 0 | 0 | 2 | 2 | 0.8050 |
| Nu SVC[2] | 2 | 1 | 2 | 2 | 2 | 2 | 1 | 2 | 2 | 0.9146 |
| Random Forest | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.9493 |
| SGD[3] | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 0.9379 |
| SVC | 2 | 1 | 2 | 0 | 2 | 2 | 0 | 2 | 2 | 0.9489 |
| XGBoosting | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 1 | 1 | 0.9667 |
| Total | 28 | 21 | 25 | 28 | 28 | 27 | 8 | 27 | 38 | |
| Frequência | 0.737 | 0.553 | 0.658 | 0.737 | 0.737 | 0.711 | 0.211 | 0.711 | | |

conjunto de números (demais características), portanto adaptações foram feitas para que ambos dados fossem combinados e utilizados nos algoritmos de classificação. A biblioteca do *sklearn* possui uma extensão para tratar dessa necessidade e se chama *make column transformer*. Uma característica recorrente em todos os modelos de aprendizado de máquina foram as *hashtags* (99%), outras características com bastante ocorrência foi o criador das postagens (82%). Com exceção do classificador Bernoulli (*Naive Bayes*), todos os outros classificadores atingiram os 99% de acurácia. Os classificadores *Bagging* e SVC foram os classificadores que mais tiveram modelos que atingiram essa marca sendo 1.103 e 687, ou seja dois terços de todos os 2.143 modelos que tiveram um bom desempenho desse grupo.

Pelas Figuras 6 e 7 pode-se afirmar que entre os dois grupos (postagens verdadeiras e falsas), o fluxo de postagens não segue o mesmo padrão. Enquanto postagens com conteúdo verdadeiro possuem um fluxo modero com poucas oscilações, postagens com conteúdos falsos possuem um fluxo intenso com muitas oscilações. Na Figura 6 o período com maior disseminação de postagens falsas das postagens coletadas começa em fevereiro e tem seu pico em Abril. Esse período conhecido também com o momento político em que o país passava com informações sobre a pandemia sendo disseminadas e muitas vezes distorcidas. No mês de pico de disseminações de notícias falsas foi o mês ao qual o ex-ministro da Saúde Luiz Henrique Mandetta foi demitido pelo presidente da república. Também foi o mês onde se intensificou a campanha empenhada por governadores nomeada "fica em casa"e neste mesmo período começaram a se difundir teorias conspiratórias que o vírus COVID-19 era uma arma biológica da China.

As *hashtags* expandem o limite de uma publicação, quanto mais *hashtags* utilizadas mais

**Tabela 12 – Tabela com os melhores modelos do grupo de características de metadados das postagens**

| Classificador | F9 | F10 | F11 | F12 | F13 | F14 | F15 | F16 | F17 | F18 | F19 | F20 | Qtd | Acurácia |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| AdaBoost | 82 | 46 | 26 | 41 | 8 | 0 | 20 | 12 | 26 | 20 | 32 | 82 | 82 | 0.9995 |
| Bagging | 641 | 343 | 345 | 381 | 346 | 338 | 425 | 330 | 442 | 465 | 294 | 687 | 687 | 0.9982 |
| Bernoulli (Naive Bayes) | 2 | 0 | 5 | 5 | 5 | 3 | 4 | 0 | 5 | 5 | 5 | 5 | 5 | 0.9511 |
| Complement (Naive Bayes) | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 0.9909 |
| DecisionTree | 2 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 2 | 0.9995 |
| ExtraTree | 11 | 10 | 10 | 12 | 8 | 10 | 5 | 8 | 5 | 6 | 8 | 13 | 13 | 0.9991 |
| ExtraTrees | 25 | 28 | 18 | 15 | 12 | 18 | 0 | 9 | 0 | 6 | 6 | 32 | 32 | 0.9995 |
| Gradient Boosting | 16 | 0 | 0 | 8 | 4 | 0 | 4 | 0 | 4 | 2 | 4 | 16 | 16 | 0.9995 |
| K Neighbors | 2 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0.9941 |
| Logistic Regression | 97 | 97 | 48 | 49 | 64 | 65 | 49 | 1 | 1 | 49 | 49 | 97 | 97 | 0.9991 |
| MLP Classifier | 11 | 12 | 7 | 7 | 2 | 4 | 6 | 7 | 9 | 11 | 2 | 12 | 12 | 0.9968 |
| Multinomial (Naive Bayes) | 3 | 3 | 2 | 0 | 1 | 2 | 3 | 0 | 3 | 3 | 3 | 3 | 3 | 0.9909 |
| Nu SVC | 35 | 20 | 18 | 27 | 3 | 0 | 29 | 8 | 2 | 20 | 14 | 35 | 35 | 0.9977 |
| Random Forest | 3 | 1 | 1 | 2 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 3 | 3 | 0.9995 |
| SGD | 3 | 2 | 1 | 1 | 2 | 0 | 1 | 3 | 1 | 2 | 3 | 3 | 3 | 0.9991 |
| SVC (SVM) | 788 | 503 | 551 | 620 | 559 | 553 | 692 | 565 | 757 | 677 | 555 | 1103 | 1103 | 0.9982 |
| XGBoosting | 47 | 39 | 20 | 15 | 16 | 22 | 11 | 19 | 8 | 26 | 6 | 47 | 47 | 0.9995 |
| Total | 1769 | 1105 | 1053 | 1185 | 1031 | 1017 | 1250 | 963 | 1264 | 1294 | 982 | 2141 | 2143 | |
| Frequência | 0.825 | 0.516 | 0.491 | 0.553 | 0.481 | 0.475 | 0.583 | 0.449 | 0.590 | 0.604 | 0.458 | 0.999 | | |

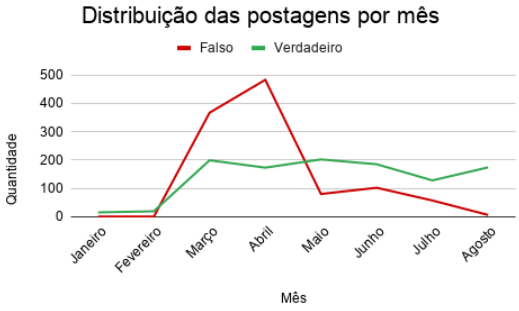


**Figura 6 – Postagens por mês**

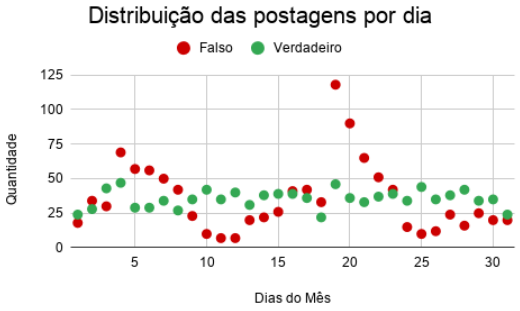


**Figura 7 – Postagens por dia**

conexões e indicações são feitas para pessoas que pesquisam ou postam os mesmos termos. A Figura 8 agrupa as 20 *hashtags* mais utilizadas nas postagens verdadeiras e a Figura 9 ilustra qwv*hashtags* mais utilizadas nas postagens falsas. Nas *hashtags* com as informações verdadeiras, as palavras com maior frequência são os nomes dos jornais (Uol, CNN Brasil etc.), nos quais as postagens foram extraídas. Uma outra *hashtag* utilizada é a PraCegoVer, esse é um recurso ao qual páginas jornalísticas e outras utilizam para facilitar o acesso de deficientes audiovisuais ao conteúdo da imagem. A Figura 9 possui inúmeros termos. dentre os quais alguns inclusive não possuem ligação direta com o foco da postagem que é o assunto sobre a pandemia de Covid-19. Algumas dessas *hashtags* SomosTodosLavaJato, MaiaInimigoDoBrasil e muitas outras que seriam de apoio ao presidente.

### 4.4.3 Texto das Imagens e Legendas

Os modelos de aprendizado de máquina baseados em características dos textos foram aplicados nas amostras da legenda e dos textos extraídos das imagens pelo processo de OCR.

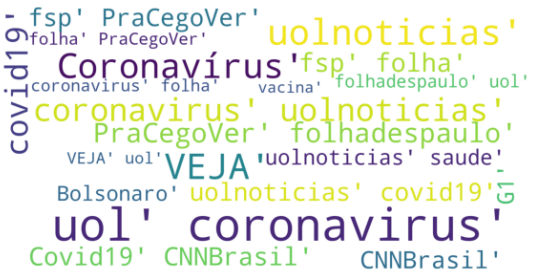


**Figura 8 – Hashtags em postagens verdadeiras**


**Figura 9 – Hashtags em postagens falsas**

Foram realizados por meio das 10 possibilidades de características textuais, produzindo 108 modelos em cada um dos 17 algoritmos de classificação gerando um total de 1.836 modelos para legendas e outros 1.836 para os textos contidos nas imagens. Os melhores modelos foram selecionados de acordo com a acurácia mais alta que cada classificador atingiu. Em alguns casos ocorreu de um classificador ter mais de um experimento que atingiu o mesmo valor.

Nas Tabelas 13 e 14 está o índice de acurácia e a quantidade de experimento que atingiu aquela acurácia de acordo com o classificador. Por exemplo, com o Algoritmo AdaBoost o número total de experimentos foram 2 e o de Bernoulli uma adaptação do tradicional *Naive Bayes* foram 4 modelos que atingiram a maior acurácia. Porém, nem todos os modelos utilizaram as mesmas características. Dos quatro modelos apenas 2 usaram características *vector tfidfvectorizer*, *vector countvectorizer*, texto em forma bruta (com letras maiúsculas e minúsculas) e palavras convertidas para minúsculas. Entretanto, em todos os modelos foram utilizadas as características técnicas de *n-Gram* para palavras e *stopwords* para o texto das legendas de acordo com a Tabela 13. A média de acurácia dos elementos da tabela citada anteriormente são de 96% à 99%. Os algoritmos de XGBoosting, Random Forest, Decision Tree e AdaBoost foram os que atingiram níveis mais altos. Algumas características tiveram pouca ocorrência nos modelos selecionados, a saber: *vector hashingvectorizer* e *n-Gram* em caracteres limitados dentro de cada palavra. Por outro lado, a característica técnica de *n-Gram* aplicada as palavras teve uma ocorrência maior dentro dos palavras convertidas para minúsculas, gerando melhores modelos selecionados e depois *stopwords*, texto em forma bruta (com letras maiúsculas e minúsculas) e *n-Gram* convencional (*bi-Gram* e *tri-Gram*).

