UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Detecção de postagens com informações falsas sobre a pandemia da COVID-19 na rede social Facebook

Leonardo Cappellesso Bigolin

Trabalho de Conclusão de Curso
MBA em Inteligência Artificial e Big Data

ICMC USP
SÃO CARLOS

UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Detecção de postagens com informações falsas sobre a pandemia da COVID-19 na rede social Facebook

Leonardo Cappellesso Bigolin

USP - São Carlos

2022

Leonardo Cappellesso Bigolin

Detecção de postagens com informações falsas sobre a pandemia da COVID-19 na rede social Facebook

Trabalho de conclusão de curso apresentado ao Departamento de Ciências de Computação do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, como parte dos requisitos para obtenção do título de Especialista em Inteligência Artificial e Big Data.

Área de concentração: Inteligência Artificial

Orientador: Prof. Dr. Ricardo Rodrigues Ciferri

USP - São Carlos

Agosto/2022

Bigolin, Leonardo
B594d Detecção de postagens com informações falsas sobre a pandemia da COVID-19 na rede social Facebook / Leonardo Bigolin; orientador Ricardo Rodrigues Ciferri. -- São Carlos, 2022.
60 p.

Trabalho de conclusão de curso (MBA em Inteligência Artificial e Big Data) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2022.

1. Conteúdo Falso. 2. Fake News. 3. Redes Sociais. 4. Facebook. 5. COVID-19. I. Rodrigues Ciferri, Ricardo. II. Detecção de postagens com informações falsas sobre a pandemia da COVID-19 na rede social Facebook.

RESUMO

Este trabalho de conclusão de curso aborda a detecção de informações falsas no Facebook, rede social que é muito utilizada para o compartilhamento de informação através de conteúdo multimídia. A intensidade de propagação de conteúdo falso se intensifica em épocas em que ocorrem acontecimentos de grandes proporções devido à maior busca de informações a respeito do assunto. Por esse motivo, este trabalho se preocupou em abordar sobre a pandemia da COVID-19, acontecimento que impactou todos os países do globo. Há diversas formas para tentar detectar informações falsas em redes sociais. Alguns métodos enfocam na análise de sentimentos, outros realizam uma análise temporal sobre a dinâmica da postagem e há também a possibilidade, como é o caso deste trabalho, de estudar o conteúdo da informação da postagem, seja uma foto, um vídeo, uma legenda em forma de texto ou todos esses formatos juntos. Dada a complexidade do assunto, para este trabalho optou-se por criar um modelo de aprendizado de máquina que permite a classificação de notícias sobre a COVID-19 como verdadeiras ou falsas utilizando-se apenas a legenda da postagem feita no Facebook. E por conta da escassez de dados sobre o assunto em português, foi necessário a construção de uma base de dados que contivesse um conjunto de legendas de postagens publicadas e a sua respectiva classificação, base de dados essa que está disponibilizada para futuras investigações sobre detecção de informações falsas no Facebook. Foram realizados testes experimentais sobre o modelo construído com essa base de dados construída e se obteve resultados de acurácia que variaram de 60% a 88% na detecção de postagens com informações falsas sobre a COVID-19 no Facebook.

Palavras-chave: informações falsas, conteúdo falso, fake News, COVID-19, pandemia, redes sociais, Facebook.

ABSTRACT

This Final Paper addresses the detection of false information at Facebook, social network that is very used to share information through multimedia content. The intensity of the propagation of false information increases when occurs major events due to the increased search for information on the subject. For this reason, this work is concerned with addressing the COVID-19 pandemic, an event that impacted all countries around the globe. There are several ways to try to detect false information on social networks. Some methods focus on the analysis of sentiments, others perform a temporal analysis on the dynamics of the post and there is also the possibility, as is the case of this work, to study the content of the post's information, be it a photo, a video, a caption in text or all these formats together. Given the complexity of the subject, for this work we chose to create a machine learning model that allows the classification of news about COVID-19 as true or false using only the text of the post made on Facebook. And due to the scarcity of data on the subject in Portuguese, it was necessary to build a database that contained a set of texts of published posts and their respective classification. This database is available for future investigations on detection of false information on Facebook. Experimental tests were carried out on the model, built with this database, and accuracy results were obtained that ranged from 60% to 88% in detection of posts with false information about COVID-19 on Facebook.

Keywords: fake news, social network, Facebook, COVID-19.