Os modelos que analisaram os textos extraídos das imagens tiveram uma acurácia de 87% à 96%, valores estes que podem ser verificados na Tabela 14. O Melhor classificador foi o SVC e logo depois *Bagging*, *XGBoosting*, *ExtraTrees*, *Logistic Regression* e *Random Forest*. os quais chegaram mais próximos dos 96% que SVC teve como desempenho. As características tiveram o mesmo desempenho que os modelos dos textos das legendas técnica de *n-Gram* aplicada as palavras, *stopwords*, texto em forma bruta (com letras maiúsculas e minúsculas) e *n-Gram* convencional (*bi-Gram* e *tri-Gram*).

**Tabela 13 – Tabela com os melhores modelos do grupo de características do texto da legenda**

| Classificador | F21 | F22 | F23 | F24 | F25 | F26 | F27 | F28 | F29 | F30 | F31 | Qtd | Acurácia |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| AdaBoost | 0 | 2 | 0 | 2 | 0 | 0 | 2 | 0 | 0 | 2 | 1 | 2 | 0.9991 |
| Bagging | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 0.9918 |
| Bernoulli (Naive Bayes) | 2 | 2 | 0 | 0 | 4 | 0 | 2 | 0 | 2 | 0 | 4 | 4 | 0.9795 |
| Complement (Naive Bayes) | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 0.9845 |
| DecisionTree | 2 | 0 | 0 | 0 | 2 | 0 | 1 | 0 | 1 | 2 | 0 | 2 | 0.9991 |
| ExtraTree | 2 | 0 | 0 | 0 | 2 | 0 | 1 | 0 | 1 | 2 | 2 | 2 | 0.9667 |
| ExtraTrees | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 1 | 1 | 0.9977 |
| Gradient Boosting | 0 | 2 | 0 | 2 | 0 | 0 | 2 | 0 | 0 | 2 | 1 | 2 | 0.9991 |
| K Neighbors | 0 | 0 | 2 | 0 | 2 | 0 | 1 | 0 | 1 | 0 | 2 | 2 | 0.9831 |
| Logistic Regression | 0 | 2 | 0 | 0 | 2 | 0 | 1 | 0 | 1 | 0 | 2 | 2 | 0.9973 |
| MLP Classifier | 0 | 2 | 0 | 2 | 0 | 0 | 2 | 0 | 0 | 2 | 1 | 2 | 0.9959 |
| Multinomial (Naive Bayes) | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 0.9845 |
| Nu SVC | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0.9822 |
| Random Forest | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 1 | 1 | 0.9991 |
| SGD | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0.9963 |
| SVC (SVM) | 2 | 0 | 0 | 0 | 2 | 0 | 1 | 0 | 1 | 0 | 0 | 2 | 0.9932 |
| XGBoosting | 2 | 0 | 0 | 0 | 0 | 2 | 2 | 0 | 0 | 2 | 1 | 2 | 0.9991 |
| Total | 14 | 13 | 2 | 6 | 21 | 2 | 16 | 6 | 7 | 15 | 18 | 29 | |
| Frequência | 0.483 | 0.448 | 0.069 | 0.207 | 0.724 | 0.069 | 0.552 | 0.207 | 0.241 | 0.517 | 0.621 | | |

**Tabela 14 – Tabela com os melhores modelos do grupo de características do texto extraído das imagens**

| Classificador | F21 | F22 | F23 | F24 | F25 | F26 | F27 | F28 | F29 | F30 | F31 | Qtd | Acurácia |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| AdaBoost | 0 | 2 | 0 | 2 | 0 | 0 | 2 | 0 | 0 | 2 | 1 | 2 | 0.9479 |
| Bagging | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 0.9598 |
| Bernoulli (Naive Bayes) | 2 | 2 | 0 | 0 | 4 | 0 | 2 | 0 | 2 | 0 | 4 | 4 | 0.8776 |
| Complement (Naive Bayes) | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 0.9388 |
| DecisionTree | 2 | 0 | 0 | 0 | 2 | 0 | 1 | 0 | 1 | 2 | 0 | 2 | 0.9301 |
| ExtraTree | 2 | 0 | 0 | 0 | 2 | 0 | 1 | 0 | 1 | 2 | 2 | 2 | 0.8968 |
| ExtraTrees | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 1 | 1 | 0.9584 |
| Gradient Boosting | 0 | 2 | 0 | 2 | 0 | 0 | 2 | 0 | 0 | 2 | 1 | 2 | 0.9525 |
| K Neighbors | 0 | 0 | 2 | 0 | 2 | 0 | 1 | 0 | 1 | 0 | 2 | 2 | 0.8954 |
| Logistic Regression | 0 | 2 | 0 | 0 | 2 | 0 | 1 | 0 | 1 | 0 | 2 | 2 | 0.9594 |
| MLP Classifier | 0 | 2 | 0 | 2 | 0 | 0 | 2 | 0 | 0 | 2 | 1 | 2 | 0.9521 |
| Multinomial (Naive Bayes) | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 0.9388 |
| Nu SVC | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0.9580 |
| Random Forest | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 1 | 1 | 0.9562 |
| SGD | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0.9502 |
| SVC (SVM) | 2 | 0 | 0 | 0 | 2 | 0 | 1 | 0 | 1 | 0 | 0 | 2 | 0.9639 |
| XGBoosting | 2 | 0 | 0 | 0 | 0 | 2 | 2 | 0 | 0 | 2 | 1 | 2 | 0.9575 |
| Total | 14 | 13 | 2 | 6 | 21 | 2 | 16 | 6 | 7 | 15 | 18 | 29 | |
| Frequência | 0.483 | 0.448 | 0.069 | 0.207 | 0.724 | 0.069 | 0.552 | 0.207 | 0.241 | 0.517 | 0.621 | | |

Outro ponto importante de ser destacado é que os vetores v*ector tfidfvectorizer* e *vector countvectorizer*, ambos nos grupos texto da legenda e texto da imagem, tiveram quase a mesma porcentagem de ocorrência, enquanto a terceira opção o *vector hashingvectorizer* teve 0,07% de ocorrência e os outros dois juntos 93%.

## 4.5 Resultados Obtidos

Essa seção discute os resultados obtidos nos diversos experimentos descritos anteriormente nesse Capítulo.

A primeira etapa do projeto foi a extração das postagens da rede social Instagram, por

meio de técnicas de *scraping* todas as informações necessárias das postagens foram extraídas e armazenadas para uso de pesquisa. Os dados foram extraídos em formato JSON (metadados e as legendas) e as imagens em formato JPG. Para uma melhor identificação dos dados foram tratados na etapa 2 e armazenados em um banco de dados orientados a documentos (Mongo DB) para serem usados na próxima etapa que é a extração de características dos metadados e das imagens. Dessa forma, tais informações foram inseridas na etapa 4 que foi o aprendizado de máquina. Os modelos foram executados em uma máquina HP ProDesk 400 G6 (9EC17LA) com processador Intel Core i3-9100 de 9ª geração, 3,6 GHz de frequência básica e Cache de 6 MB, 4 núcleos, 4 threads. O Sistema Operacional utilizado foi Ubuntu 20.04 em um HD SSD Samsung de 1 TB e memória RAM DDR4 de 20GB.

Foram utilizados os 17 classificadores da biblioteca *sklearn*, eles são utilizados para aprendizado supervisionado e a fonte de dados tanto para o treino quanto para o teste são rotulados (como verdadeiro e falso). Todos os classificadores tiveram o mesmo agrupamento de dados e características, dessa forma foi possível avaliar a média de desempenho desses classificadores.

Algumas limitações foram encontradas ao longo do projeto, uma delas foi a falta de uma base contendo os dados já curados prontos para aplicar os modelos de aprendizado de máquina. Neste projeto apenas o autor que aferiu as imagens e as classificou, por isso apenas cerca de 10% das imagens foram classificadas. Outras limitações encontradas é a forma com as quais o Instagram dificulta o acesso aos dados, as postagens são de domínio público, entretanto a plataforma restringe as ferramentas que utilizam técnicas de *scraping* (extrair dados de determinada plataforma por meio de ferramentas automatizadas) possuem baixo desempenho (demora muito tempo para extrair as informações).

Um total de 77.622 modelos foram analisados por meio de modelos de aprendizado de máquina, e 2.239 foram agrupados e destacados na seção anterior pois obtiveram a melhor acurácia de acordo com o classificador. Para uma análise geral de desempenho dos classificadores foi feito a média das principais métricas de cada um dos 17 classificadores. Foi considerado um bom desempenho os algorítimos que superavam a marca de 90% de F1-Score e Acurácia que são duas medidas muito utilizadas para avaliar experimentos de aprendizado supervisionado de máquina.

Nas Figuras 10, 11, 12, 13 temos no primeiro gráfico a avaliação dos algoritmos para classificar a postagem como falso e o segundo a exatidão de como classificou os verdadeiros. Cada uma das quatro figuras representa cada uma das características avaliadas neste projeto: texto da imagem, texto da legenda, imagem e metadados.