LISTA DE ABREVIATURAS E SIGLAS

AM	Aprendizado de Máquina
AUC	*Area Under the Curve*
BOW	*Bag of Words*
CSV	Comma Separated Value
FN	Falso Negativo
FP	Falso Positivo
GB	*Gigabytes*
GHz	*Gigahertz*
HD	*Hard Drive*
IA	Inteligência Artificial
KNN	*K Nearest Neighbor*
MHz	Megahertz
NLTK	*Natural Language Toolkit*
OCR	*Optical Character Recognition*
PLN	Processamento de Linguagem Natural
ROC	*Receiver Operating Characteristic*
SVM	*Support Vector Machine*
TB	*Terabytes*
VN	Verdadeiro Negativo
VP	Verdadeiro Positivo

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

1.1 Considerações Iniciais

O objetivo deste capítulo inicial é contextualizar o problema de postagens e detecção de informações falsas em redes sociais (seção 1.2), demonstrar quais são as motivações para o desenvolvimento do presente trabalho de conclusão de curso (seção 1.3) e, por fim, evidenciar para o leitor quais sãos as questões de pesquisa que tentarão ser respondidas no trabalho e os seus objetivos (seção 1.4).

1.2 Contextualização do Problema

Há anos observa-se o crescimento no uso de mídias sociais, cujo conceito abrange desde *e-mails* e *blogs* até plataformas como Facebook, Instagram, WhatsApp, Youtube, Twitter e TikTok. Esse crescimento é verificado tanto nos usuários mais jovens quanto nos mais idosos. Por permitirem o compartilhamento de textos e mídias, as redes sociais são utilizadas não apenas para conectar virtualmente os usuários, mas também para transmitir informações. Dessa forma, muitos indivíduos passaram a consumir notícias usando basicamente apenas as redes sociais.

Porém, como o conteúdo das postagens feitas nas plataformas de redes sociais não passam por um filtro imediato, há a possibilidade de qualquer usuário publicar as postagens que desejar, sem verificação da sua veracidade, diferentemente do que ocorre em empresas jornalísticas sérias. Essa liberdade das redes sociais e o consumo de notícias por meio das mesmas resulta na propagação de postagens com informações falsas.

Como o lucro das empresas criadoras de redes sociais é apoiado basicamente em anúncios de outras empresas, a maioria dessas redes sociais atuais possui um algoritmo que capta os principais interesses de cada usuário e filtra os anúncios e informações que são disponibilizadas aos usuários em conformidade com o seu interesse. Isso faz com que usuários que possuam interesses parecidos sejam agrupados e tenham acesso a informações parecidas. Por esse motivo, a disseminação de notícias falsas se torna muito preocupante, uma vez que esse tipo de notícia possui conteúdo apelativo com a intenção de captar a atenção de um maior número de indivíduos e tem tendência de ser disponibilizada para indivíduos que se identificam com a informação que está sendo passada gerando, dessa forma, polarização na população.

Uma vez que verificar a veracidade de notícias postadas em redes sociais exige um custo humano elevado, tem sido desenvolvidos algoritmos que, por meio do Processamento de

Linguagem Natural (PLN) e uso de Aprendizado de Máquina (AM) supervisionado, extraem a informação textual contida nas notícias postadas e tentam classifica-las como verdadeiras ou falsas com a intenção de controlar a postagem ou mitigar os efeitos negativos das informações falsas.

Dessa forma, esta pesquisa de Trabalho de Conclusão de Curso será uma continuidade da pesquisa de Mestrado de Mateus Oliveira Cabral (CABRAL, 2021), orientado pelo Prof. Dr. Ricardo Rodrigues Ciferri, que teve como objetivo a detecção de postagens com informações falsas sobre a pandemia da COVID-19 na rede social Instagram. De forma distinta, o presente trabalho será focado na rede social Facebook. Para isso, serão utilizadas as técnicas de PLN e os algoritmos de AM para classificação de postagens falsas que Mateus Oliveira Cabral utilizou em sua pesquisa de mestrado para o Instagram e adaptá-los para o desenvolvimento de uma aplicação para classificação de notícias com informações falsas sobre a pandemia da COVID-19 postadas no Facebook e, com isso, espera-se contribuir para diminuir os esforços para a rotulação de notícias falsas.

1.3 Motivação

Nos últimos anos, a propagação de notícias falsas (*Fake News*) se tornou um enorme problema devido ao fato de despertarem curiosidade nas pessoas pelo seu conteúdo apelativo e, consequentemente, se espalhar de forma rápida para um grande número de indivíduos. De acordo com o relatório de notícias digitais de 2021 do *Reuters Institute* as preocupações com informações falsas aumentaram naquele ano e o Brasil foi o país mais preocupado com o assunto. Dessa forma, detectar com presteza e eficácia a postagem de notícias falsas se torna fundamental uma vez a propagação generalizada de notícias falsas pode culminar em problemas como a perda de confiança generalizada nos meios de informação e prejuízos causados por decisões tomadas por indivíduos que foram baseadas em informações que não refletem a realidade.