Em ambos rótulos (verdadeiros e falsos) e nos grupos de abordagem, os algoritmos de classificação que aparecem com um alto desempenho nos quatro grupos foram: *Extra Trees*, *Gradient Boosting*, Floresta Aleatória e *XGBoosting*. As médias de acurácia e F1 que podem medir a exatidão do processo de classificação variaram de 90% à 97% com os algorítimos

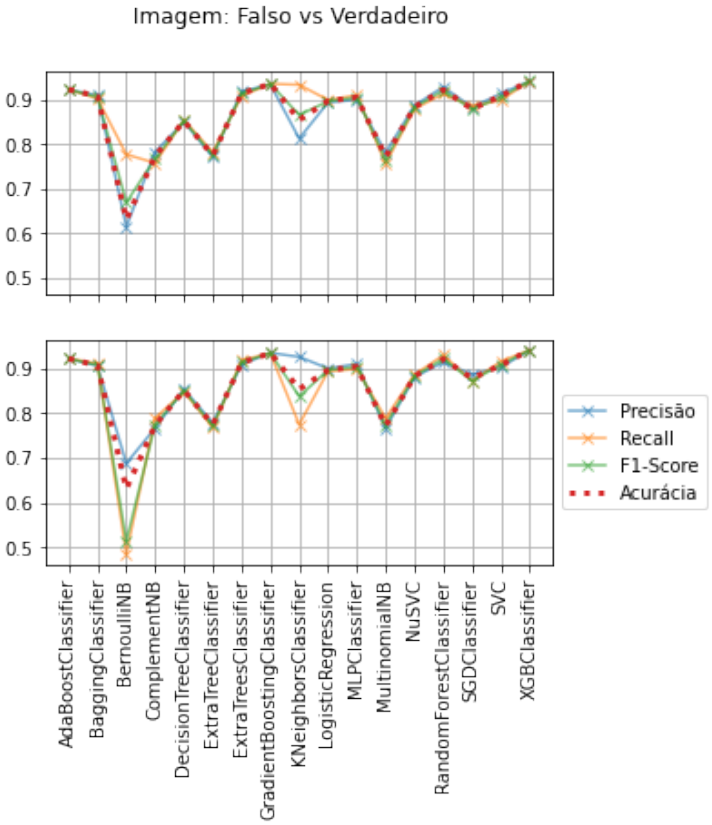


**Figura 10 – Média dos Classificadores para característica de imagens**

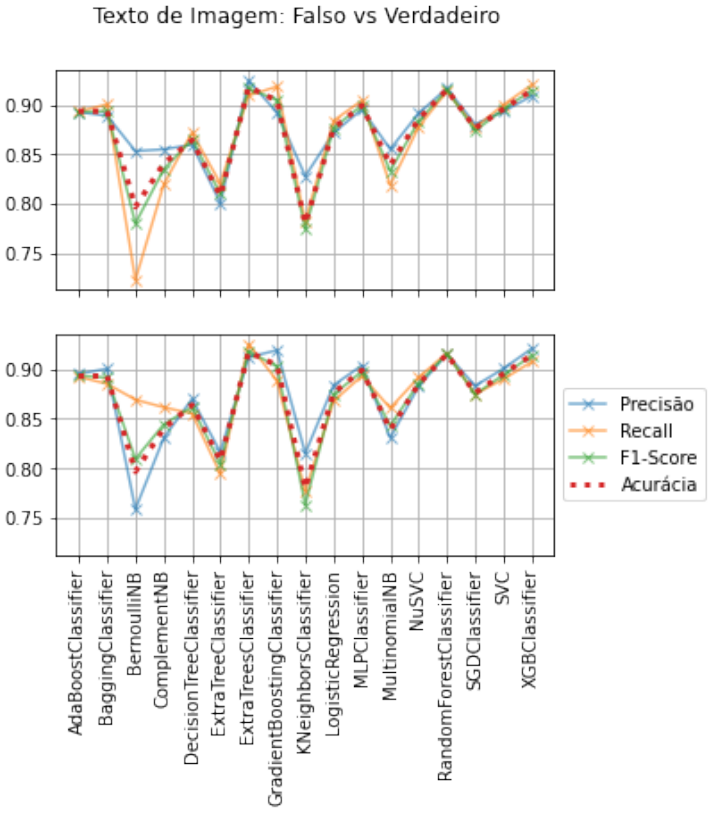


**Figura 11 – Média dos Classificadores para característica do texto das imagens**

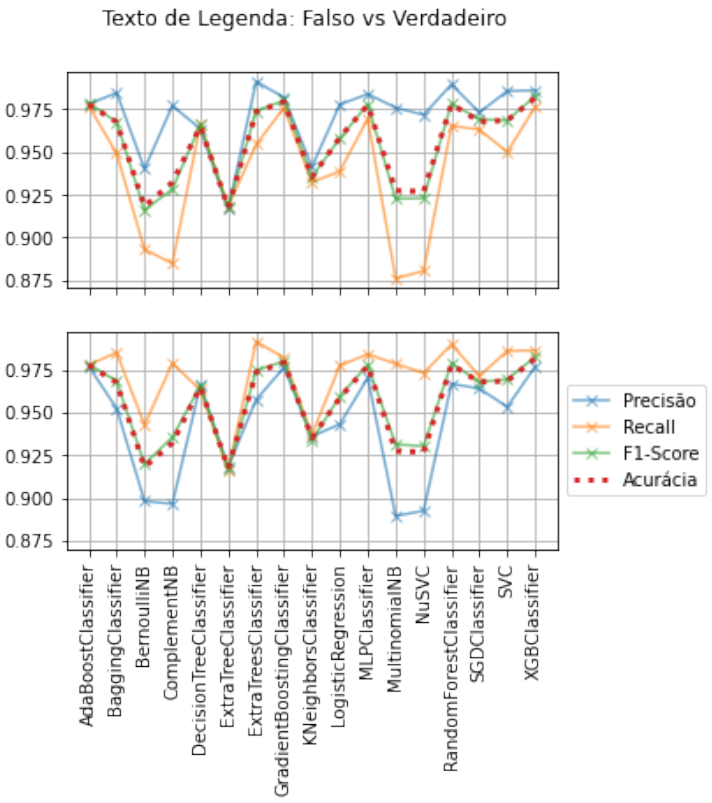


**Figura 12 – Média dos Classificadores para característica do texto das legendas**

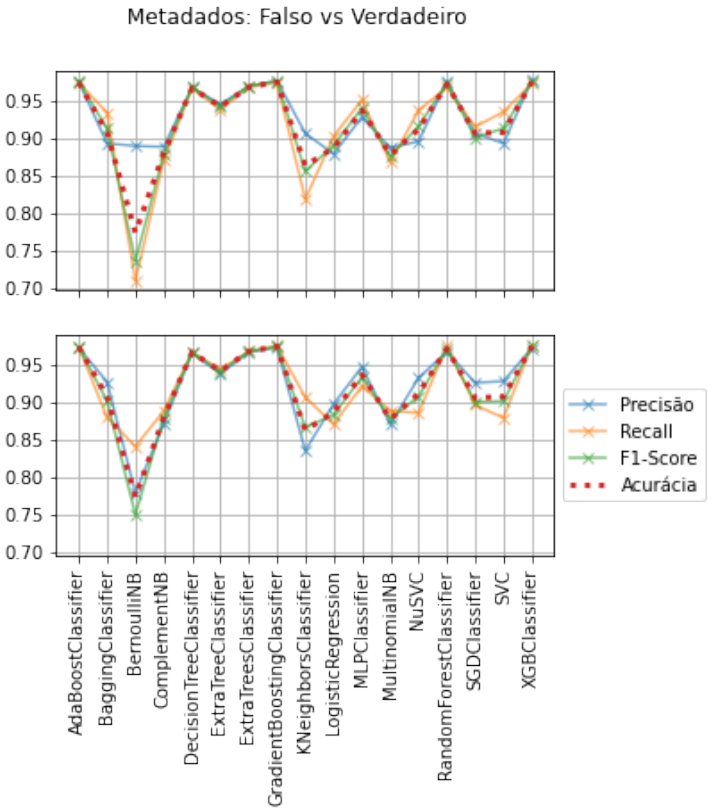


**Figura 13 – Média dos Classificadores para característica do metadados**

citados anteriormente em diferentes grupos. Outro algoritmo que é importante de ser destacado é o *Perceptron* Multicamadas, ou conhecido também como *MultiLayer Perceptron* (MLP), que proporcionou um bom desempenho no grupo de texto das imagens, texto da legenda e imagens. *AdaBoosting* também teve um bom desempenho em grupos diferentes como Texto da Legenda, Metadados e Imagens. Alguns algoritmos tiveram desempenhos bons. Porém, em único grupo como o *Extra Tree* no grupo de Metadados e Regressão Logística e *Support Vector Clustering* (SVC) para o grupo de Imagem.

Como foi avaliada a média dos resultados produzidos pelos algoritmos de classificação, outra avaliação realizada foi o desempenho dos melhores experimentos de cada grupo. Tal avaliação foi feita por meio da projeção do gráfico de Características de Operação do Receptor, também conhecida como ROC, utilizada para ilustrar o desempenho de classificações binárias (no caso desse projeto Verdadeiro e Falso). Para poder gerar o gráfico, os experimentos foram executados novamente, e como forma de chegar cada vez mais perto de um resultado preciso e mais perto da realidade, os dados de treino e teste foram divididos em 10 grupos com quantidades próximas de cada rótulo. Nas imagens 14, 15, 16 são ilustradas por método de classificação o valor de acurácia, sendo que em ambos experimentos foi utilizado o algoritmo XGBoosting.