Para agravar a situação, passamos pela crise da COVID-19, que mostrou o impacto que a reprodução descontrolada de notícias falsas pode causar na área da saúde e no bem-estar da população. A conjectura se torna ainda mais sensível pelo fato de a própria pandemia aumentar em frequência e em quantidade o consumo de notícias, expandindo ainda mais a velocidade de reprodução de informações. No entanto, existem plataformas que são utilizadas para a verificação de fatos que, em geral, trabalham de forma independente da mídia jornalística. Alguns exemplos de plataformas que se dedicam a essa finalidade são as agências Lupa, Aos

Fatos, CNN Brasil, Covid Verificado, Estadão, Folha de São Paulo, Portal G1, Projeto Comprova, Istoé, Uol Notícias, revista Veja e Polifact (CABRAL, 2021).

Diante desse cenário de propagação de grandes volumes de dados sendo enviados ininterruptamente, a verificação de todos os fatos por parte dos seres humanos se torna inviável, e deve-se desenvolver tecnologias capazes de ajudar na verificação da veracidade de notícias propagadas nas redes sociais. Nesse âmbito já existem diversas pesquisas sendo desenvolvidas com a finalidade de desenvolver um modelo computacional que faça ou ajude nessa verificação de forma automática. O método de verificação pode variar para cada projeto. Algumas se utilizam dos recursos linguísticos utilizados para a escrita da notícia, outras podem utilizar o caminho percorrido pela notícia quando foi propagada pelos usuários, ou até mesmo reação dos usuários após a postagem (CABRAL, 2021).

O Facebook é visto como a principal canal para disseminação de notícias falsas do mundo, conforme aponta o Digital News Report 2021, relatório do Instituto Reuters, e essa rede social conta com ferramentas capazes de aumentar ainda mais a dispersão de informações, como exemplo o uso de robôs ou retransmissão de notícias em cascata.

1.4 Questões de Pesquisa e Objetivos

Este Trabalho de Conclusão de Curso consiste em uma complementação do tema de pesquisa desenvolvido por Mateus Oliveira Cabral, orientado pelo Prof. Dr. Ricardo Rodrigues Ciferri, em sua dissertação de mestrado, que teve como objetivo a detecção de postagens com informações falsas sobre a pandemia da Covid-19 na rede social Instagram.

No entanto, a rede social de estudo para o presente trabalho será o Facebook. Dessa forma, o principal objetivo desta monografia é a detecção de postagens com notícias falsas a respeito do COVID-19 publicadas na rede social Facebook. Com isso, espera-se poder classificar as postagens utilizando, de forma adaptada, as soluções de (CABRAL, 2021) com grau aceitável de eficácia, as notícias postadas no Facebook em duas categorias: verdadeiras ou falsas. Dessa forma, formulou-se a seguinte questão de pesquisa para nortear este projeto:

> "*O algoritmo de classificação utilizado para detecção de notícias falsas no Instagram é também capaz de detectar a veracidade de notícias postadas na rede social Facebook com eficácia aceitável?*"

Com essa questão de pesquisa, pode-se definir os seguintes objetivos para o desenvolvimento deste trabalho:

- Modelar um esquema para o armazenamento dos conteúdos do Facebook que facilite o processamento das postagens e a exportação delas para um banco de dados.
- Utilizar os algoritmos de classificação usados por Mateus Oliveira Cabral para detecção de notícias falsas no Instagram e aplicá-los para o Facebook.
- Verificar se o desempenho dos algoritmos traz resultados de eficácia satisfatória ao serem aplicados ao Facebook.

2 FUNDAMENTAÇÃO TEÓRICA

2.1 Considerações Iniciais

A intenção do desenvolvimento de um algoritmo que detecte informações falsas em redes sociais como o Facebook é que essa atividade seja feita de forma automatizada, com nenhuma ou com o mínimo de intervenção humana possível. Dessa forma, para que seja possível o entendimento do processo de detecção de informações falsas, as tecnologias envolvidas devem ser de conhecimento do desenvolvedor ou usuário.

A detecção de notícias pelas máquinas passa por análises dos textos que foram publicados nas redes sociais, dessa forma, deve ser possível para a máquina conseguir extrair significado dessas postagens. Isso é feito através do Processamento da Linguagem Natural, um dos ramos da IA que será visto com mais detalhes no subtópico 2.2 deste capítulo.

Uma vez que a máquina consegue extrair informações de um grupo de dados textuais é possível aplicar técnicas de aprendizado de máquina com o objetivo de classificar as informações postadas nas redes sociais em verdadeiras ou falsas de forma automatizada. Esse tema será aprofundado no subtópico 2.3 do presente capítulo.

No subtópico 2.4 será abordado como a publicação de informações falsas se tornou um problema no âmbito das redes sociais e algumas características presentes nesse tipo de postagem que ajudam na classificação de notícias falsas.

Por fim, como este trabalho se preocupa na detecção de notícias falsas no Facebook, no subtópico 2.5 serão abordadas algumas características desta rede social.

2.2 Processamento de Linguagem Natural

O Processamento de Linguagem Natural (PLN) é um ramo da Inteligência Artificial, da Ciência da Computação e da Linguística que busca processar e analisar a linguagem natural tanto na forma escrita como na forma falada. Dessa forma, o PLN basicamente tenta fazer com que as máquinas sejam capazes de compreender a linguagem natural e realizar tarefas envolvendo essa linguagem (VAJJALA, MAJUMDER, *et al.*, 2020; VASILIEV, 2020)

Os assistentes de voz como Apple Siri, o assistente do Google e a Amazon Alexia talvez sejam as aplicações mais conhecidas que se utilizam do PLN, no entanto o seu uso se estende também, por exemplo, nos campos de tradução, classificação e resumo de textos, extração de informações, recuperação de informações em documentos, modelagem de perguntas e de tópicos (VAJJALA, MAJUMDER, *et al.*, 2020).

Conforme define VAJJALA et al. (2020), a linguagem natural "é um sistema estruturado de comunicação que envolve combinações complexas de seus componentes constituintes, como caracteres, palavras, sentenças, etc.". Dada a complexidade do campo, para desenvolver aplicações de PLN são necessárias grandes quantidades de dados textuais ou falados (que são comumente chamados de *corpus*) e a ajuda de tecnologias de aprendizado de máquina assim como de análises estatísticas e de definições de regras gramaticais para dar à máquina condições de detectar padrões na linguagem e retornar um resultado aceitável.

Um exemplo da utilização de técnicas de aprendizado de máquina supervisionado para classificação de dados textuais consiste na categorização de artigos de notícia de acordo com o seu assunto, como esportes ou política, por exemplo. Por outro lado, também é possível treinar a máquina para realizar técnicas de regressão em dados textuais. Um exemplo da aplicação dessa última técnica seria a previsão do preço de uma ação com base na discussão a respeito da mesma em redes sociais. Além disso, técnicas de aprendizado não supervisionado também podem ser utilizadas, por exemplo, para agrupar dados de texto (VAJJALA, MAJUMDER, *et al.*, 2020).

Antes de submeter o *corpus* a alguma atividade em aprendizado de máquina podem ser utilizadas diversas técnicas de pré-processamento. Essas técnicas podem ser aplicadas sobre todo o texto, sobre as sentenças do texto e/ou sobre suas palavras separadamente. Inicialmente segmenta-se o texto em sentenças e as sentenças em palavras. A análise é feita sobre essas partes do texto separadamente (sentenças ou palavras) e cada parte desse texto é comumente chamada de *token*, por isso essa atividade de dividir o texto em sentenças ou palavras é chamada de tokenização. Em seguida, pode-se realizar normalizações nas sentenças desse texto, como por exemplo, a remoção da pontuação e a correção de erros gramaticais e abreviaturas. Em nível de palavra, pode-se realizar o processo de *stemização*, que consiste na remoção de sufixos e a redução de cada palavra a uma forma base que gera todas as demais variantes dessa palavra, e de lematização, que consiste no mapeamento de todos os diferentes formatos de uma palavra a uma palavra base ou lema, ou seja, está relacionada à derivação linguística das palavras. Apesar de stemização e lematização parecerem similares, estes não o são. Por exemplo ao aplicar a stemização à palavra "amiga" se obteria a "amig", mas ao se aplicar a lematização à essa palavra se obteria a palavra "amizade" (VAJJALA, MAJUMDER, *et al.*, 2020).

Por fim, uma vez realizada a etapa de pré-processamento, diversas técnicas podem ser aplicadas no *corpus* a depender da finalidade desejada. Uma dessas técnicas é a vetorização de palavras, que consiste em atribuir um valor inteiro único a cada palavra do vocabulário. Isso torna possível a realização de operações matemáticas e cálculos estatísticos sobre os

documentos de texto. Exemplos de técnicas de vetorização são a codificação *one-hot*, saco de palavras (*bag-of words*), saco de N-*Grams* e TF-IDF (CABRAL, 2021).