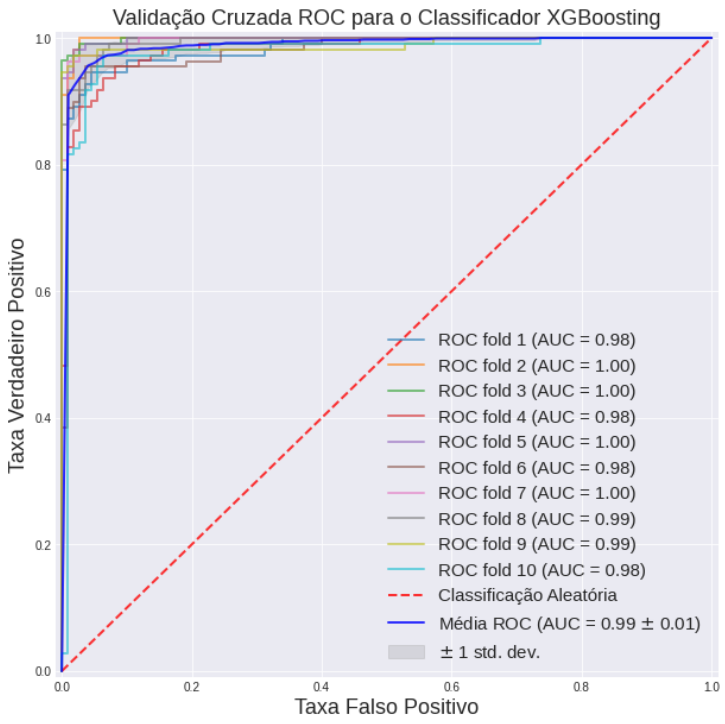


**Figura 14 – Experimento de Característica de Imagem**

O experimento das características de imagem, texto da legenda e metadados tiveram uma das acurácias mais altas de 99%. O experimento com texto nas imagens que está descrito na Figura 17 produziu de 98% de acurácia. Medir a curva ROC é importante para acompanhar o quanto de variações um mesmo grupo obteve na classificação. Por isso, as variações como no caso da classificação da imagem. O ideal do experimento que tenha um ótimo desempenho é

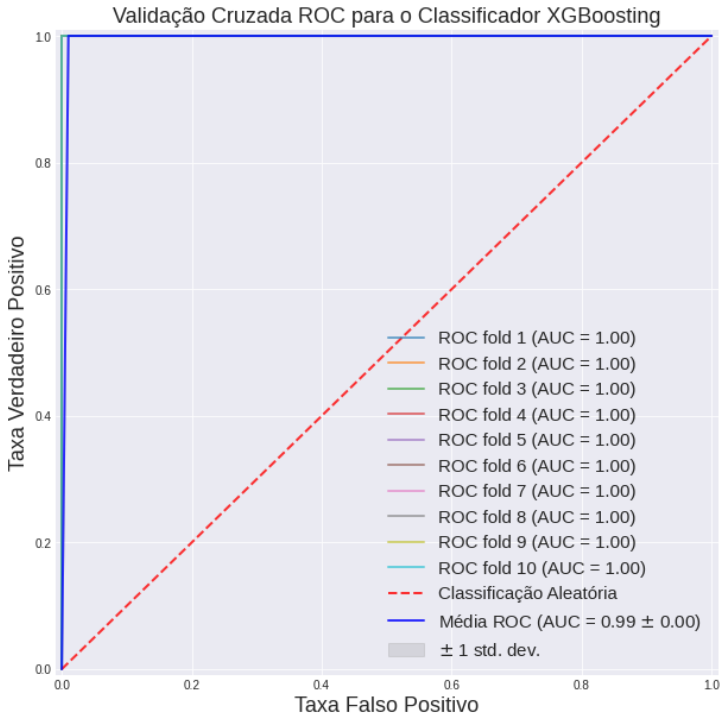


**Figura 15 – Experimento de Característica de Metadados**

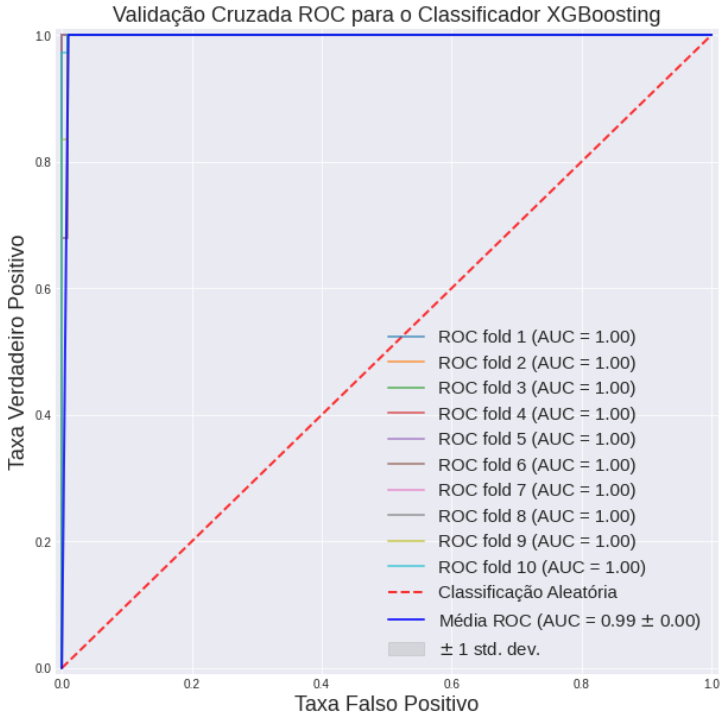


**Figura 16 – Experimento de Característica de Texto da Legenda**

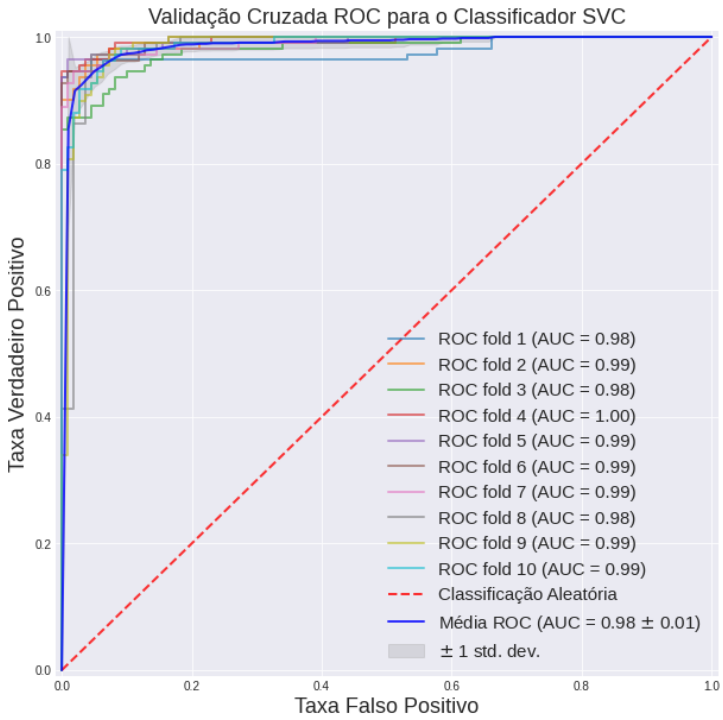


**Figura 17 – Experimento de Característica de Texto da Imagem**

quando a curva ROC segue superiormente a linha vermelha e próxima da borda superior do gráfico, essa área do gráfico mostra o quanto de acertos o classificador teve, ou como é chamado em aprendizado de máquina ilustra a Taxa de Verdadeiros Positivos, que são os dados que foram corretamente rotulados pelo algoritmo.

Em resumo, os experimentos demonstram que todos os grupos de características das postagens da plataforma Instagram podem trazer desempenhos similares de acurácia. Mesmo as características de texto da Imagem tendo valores pouco menores é uma das características mais importantes que devem ser avaliadas quando o assunto é informação falsa, pois esse método vem crescendo muito, não apenas na Plataforma do Instagram ao se compartilhar tais imagens nos *feeds*, como nos *stories* e também em outras mídias sócias. Portanto, não deve ser um grupo de características que devam ser descartados. Além disso, é possível afirmar que utilizar as técnicas de OCR para extrair os textos das imagens com fins de aplicá-las em processos de PLN para os modelos de aprendizado de máquina são tarefas essenciais quando a tarefa for classificar as postagens de redes sociais com características de multimídia entre elas a utilizada nesse trabalho, a rede social Instagram.

# Capítulo 5
# CONCLUSÃO

## 5.1 Considerações Iniciais

Neste capítulo são apresentadas as conclusões sobre a pesquisa de Mestrado. O Capítulo está dividido em três seções, sendo que na primeira são discutidos os resultados dos experimentos. Na segunda seção são apresentadas todas as dificuldades e limitações que foram surgindo no decorrer do projeto. Por fim, na terceira e última seção são apresentados os trabalhos futuros indicados para serem realizados como continuação a esta pesquisa de Mestrado.

Neste trabalho foram propostos e usados vários modelos de aprendizado de máquina que analisam e classificam postagens com informações falsas e verdadeiras em português na plataforma Instagram. Tal ação se faz necessária devido ao alto número de informações que são compartilhadas nessa rede social. Se torna uma dificuldade maior ainda quando as informações são compartilhadas em imagens, o que torna difícil para os algoritmos da plataforma de identificar palavras chaves que considerem indevidas de serem usadas na plataforma, ou até mesmo para problemas de acessibilidade. Como solução esse trabalho utilizou o sistema OCR Tesseract para extrair os textos das imagens, e eles foram depois manualmente verificados para conter exatamente o mesmo texto da imagem, para no fim do processo poder estipular o quanto de acerto tal processo obteve em extrair os textos das imagens.

## 5.2 Contribuições

Para a execução dos experimentos foi necessário extrair dados e realizar uma filtragem nos dados para que os dados ficassem corretos e em conformidade com os formatos e padrões de confidencialidade exigidos para a aplicação dos modelos. Então todas as imagens das postagens foram verificadas, e as que eram de cunho pessoal (*selfies* ou com apenas rosto de pessoas não públicas) foram retiradas. Todos os textos das imagens foram extraídos por meio do uso da técnica de OCR e depois verificados individualmente para verificar se o texto precisava de correção ou não. Nesse sentido, um total de 23.951 imagens com textos passaram por esse processo. Esse conjunto de dados foi guardado em um banco de dados orientado a documentos.

As imagens foram guardadas em arquivos a parte devido ao seu tamanho. Porém, todas as imagens foram devidamente anexadas nos documentos do banco de dados, facilitando assim quando necessário visualizar todos os dados de uma postagem incluindo a sua foto.

Uma parte das postagens foi também classificada manualmente em dois grupos como verdadeiros e falsos, seguindo informações de páginas de verificação de conteúdo. Com esse grupo de dados também podem ser feitas outras classificações para as postagens como se são uma sátira, pelo simples fato de ter muitas charges em postagens com a *hastag* vírus chinês.