2.3 Aprendizado de Máquina

Com a abundância de dados estruturados e não estruturados que a humanidade tem à disposição na atual era da tecnologia moderna, foi possível o desenvolvimento de um ramo da IA no qual algoritmos extraem automaticamente conhecimento desses dados para fazer predições e, ao mesmo tempo, melhorar o seu desempenho para predições futuras. Dessa forma, é possibilitado às máquinas a aprender a partir dos dados que são colocados à sua disposição (RASHKA e MIRJALILI, 2019).

O aprendizado de máquina tem sido utilizado cada vez mais em pesquisas na ciência da computação e já tem um papel importante no dia-a-dia das pessoas como por exemplo nos filtros de *spam*, em *software* de reconhecimento de texto ou voz, em *sites* de busca na *Internet* e em até programas de jogos de xadrez (RASHKA e MIRJALILI, 2019).

Existem três principais tipos de aprendizado de máquina: aprendizado supervisionado, aprendizado não supervisionado, aprendizado semi-supervisionado e aprendizado por reforço. No aprendizado supervisionado, o sistema aprende a fazer previsões utilizando-se de dados de treinamento cujo rótulo de classificação desejado já é conhecido e, posteriormente, utiliza esse conhecimento adquirido para fazer previsões de novos casos em que o rótulo de classificação é desconhecido a priori. Um exemplo de uso de aprendizado supervisionado é a utilização de uma série de *e-mails* pré-classificados como “spam” ou “não spam” para treinar um sistema para classificar novos e-mails em uma dessas duas categorias (RASHKA e MIRJALILI, 2019).

Já no aprendizado não supervisionado, utiliza-se dados em que o rótulo de classificação é desconhecido e o sistema tenta gerar agrupamentos desses dados de acordo com a sua similaridade. Um exemplo de aplicação do aprendizado não supervisionado seria a exploração de características em comum de um grupo de consumidores levando em consideração os seus interesses em comum. É importante mencionar que é possível a utilização de dados que estão parcialmente rotulados e fazer uma combinação de métodos supervisionados e não supervisionados, isso é chamado de aprendizado semi-supervisionado (FENNER, 2019; RASHKA e MIRJALILI, 2019).

Por fim, no aprendizado por reforço o agente de aprendizagem capta dados do ambiente e recebe recompensas positivas ou negativas conforme vai se aproximando ou se distanciando

de seu objetivo, dessa forma, o sistema aprende qual é a estratégia na qual a recompensa, no final do processo, seja a maior possível (FENNER, 2019).

Como o intuito do presente trabalho é a classificação de notícias postadas no Facebook entre verdadeiras e falsas, utilizar-se-á o método de aprendizagem supervisionada, e nos seguintes tópicos serão discutidos os principais algoritmos utilizados neste tipo de aprendizagem, a importância do uso de dados de treinamento e dados de teste e as principais medidas de desempenho dos resultados obtidos por algoritmos de aprendizagem supervisionada.

2.3.1 Tarefa de Classificação

Nesta subseção são destacados alguns algoritmos de aprendizado de máquina supervisionado utilizados em tarefas de classificação.

2.3.1.1 Support Vector Machine (SVM)

A Máquina Vetorial de Suporte ou *Support Vector Machine* (SVM) é um modelo muito utilizado no aprendizado de máquina e pode ser considerado uma extensão do *perceptron*. Pode ser utilizado tanto para a resolução de problemas linearmente separáveis como no de problemas não linearmente separáveis (não lineares). O seu objetivo é determinar um hiperplano em um espaço com N dimensões (onde N é o número de atributos) que melhor classifica o conjunto de dados. Os vetores de suporte são dados que estão localizados próximos a esse hiperplano e o classificador SVM tenta maximizar a distância (margem) entre esses vetores de suporte de maneira que a divisão entre as classes definida pelo hiperplano seja a melhor possível (RASHKA e MIRJALILI, 2019).