Com realização a implementação, também foi produzida uma lista de *notebooks* que detalha todo o processo de Transformação e Carregamento dos dados para o Banco de dados orientado a documentos, esses arquivos estão no formato *ipynb*, e a divisão foi feita pelas etapas realizadas neste trabalho. Todos os arquivos têm comentários e bibliotecas utilizadas já inclusas. Os *notebooks* foram adaptados para realizar esses experimentos em outros tipos de classificação e não necessariamente apenas classificações binarias (falso e verdadeiro).

Além dessas contribuições anteriormente citadas, essa pesquisa de Mestrado também produziu as seguintes contribuições:

- Construção de modelos de detecção de informações falsas em português com enfoque específico e original para a rede social Instagram.
- Tratamento da detecção de notícias falsas por meio do processamento de imagens usadas no compartilhamento de conteúdo falsas.
- Proposta de uma arquitetura para um ambiente que processe e detecte notícias falsas de postagens de redes sociais.
- Técnica que pode ser de interesse para a comunidade do Instagram, oferecendo uma solução para que postagens falsas sejam identificadas, expostas e que não alcance um alto nível de difusão.
- Realização de experimentos intensivos e abrangentes com ferramentas de PLN para produzir análises, contendo a acurácia dos resultados para que outros estudos sobre redes sociais possam identificar padrões de comportamento de usuários ou postagens em diferentes redes sociais.
- Identificação das principais notícias já espalhadas sobre o COVID-19, para combater a desinformação que pode causar danos a toda sociedade.
- Escrita de artigo científico para evento qualificado na área de Computação contendo os resultados obtidos (SBBD 2021).

## 5.3 Trabalhos Futuros

A seguir são enumeradas algumas sugestões de trabalhos futuros:

- Rede de compartilhamento de postagens e analise de sentimentos dentro do Instagram: Uma das verificações que foram possíveis de saber no âmbito dessa pesquisa de Mestrado é o comportamento em termos de volume da postagem em uma linha do tempo, ou em outras palavras como foi a variação da quantidade de postagem dentro do período investigado. Outra abordagem que poderia ser trabalhada é a formação das redes de compartilhamento, por exemplo, como um perfil que segue outro perfil se comporta em diferentes postagens. Analisando também os comentários de cada uma das postagens. Neste trabalho os endereços são disponibilizados, portanto apenas haveria a necessidade de extrair os comentários das postagens para analisar e classificar os comentários como positivos ou negativos das páginas de Instagram dos principais jornais do país.

- Otimizar o uso do sistema OCR: a ferramenta de OCR foi utilizada nesse trabalho como um apoio para a extração dos textos das imagens, por isso o grau de erro de 22% não foi um fator que precisou ser otimizado nesta pesquisa. Porém. com treinamento o índice pode atingir melhores níveis de exatidão nas extrações. Na base de dados compartilhada, existe um campo de gabarito que contém exatamente o texto reproduzido na imagem, esse campo foi supervisionado e verificado individualmente imagem por imagem para garantir a integridade do texto. Alguns trabalhos podem utilizar esse gabarito e as suas extrações de texto pata aperfeiçoar o método de OCR.

- Aplicação dos Modelos em um ambiente *Web*: com as análises de características e classificadores que foram utilizadas nesse trabalho é possível que a própria rede social, fomente uma base de dados com as páginas dos jornais e verificadores de conteúdo e por meio de um sistema de usuários autenticados (usuários que são perfis reais e comprovam a sua identidade junto a plataforma) possa alimentar a base verdadeira, e com as informações extraídas dessa base, as postagens que forem marcadas como um conteúdo falso podem ter uma pré-triagem nessa base de conhecimento já pré-classificando a postagem em dúvida, economizando muitas vezes postagens duplicadas e a necessidade de verificar manualmente as informações já apuradas.

- Produção de uma produção técnica que permita que usuários enviem postagens do Instagram e o sistema informe se a postagem é falsa ou verdadeira. O uso de lógica *fuzzy* pode adicionar incerteza para esse sistema e produzir graus de certeza sobre a postagem sendo verificada.