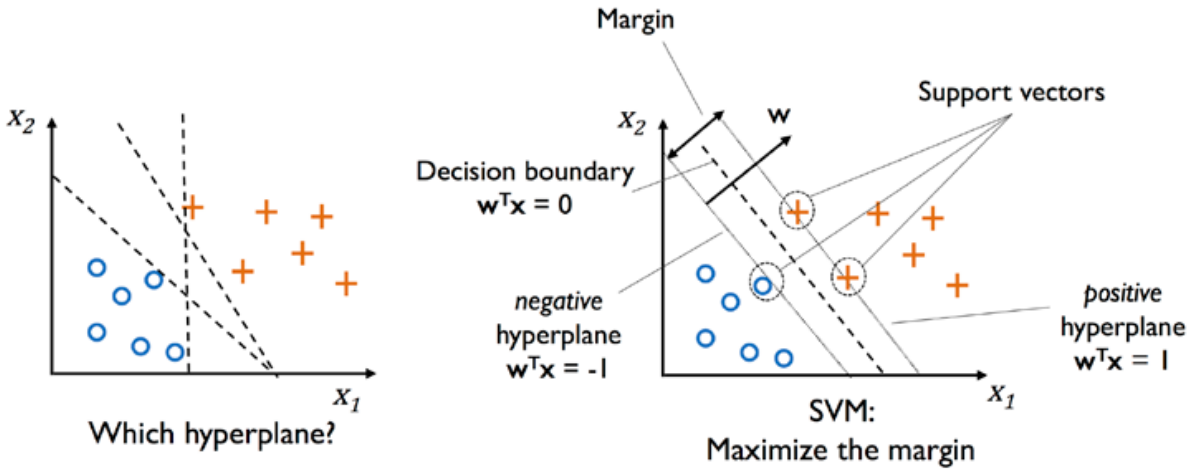

Figura 1 - Esquema básico de funcionamento do classificador SVM
Fonte: *(RASHKA e MIRJALILI, 2019)*

Observa-se que vetores de suporte que possuem grandes margens tendem a apresentar menores erros de generalização, enquanto que modelos com vetores de suporte com margens pequenas estão mais suscetíveis a apresentar *overfitting* de dados (RASHKA e MIRJALILI, 2019).

2.3.1.2 Árvores de Decisão

As Árvores de Decisão são algoritmos de aprendizado de máquina supervisionado muito utilizados em problemas de classificação e podem ser aplicados tanto em variáveis contínuas como em variáveis categóricas. Esse tipo de algoritmo funciona identificando as variáveis mais significantes para realizar a classificação desejada e, em seguida, divide o conjunto de dados em conjuntos de dados mais homogêneos possíveis (SULLIVAN, 2017).

Como exemplo pode-se citar o uso das Árvores de Decisão para a classificação de flores. Em um primeiro momento o algoritmo pode dividir o grupo de dados em dois: de um lado as flores que possuírem largura de pétala maior que 0,75 cm e de outro aquelas que possuírem valor de largura de pétala menor ou igual a 0,75 cm. Esse primeiro nó é chamado de nó raiz. A partir da divisão feita pelo nó raiz, novas divisões podem ser realizadas ao serem definidos novos nós. Como exemplo de um novo nó, o algoritmo poderia dividir o grupo de pétalas que possuem largura de pétala maior que 0,75 cm em outros dois novos grupos: aquelas pétalas que possuírem comprimento maior que 4,75 cm e aquelas que não se enquadrarem nessa categoria. Novos nós podem ser definidos até alcançar a classificação desejada. Abaixo evidencia-se o esquema do exemplo citado e as classificações obtidas (RASHKA e MIRJALILI, 2019).

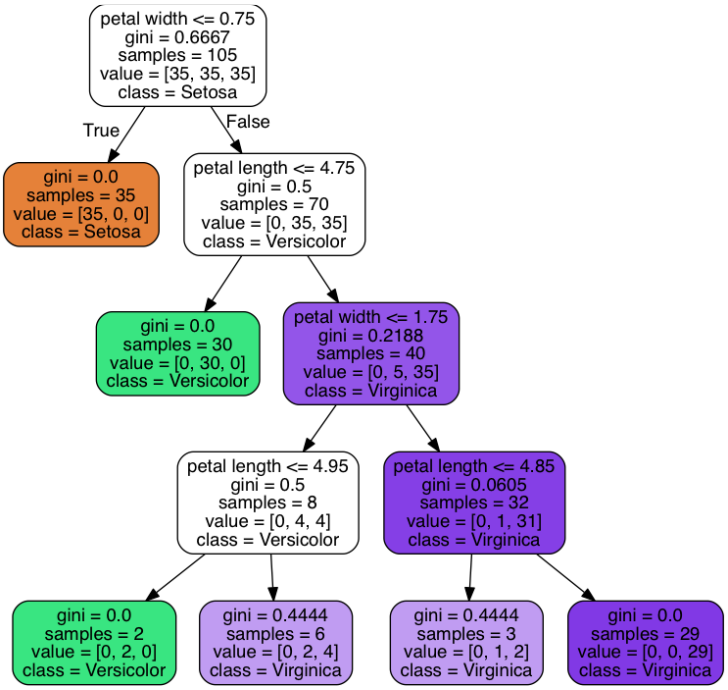

Figura 2 - Exemplo de classificação de flores utilizando Árvores de Decisão
Fonte: *(RASHKA e MIRJALILI, 2019)*