# Apêndice A

## IMAGENS COM INFORMAÇÕES VERDADEIRAS

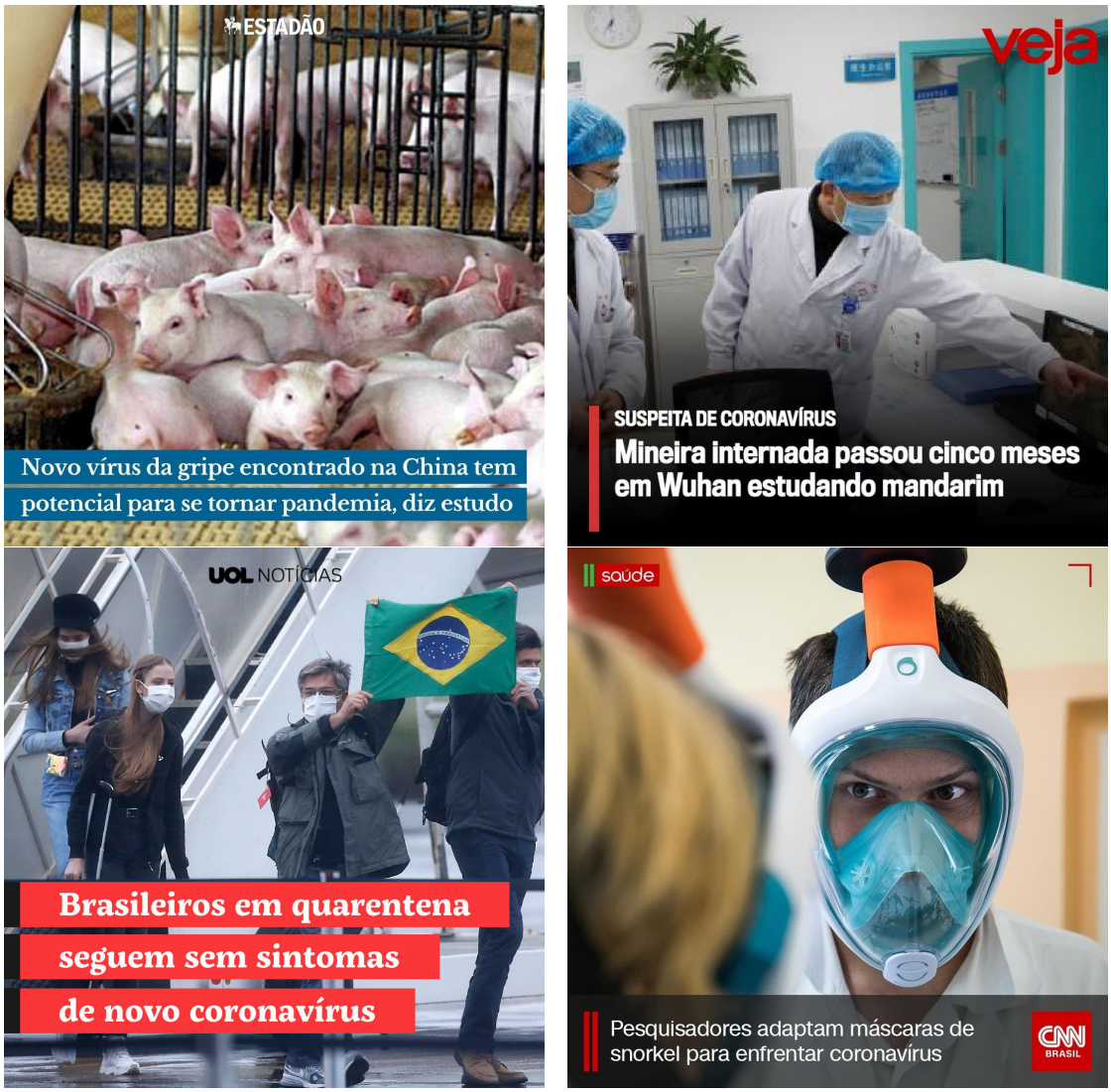


**Figura 18 – Imagens compartilhadas por jornais**

# Apêndice B

## IMAGENS COM INFORMAÇÕES FALSAS

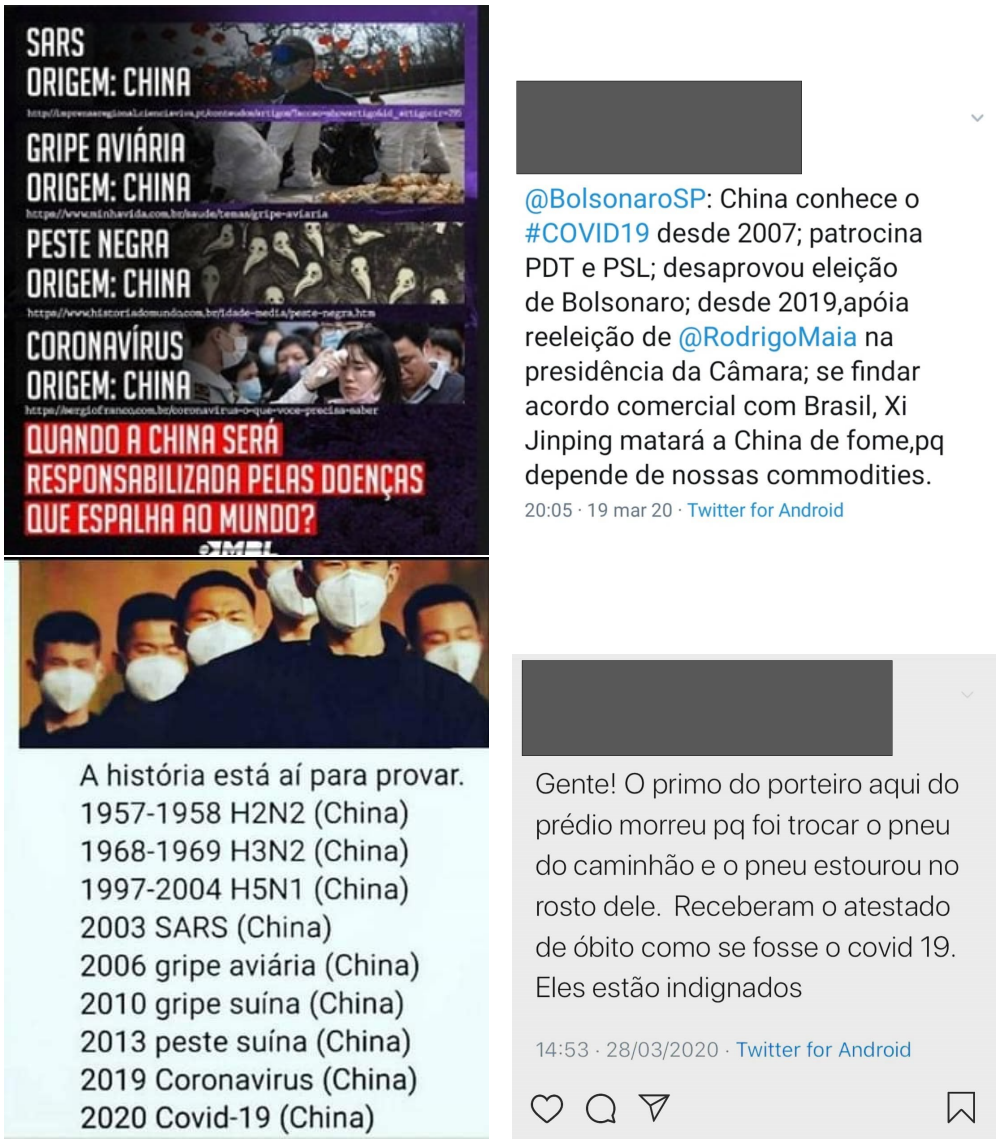


**Figura 19 – Imagens contendo informações falsas**

# REFERÊNCIAS

ACERBI, A. Cognitive attraction and online misinformation. v. 5, n. 1, 2019. ISSN 2055-1045. Disponível em: <http://www.nature.com/articles/s41599-019-0224-y>. Citado na página 79.

AGRAWAL, T. *Hyperparameter Optimization in Machine Learning: Make Your Machine Learning and Deep Learning Models More Efficient*. [S.l.]: Apress, 2020. Citado 3 vezes nas páginas 42, 44 e 45.

AHMED, S.; HINKELMANN, K.; CORRADINI, F. *Combining machine learning with knowledge engineering to detect fake news in social networks - A survey*. 2019. Disponível em: <www.fiskkit.com>. Citado 2 vezes nas páginas 50 e 58.

APUKE, O. D.; OMAR, B. Fake news and covid-19: modelling the predictors of fake news sharing among social media users. *Telematics and Informatics*, Elsevier Ltd, v. 56, p. 101475, 1 2021. ISSN 07365853. Citado 2 vezes nas páginas 65 e 69.

ATODIRESEI, C.-S.; TĂNĂSELEA, A.; IFTENE, A. Identifying fake news and fake users on twitter. v. 126, p. 451–461, 2018. ISSN 1877-0509. Disponível em: <http://www.sciencedirect.com/science/article/pii/S1877050918312559>. Citado 3 vezes nas páginas 49, 56 e 76.

BAGADE, A.; PALE, A.; SHETH, S.; AGARWAL, M.; CHAKRABARTI, S.; CHEBROLU, K.; SUDARSHAN, S. The kauwa-kaate fake news detection system: Demo. In: *Proceedings of the 7th ACM IKDD CoDS and 25th COMAD*. Association for Computing Machinery, 2020. (CoDS COMAD 2020), p. 302–306. ISBN 978-1-4503-7738-6. Disponível em: <https://doi.org/10.1145/3371158.3371402>. Citado 3 vezes nas páginas 49, 56 e 76.

BANG, Y.; ISHII, E.; CAHYAWijAYA, S.; JI, Z.; FUNG, P. Model generalization on covid-19 fake news detection. 2021. Disponível em: <https://www.bbc.com/news/world-53755067>. Citado 3 vezes nas páginas 64, 65 e 66.

BARRETT, P. M. *Disinformation and the 2020 Election: How the Social Media Industry Should Prepare*. 2019. Disponível em: <https://issuu.com/nyusterncenterforbusinessandhumanri/docs/nyu_election_2020_report/1>. Citado 4 vezes nas páginas 16, 20, 46 e 79.

BORISYUK, F.; GORDO, A.; SIVAKUMAR, V. Rosetta: Large scale system for text detection and recognition in images. In: *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 2018. (KDD '18), p. 71–79. ISBN 978-1-4503-5552-0. Event-place: London, United Kingdom. Disponível em: <http://doi.acm.org/10.1145/3219819.3219861>. Citado na página 81.

BOVET, A.; MAKSE, H. A. Influence of fake news in twitter during the 2016 US presidential election. v. 10, n. 1, p. 1–14, 2019. ISSN 2041-1723. Disponível em: <https://www.nature.com/articles/s41467-018-07761-2>. Citado 3 vezes nas páginas 46, 49 e 57.

BUNKE, H.; WANG, P. S. P. WORLD SCIENTIFIC, 1997. ISBN 978-981-02-2270-3 978-981-283-096-8. Disponível em: <https://www.worldscientific.com/worldscibooks/10.1142/2757>. Citado 6 vezes nas páginas 22, 23, 24, 26, 27 e 28.

CHAUDHURI, A.; MANDAVIYA, K.; BADELIA, P.; GHOSH, S. K. *Optical Character Recognition Systems for Different Languages with Soft Computing*. Springer International Publishing, 2017. (Studies in Fuzziness and Soft Computing). ISBN 978-3-319-50251-9. Disponível em: <https://www.springer.com/gp/book/9783319502519>. Citado 4 vezes nas páginas 23, 26, 27 e 28.

CHEN, B.; CHEN, B.; GAO, D.; CHEN, Q.; HUO, C.; MENG, X.; REN, W.; ZHOU, Y. Transformer-based language model fine-tuning methods for covid-19 fake news detection. 2021. Disponível em: <https://github.com/huggingface/tokenizers>. Citado 3 vezes nas páginas 64, 65 e 66.

CONROY, N. J.; RUBIN, V. L.; CHEN, Y. Automatic deception detection: Methods for finding fake news. *Proceedings of the Association for Information Science and Technology*, v. 52, p. 1–4, 2015. ISSN 23739231. Tem o artigo impresso e já comentado. Citado 2 vezes nas páginas 50 e 59.

CORREIA, P. H. B.; RIVERA, G. A. R. Evaluation of OCR free software applied to old books. n. 26, 2018. ISSN 2596-1969. Disponível em: <https://econtents.bc.unicamp.br/eventos/index.php/pibic/article/view/1132>. Citado na página 46.

CURY, M. E. WhatsApp confirma envio ilegal de mensagens por grupos políticos em 2018. 2019. Disponível em: <https://exame.abril.com.br/tecnologia/whatsapp-confirma-envio-ilegal-de-fake-news-por-grupos-politicos-em-2018/>. Citado na página 19.

DALAL, N.; TRIGGS, B. Histograms of oriented gradients for human detection. In: *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*. [S.l.: s.n.], 2005. v. 1, p. 886–893 vol. 1. Citado na página 83.

DAS, S. D.; BASAK, A.; DUTTA, S. A heuristic-driven ensemble framework for covid-19 fake news detection. 2021. Disponível em: <https://competitions.codalab.org/competitions/26655>. Citado 3 vezes nas páginas 64, 65 e 67.

DATASENADO. *Redes sociais influenciam voto de 45% da população, indica pesquisa do DataSenado*. 2019. Disponível em: <https://www12.senado.leg.br/noticias/materias/2019/12/12/redes-sociais-influenciam-voto-de-45-da-populacao-indica-pesquisa-do-datasenado>. Citado na página 18.

DONG, X.; VICTOR, U.; QIAN, L. Two-path deep semi-supervised learning for timely fake news detection. 2020. Disponível em: <http://arxiv.org/abs/2002.00763>. Citado 2 vezes nas páginas 49 e 57.

ELHADAD, M. K.; LI, K. F.; GEBALI, F. Fake news detection on social media: A systematic survey. In: . [S.l.]: Institute of Electrical and Electronics Engineers Inc., 2019. ISBN 9781728127941. Artigo está impresso e comentado. Citado 2 vezes nas páginas 50 e 63.

ESPOSITO, F. E. D. *Introducing Machine Learning*. [S.l.]: Microsoft Press, 2020. Citado na página 38.

FABIO, A. C. *O que é 'pós-verdade', a palavra do ano segundo a Universidade de Oxford*. 2016. Disponível em: <https://www.nexojornal.com.br/expresso/2016/11/16/O-que-%C3%A9-%E2%80%98p%C3%B3s-verdade%E2%80%99-a-palavra-do-ano-segundo-a-Universidade-de-Oxford>. Citado na página 17.

FAUSTINI, P.; COVõES, T. F. Fake news detection using one-class classification. In: *2019 8th Brazilian Conference on Intelligent Systems (BRACIS)*. [S.l.: s.n.], 2019. p. 592–597. ISSN: 2643-6256. Citado 2 vezes nas páginas 49 e 55.

FENNER, M. *Machine Learning with Python for Everyone*. [S.l.]: Addison-Wesley Professional, 2019. Citado 3 vezes nas páginas 23, 38 e 39.

FERREIRA, J. R. S.; LIMA, P. R. S.; SOUZA, E. D. de. Desinformação, infodemia e caos social: impactos negativos das fake news no cenário da covid-19. *Em Questão*, v. 27, p. 30–53, 2020. ISSN 1807-8893. Citado 2 vezes nas páginas 65 e 70.

G1. *Marielle engravidou aos 16? Foi casada com o traficante Marcinho VP? Ignorava as mortes de policiais? Não é verdade!* 2018. Disponível em: <https://g1.globo.com/e-ou-nao-e/noticia/marielle-engravidou-aos-16-foi-casada-com-o-traficante-marcinho-vp-ignorava-as-mortes-de-policiais-nao-ghtml>. Citado na página 19.