Entre as vantagens citadas para as Árvores de Decisão estão a facilidade de entendimento, a sua utilidade para exploração de dados, a pequena necessidade de limpeza dos dados, sua versatilidade e o fato de ser um método não paramétrico. Por outro lado, como desvantagens pode-se citar a sua tendência em apresentar *over fitting* sobre os dados e frequentemente não apresenta resultados satisfatórios quando utilizado sobre dados contínuos (SULLIVAN, 2017).

2.3.1.3 Florestas Aleatórias

As Florestas Aleatórias ou *Random Forest* é uma técnica *emsemble* que combina múltiplas Árvores de Decisão. Cada Árvore de Decisão é treinada utilizando um subconjunto selecionado aleatoriamente no conjunto de dados e, para realizar previsões, obtém-se a previsão de cada Árvore de Decisão e observa-se a classificação que obteve mais votos. As Florestas Aleatórias normalmente apresentam melhor desempenho de generalização devido à aleatoriedade de formação dos conjuntos de treinamento de cada Árvore de Decisão. Outra vantagem é a pouca sensibilidade a *outliers* e o fato de não necessitar de muitos ajustes de parâmetros, sendo o número de Árvores de Decisão, tipicamente, o único parâmetro a ser definido (RASHKA e MIRJALILI, 2019).

2.3.1.4 AdaBoost e Gradient Boosting

O nome *AdaBoost* (*Adaptive Boosting)* e o *Gradient Boosting* são técnicas que combinam dois ou mais classificadores com a finalidade de melhorar o desempenho do sistema. Primeiramente é treinado um classificador qualquer, por exemplo o SVM, em seguida, observa-se em quais instâncias o classificador não obteve desempenho adequado de classificação e altera-se o peso relativo dos classificadores nessas instâncias. Ocorre o treinamento de um outro classificador e utiliza-se os pesos alterados nas instâncias mal classificadas pelo primeiro classificador para fazer novas previsões a respeito dessas instâncias (RASHKA e MIRJALILI, 2019).

2.3.2 Treinamento e Teste

Ao implementar um algoritmo de aprendizado de máquina, é desejável determinar o desempenho deste algoritmo ao classificar os dados fornecidos para o aprendizado de máquina afim de conferir se o resultado apresentado possui acurácia aceitável. Além disso, deve-se garantir que o algoritmo seja capaz de classificar com precisão aceitável dados diferentes dos apresentados na etapa de treinamento, ou seja, novos dados (CABRAL, 2021).

Por esse motivo, os dados fornecidos para realizar o aprendizado de máquina são divididos em dois conjuntos: um conjunto de treinamento e um conjunto de teste. Como o próprio nome sugere, o conjunto de treinamento é utilizado para treinar o modelo enquanto que o conjunto de teste é utilizado para testá-lo. Quando a taxa de acertos nos dados de treinamento é alta, mas a taxa de acertos nos dados de teste não alcança o mesmo desempenho, diz-se que o modelo se encontra sobre ajustado para os dados de treinamento. Além disso, diz-se que o modelo foi capaz de generalizar quando atinge bom desempenho na classificação de novos casos (CABRAL, 2021; FENNER, 2019).

Com o intuito de reduzir o viés das medidas de desempenho dos algoritmos de aprendizado de máquina, frequentemente utiliza-se a técnica de *k-fold cross validation*. Esta técnica consiste em dividir o conjunto total de dados disponíveis em k subconjuntos sem reposição e de mesmo tamanho e, em seguida, utiliza-se um desses subconjuntos para realizar o teste do modelo e os k-1 restantes para o seu treinamento. Esse procedimento é repetido k vezes de maneira que se obtém k grupos de medidas de desempenho. Por fim, calcula-se a média do desempenho do modelo utilizando as medidas de desempenho obtidas para cada subgrupo (RASHKA e MIRJALILI, 2019).