GABASIO, A. Comparison of optical character recognition (OCR) software. p. 95, 2013. Citado na página 46.

GALHARDI, C. P.; FREIRE, N. P.; MINAYO, M. C. de S.; FAGUNDES, M. C. M. Fact or fake? an analysis of disinformation regarding the covid-19 pandemic in brazil. 2020. Disponível em: <https://orcid.org/0000-0002-9038-9974>. Citado 2 vezes nas páginas 64 e 68.

GRAGNANI, J. Pesquisa inédita identifica grupos de família como principal vetor de notícias falsas no whatsapp. 2018. Disponível em: <https://www.bbc.com/portuguese/brasil-43797257>. Citado na página 18.

GUNDAPU, S.; MAMIDI, R. Transformer based automatic covid-19 fake news detection system. 2021. Disponível em: <https://www.worldometers.info/coronavirus/>. Citado 3 vezes nas páginas 64, 65 e 66.

GUPTA, A.; SUKUMARAN, R.; JOHN, K.; TEKI, S. Hostility detection and covid-19 fake news detection in social media. 2021. Disponível em: <https://pypi.org/project/tweet-preprocessor/>. Citado 2 vezes nas páginas 64 e 68.

HU, M.-K. Visual pattern recognition by moment invariants. *IRE Transactions on Information Theory*, v. 8, n. 2, p. 179–187, 1962. Citado na página 83.

HUTCHINSON, D. M. P. *Brilliant NLP*. [S.l.]: Pearson Education Limited, 2015. Citado 3 vezes nas páginas 22, 31 e 33.

INUWA-DUTSE, I. Towards combating pandemic-related misinformation in social media. 2021. Disponível em: <www.worldometers.info>. Citado 2 vezes nas páginas 65 e 72.

JAIN, A.; KASBE, A. Fake news detection. In: *2018 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS)*. [S.l.: s.n.], 2018. p. 1–5. ISSN: null. Citado 2 vezes nas páginas 49 e 53.

JALONICK, M. C. *House Democrats release more than 3,500 Facebook ads created by Russians*. 2018. Disponível em: <https://www.pbs.org/newshour/politics/house-democrats-release-more-than-3500-facebook-ads-created-by-russians>. Citado na página 19.

JOSHI, P. *Artificial Intelligence with Python*. [S.l.]: Packt Publishing, 2017. Citado 2 vezes nas páginas 32 e 33.

JúNIOR, J. H. de S.; RAASCH, M.; SOARES, J. C.; RIBEIRO, L. V. H. A. de S. Da desinformação ao caos: uma análise das fake news frente à pandemia do coronavírus (covid-19) no brasil. *Cadernos de Prospecção*, v. 13, p. 331 – 346, 2020. ISSN 2317-0026. Citado 2 vezes nas páginas 65 e 69.

KOLOSKI, B.; STEPIšNIK-PERDIH, T.; POLLAK, S.; BLAžšKRLJ, B. B. Identification of covid-19 related fake news via neural stacking. 2021. Disponível em: <https://gitlab.com/boshko.koloski/covid19-fake-news>. Citado 3 vezes nas páginas 64, 65 e 67.

KRISHNAN, S.; CHEN, M. Identifying tweets with fake news. In: *2018 IEEE International Conference on Information Reuse and Integration (IRI)*. [S.l.: s.n.], 2018. p. 460–464. ISSN: null. Citado na página 50.

LAZER, D. M. J.; BAUM, M. A.; BENKLER, Y.; BERINSKY, A. J.; GREENHILL, K. M.; MENCZER, F.; METZGER, M. J.; NYHAN, B.; PENNYCOOK, G.; ROTHSCHILD, D.; SCHUDSON, M.; SLOMAN, S. A.; SUNSTEIN, C. R.; THORSON, E. A.; WATTS, D. J.; ZITTRAIN, J. L. The science of fake news. v. 359, n. 6380, p. 1094–1096, 2018. ISSN 0036-8075, 1095-9203. Disponível em: <https://science.sciencemag.org/content/359/6380/1094>. Citado na página 79.

LEE, W.-M. *Python Machine Learning*. [S.l.]: Wiley, 2019. Citado 2 vezes nas páginas 40 e 41.

LI, X.; LI, S. Exploring text-transformers in aaai 2021 shared task: Covid-19 fake news detection in english. 2021. Disponível em: <https://github.com/archersama/3rd-solution-COVID19-Fake-News-Detection-in->. Citado 3 vezes nas páginas 64, 65 e 68.

LINDEN, S. van der; ROOZENBEEK, J.; COMPTON, J. Inoculating against fake news about covid-19. *Frontiers in Psychology*, Frontiers Media S.A., v. 11, 10 2020. ISSN 16641078. Citado 2 vezes nas páginas 64 e 69.

MAHID, Z. I.; MANICKAM, S.; KARUPPAYAH, S. Fake news on social media: Brief review on detection techniques. In: . [S.l.]: Institute of Electrical and Electronics Engineers Inc., 2018. ISBN 9781538671672. Citado na página 50.

MAHID, Z. I.; MANICKAM, S.; KARUPPAYAH, S. Fake news on social media: Brief review on detection techniques. In: *2018 Fourth International Conference on Advances in Computing, Communication Automation (ICACCA)*. [S.l.: s.n.], 2018. p. 1–5. ISSN: 2641-8134. Citado na página 63.

MAJUMDER ANUJ GUPTA, H. S. S. V. B. *Practical Natural Language Processing*. [S.l.]: O'Reilly Media, Inc., 2020. Citado 3 vezes nas páginas 31, 32 e 33.

MOLDEN, D. *NLP Business Masterclass: Driving peak performance with NLP*. Second edition. [S.l.]: Pearson, 2009. Citado na página 30.

MONTEIRO, R. A.; SANTOS, R. L. S.; PARDO, T. A. S.; ALMEIDA, T. A. de; RUIZ, E. E. S.; VALE, O. A. Contributions to the study of fake news in portuguese: New corpus and automatic detection results. In: VILLAVICENCIO, A.; MOREIRA, V.; ABAD, A.; CASELI, H.; GAMALLO, P.; RAMISCH, C.; OLIVEIRA, H. G.; PAETZOLD, G. H. (Ed.). *Computational Processing of the Portuguese Language*. Springer International Publishing, 2018. v. 11122, p. 324–334. ISBN 978-3-319-99721-6 978-3-319-99722-3. Disponível em: <http://link.springer.com/10.1007/978-3-319-99722-3_33>. Citado 2 vezes nas páginas 74 e 75.

MONTI, F.; FRASCA, F.; EYNARD, D.; MANNION, D.; BRONSTEIN, M. M. Fake news detection on social media using geometric deep learning. 2019. Disponível em: <http://arxiv.org/abs/1902.06673>. Citado 2 vezes nas páginas 48 e 52.

MORI, S.; NISHIDA, H.; YAMADA, H. *Optical Character Recognition*. 1 edition. ed. [S.l.]: Wiley-Interscience, 1999. ISBN 978-0-471-30819-5. Citado 3 vezes nas páginas 23, 24 e 25.

NAJAR, F.; ZAMZAMI, N.; BOUGUILA, N. Fake news detection using bayesian inference. In: *2019 IEEE 20th International Conference on Information Reuse and Integration for Data Science (IRI)*. [S.l.: s.n.], 2019. p. 389–394. ISSN: null. Citado 2 vezes nas páginas 49 e 55.

NETO, M.; GOMES, T. de O.; PORTO, F. R.; RAFAEL, R. de M. R.; FONSECA, M. H. S.; NASCIMENTO, J. Fake news no cenário da pandemia de covid-19. *Cogitare Enfermagem*, v. 25, 2020. ISSN 21769133. Citado 2 vezes nas páginas 65 e 70.

NEWNHAM, J. *Machine Learning with Core ML*. [S.l.]: Packt Publishing, 2018. Citado 3 vezes nas páginas 37, 38 e 39.

NYOW, N. X.; CHUA, H. N. Detecting fake news with tweets' properties. In: *2019 IEEE Conference on Application, Information and Network Security (AINS)*. [S.l.: s.n.], 2019. p. 24–29. ISSN: null. Citado 2 vezes nas páginas 49 e 54.

PATWA, P.; SHARMA, S.; PYKL, S.; GUPTHA, V.; KUMARI, G.; AKHTAR, S.; EKBAL, A.; DAS, A.; CHAKRABORTY, T. Fighting an infodemic: Covid-19 fake news dataset. 2020. Disponível em: <www.boomlive.in>. Citado 3 vezes nas páginas 50, 73 e 74.

PAULO, F. de S. *Veja o passo a passo da notícia falsa que acabou em tragédia em Guarujá*. 2018. Disponível em: <https://www1.folha.uol.com.br/cotidiano/2018/09/veja-o-passo-a-passo-da-noticia-falsa-que-acabou-em-tragedia-em-guaruja.shtml>. Citado na página 19.

PIERRI, F.; CERI, S. False news on social media: A data-driven survey. *SIGMOD Record*, Association for Computing Machinery, v. 48, p. 18–32, 6 2019. ISSN 01635808. Tem artigo impresso. Citado 3 vezes nas páginas 45, 50 e 63.

PREVITI, M.; RODRIGUEZ-FERNANDEZ, V.; CAMACHO, D.; CARCHIOLO, V.; MALGERI, M. Fake news detection using time series and user features classification. In: CASTILLO, P. A.; LAREDO, J. L. J.; VEGA, F. Fernández de (Ed.). *Applications of Evolutionary Computation*. [S.l.]: Springer International Publishing, 2020. (Lecture Notes in Computer Science), p. 339–353. ISBN 978-3-030-43722-0. Citado 2 vezes nas páginas 50 e 58.

PéREZ-ROSAS, V.; KLEINBERG, B.; LEFEVRE, A.; MIHALCEA, R. Automatic detection of fake news. p. 3391–3401, 2017. Disponível em: <http://arxiv.org/abs/1708.07104>. Citado 2 vezes nas páginas 50 e 59.

RASCHKA, V. M. S. *Python Machine Learning*. Third edition. [S.l.]: Packt Publishing, 2019. Citado 3 vezes nas páginas 40, 41 e 45.

RECUERO, R.; GRUZD, A.; RECUERO, R.; GRUZD, A. Cascatas de fake news políticas: um estudo de caso no twitter. n. 41, p. 31–47, 2019. ISSN 1982-2553. Disponível em: <http://www.scielo.br/scielo.php?script=sci_abstract&pid=S1982-25532019000200031&lng=en&nrm=iso&tlng=pt>. Citado 2 vezes nas páginas 48 e 53.

REIS, J. C. S.; CORREIA, A.; MURAI, F.; VELOSO, A.; BENEVENUTO, F. Explainable machine learning for fake news detection. In: *Proceedings of the 10th ACM Conference on Web Science - WebSci '19*. ACM Press, 2019. p. 17–26. ISBN 978-1-4503-6202-3. Disponível em: <http://dl.acm.org/citation.cfm?doid=3292522.3326027>. Citado 2 vezes nas páginas 49 e 53.

REIS, J. C. S.; CORREIA, A.; MURAI, F.; VELOSO, A.; BENEVENUTO, F. Supervised learning for fake news detection. v. 34, n. 2, p. 76–81, 2019. ISSN 1941-1294. Citado 3 vezes nas páginas 49, 55 e 76.

RICE, S. V.; NAGY, G.; NARTKER, T. A. *Optical Character Recognition: An Illustrated Guide to the Frontier*. Springer US, 1999. (The Springer International Series in Engineering and Computer Science). ISBN 978-0-7923-8492-2. Disponível em: <https://www.springer.com/gp/book/9780792384922>. Citado 2 vezes nas páginas 24 e 25.

ROVETTA, A.; BHAGAVATHULA, A. S. Global infodemiology of covid-19: Analysis of google web searches and instagram hashtags. *Journal of Medical Internet Research*, JMIR Publications Inc., v. 22, 8 2020. ISSN 14388871. Disponível em: <https://pubmed.ncbi.nlm.nih.gov/32748790/>. Citado 2 vezes nas páginas 65 e 71.

RUBIN, V. L.; CHEN, Y.; CONROY, N. J. Deception detection for news: Three types of fakes. *Proceedings of the Association for Information Science and Technology*, v. 52, p. 1–4, 2015. ISSN 23739231. Citado 2 vezes nas páginas 50 e 60.

SALAS, J. *Usuários transformam seus murais no Facebook em 'bolhas' ideológicas*. 2015. Disponível em: <https://brasil.elpais.com/brasil/2015/05/06/tecnologia/1430934202_446201.html>. Citado na página 17.

SAMPLE, C.; JUSTICE, C.; DARRAJ, E. Fake news: A method to measure distance from fact. In: *2018 IEEE International Conference on Big Data (Big Data)*. [S.l.: s.n.], 2018. p. 4443–4452. ISSN: null. Citado 3 vezes nas páginas 46, 49 e 54.

SANTIA, G. C.; WILLIAMS, J. R. BuzzFace: A news veracity dataset with facebook user commentary and egos. In: *Twelfth International AAAI Conference on Web and Social Media*. [s.n.], 2018. Disponível em: <https://aaai.org/ocs/index.php/ICWSM/ICWSM18/paper/view/17825>. Citado 2 vezes nas páginas 74 e 75.

SCHANTZ, H. F. *The history of OCR, optical character recognition*. [S.l.]: Recognition Technologies Users Association, 1982. OCLC: 9944667. ISBN 978-0-943072-01-2. Citado 2 vezes nas páginas 24 e 25.

SHEARER, E.; MATSA, K. E. *News Use Across Social Media Platforms 2018*. 2018. Disponível em: <https://www.journalism.org/2018/09/10/news-use-across-social-media-platforms-2018/>. Citado 2 vezes nas páginas 15 e 16.

SHU, K.; BERNARD, H. R.; LIU, H. Studying fake news via network analysis: Detection and mitigation. In: AGARWAL, N.; DOKOOHAKI, N.; TOKDEMIR, S. (Ed.). *Emerging Research Challenges and Opportunities in Computational Social Network Analysis and Mining*. Springer International Publishing, 2019, (Lecture Notes in Social Networks). p. 43–65. ISBN 978-3-319-94105-9. Disponível em: <https://doi.org/10.1007/978-3-319-94105-9_3>. Citado 2 vezes nas páginas 48 e 51.

SHU, K.; MAHUDESWARAN, D.; WANG, S.; LEE, D.; LIU, H. FakeNewsNet: A data repository with news content, social context and spatialtemporal information for studying fake news on social media. 2019. Disponível em: <http://arxiv.org/abs/1809.01286>. Citado 2 vezes nas páginas 74 e 76.

SILVA, F. C. D. da; VIEIRA, R.; GARCIA, A. C. Can machines learn to detect fake news? a survey focused on social media. In: . [S.l.]: Hawaii International Conference on System Sciences, 2019. Citado 2 vezes nas páginas 50 e 62.

SINGH, L.; BANSAL, S.; BODE, L.; BUDAK, C.; CHI, G.; KAWINTIRANON, K.; PADDEN, C.; VANARSDALL, R.; VRAGA, E.; WANG, Y. A first look at covid-19 information and misinformation sharing on twitter. 2020. Citado 2 vezes nas páginas 65 e 71.

SINGHAL, S.; SHAH, R. R.; CHAKRABORTY, T.; KUMARAGURU, P.; SATOH, S. SpotFake: A multi-modal framework for fake news detection. In: *2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM)*. [S.l.: s.n.], 2019. p. 39–47. ISSN: null. Citado 2 vezes nas páginas 48 e 51.

SMITH, R. History of the tesseract OCR engine: what worked and what didn't. p. 02, 2013. Citado na página 30.

TACCHINI, E.; BALLARIN, G.; VEDOVA, M. L. D.; MORET, S.; ALFARO, L. de. Some like it hoax: Automated fake news detection in social networks. 2017. Disponível em: <http://arxiv.org/abs/1704.07506>. Citado 2 vezes nas páginas 73 e 74.

TANDOC, E. C.; LIM, Z. W.; LING, R. Defining "fake news": A typology of scholarly definitions. *Digital Journalism*, Routledge, v. 6, p. 137–153, 2018. ISSN 2167082X. Tem artigo impresso para ler. Disponível em: <http://doi.org/10.1080/21670811.2017.1360143>. Citado 2 vezes nas páginas 50 e 61.

TEAGUE, M. Image analysis via the general theory of moments. *Journal of the Optical Society of America*, v. 70, p. 920–930, 1980. Citado na página 83.

THOMAS, A. *Natural Language Processing with Spark NLP*. [S.l.]: O'Reilly Media, Inc., 2020. Citado na página 30.

TRAYLOR, T.; STRAUB, J.; GURMEET; SNELL, N. Classifying fake news articles using natural language processing to identify in-article attribution as a supervised learning estimator. In: *2019 IEEE 13th International Conference on Semantic Computing (ICSC)*. [S.l.: s.n.], 2019. p. 445–449. ISSN: 2325-6516. Citado 2 vezes nas páginas 49 e 55.

TSCHIATSCHEK, S.; SINGLA, A.; RODRIGUEZ, M. G.; MERCHANT, A.; KRAUSE, A. Fake news detection in social networks via crowd signals. In: *Companion Proceedings of the The Web Conference 2018*. International World Wide Web Conferences Steering Committee, 2018. (WWW '18), p. 517–524. ISBN 978-1-4503-5640-4. Disponível em: <https://doi.org/10.1145/3184558.3188722>. Citado 2 vezes nas páginas 49 e 57.

VOSOUGHI, S.; ROY, D.; ARAL, S. The spread of true and false news online. *Science*, American Association for the Advancement of Science, v. 359, p. 1146–1151, 3 2018. ISSN 10959203. Disponível em: <http://science.sciencemag.org/>. Citado 2 vezes nas páginas 50 e 61.

WALT, S. van der; SCHöNBERGER, J.; NUNEZ-IGLESIAS, J.; BOULOGNE, F.; WARNER, J.; YAGER, N.; GOUILLART, E.; YU, T.; CONTRIBUTORS, t. scikit-image: Image processing in python. *PeerJ*, v. 2, 07 2014. Citado na página 84.

WANG, W. Y. "liar, liar pants on fire": A new benchmark dataset for fake news detection. 2017. Disponível em: <http://arxiv.org/abs/1705.00648>. Citado 2 vezes nas páginas 74 e 75.

WANI, A.; JOSHI, I.; KHANDVE, S.; WAGH, V.; JOSHI, R. Evaluating deep learning approaches for covid19 fake news detection. 2021. Citado 3 vezes nas páginas 64, 65 e 66.

YAMASAKI, I. Quantitative evaluation of print quality for optical character recognition systems. v. 8, n. 5, p. 371–381, 1978. ISSN 2168-2909. Citado na página 22.

YU, F. T. S.; JUTAMULIA, S. (Ed.). *Optical Pattern Recognition*. [S.l.]: Cambridge University Press, 1998. Citado na página 24.

ZANNETTOU, S.; SIRIVIANOS, M.; BLACKBURN, J.; KOURTELLIS, N. The web of false information: Rumors, fake news, hoaxes, clickbait, and various other shenanigans. *J. Data and Information Quality*, Association for Computing Machinery, New York, NY, USA, v. 11, n. 3, maio 2019. ISSN 1936-1955. Disponível em: <https://doi.org/10.1145/3309699>. Citado na página 45.

ZAREI, K.; FARAHBAKHSH, R.; CRESPI, N.; TYSON, G. A first instagram dataset on covid-19. v. 2020, p. 2–5, 2020. Disponível em: <http://arxiv.org/abs/2004.12226>. Citado 3 vezes nas páginas 74, 76 e 77.

ZHOU, X.; ZAFARANI, R. Network-based fake news detection: A pattern-driven approach. v. 21, n. 2, p. 48–60, 2019. ISSN 1931-0145. Disponível em: <https://doi.org/10.1145/3373464.3373473>. Citado 2 vezes nas páginas 48 e 52.