2.3.3 Principais Medidas de Desempenho

Existem diversas maneiras de medir o desempenho dos algoritmos de aprendizado de máquina. Nesta seção serão abordadas a matriz de confusão e as medidas de desempenho acurácia,

2.3.3.1 Matriz de Confusão

A matriz de confusão é uma matriz com duas linhas e duas colunas que evidencia o desempenho de um algoritmo em aprendizado ao mostrar o número de Verdadeiros Positivos (VP), Verdadeiros Negativos (VN), Falsos Positivos (FP) e Falsos Negativos (FN). Ocorre a contabilização de um VP quando o algoritmo prevê corretamente um dado como membro da classe e ocorre a contabilização de um VN quando o algoritmo prevê corretamente um dado como não pertencente à classe. Alternativamente, ocorre a contabilização de um FP quando o algoritmo prevê erroneamente um dado como pertencente a classe e ocorre a contabilização de um FN quando o algoritmo prevê erroneamente um dado como não pertencente a classe (CABRAL, 2021; RASHKA e MIRJALILI, 2019).

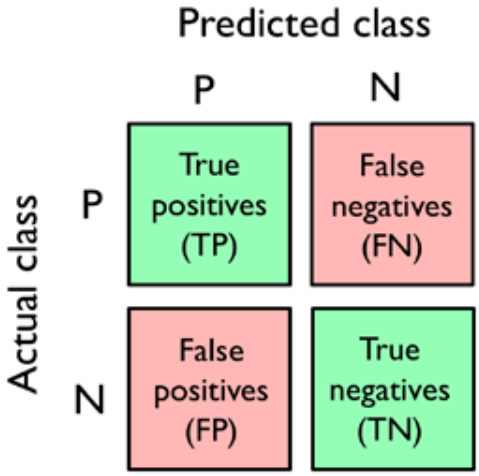

Figura 3 - Representação de uma Matriz de Confusão

Fonte: *(RASHKA e MIRJALILI, 2019)*

Com a representação da matriz de confusão, é possível realizar uma análise mais detalhada do que apenas a contagem de classificações corretas ou incorretas o que é muito útil no caso de conjuntos de dados desbalanceados, que ocorrem quando há uma grande diferença entre a quantidade de amostras em diferentes classes (CABRAL, 2021).

2.3.3.2 Acurácia

A acurácia é o número de previsões corretas dividido pelo número total de previsões realizadas, sendo calculada pela fórmula abaixo:

$$Acurácia = \frac{VP + VN}{VP + VN + FP + FN}$$

2.3.3.3 Precisão

A precisão mede o desempenho das classificações realizadas pelo algoritmo no âmbito dos dados que foram ou deveriam ter sido classificados como pertencentes à classe. Ou seja, leva em consideração apenas os Verdadeiros Positivos ou os Falsos Positivos. Essa métrica é muito utilizada, por exemplo, para aplicações como o filtro de *spam*. Para esse tipo de aplicação, o envio equivocado de um *e-mail* para a caixa de *spam* é muito mais danoso do que a manutenção de um *spam* na caixa de entrada. Dessa forma, a presença de Falsos Positivos é menos desejada do que a de Falsos Negativos (CABRAL, 2021).

O cálculo da Precisão é dado pela fórmula abaixo:

$$Precisão = \frac{VP}{VP + FP}$$

2.3.3.4 Revocação

Diferentemente da Precisão, a Revocação leva em consideração apenas o número de casos enquadrados como Verdadeiros Positivos e o número de casos classificados como Falsos Negativos. A Revocação é muito utilizada quando a presença de Falsos Negativos é indesejada como exemplo em exames de detecção de doenças (CABRAL, 2021).

O cálculo da Revocação é dado pela fórmula abaixo:

$$Revocação = \frac{VP}{VP + FN}$$

2.3.3.5 Pontuação F1

Com a finalidade de balancear os lados negativos e positivos da Revocação e da Precisão, utiliza-se uma combinação das duas medidas que é chamada de Pontuação F1. Normalmente o cálculo da Pontuação F1 é feita por meio da média harmônica da Revocação e da Precisão, conforme descreve a fórmula abaixo. A Pontuação F1 se mostra menos suscetível de distorções diante de *outliers* (RASHKA e MIRJALILI, 2019; CABRAL, 2021).

$$F1 = 2 * \frac{Revocação * Precisão}{Revocação + Precisão}$$

2.3.3.6 Curva ROC e AUC

O nome da curva ROC vem do inglês *Receiver Operating Characteristic* (ROC) e é uma ferramenta muito utilizada para selecionar modelos para classificação levando em consideração o desempenho em relação às taxas de Falsos Positivos e Verdadeiros Positivos obtida. A diagonal de uma curva ROC pode ser interpretada como uma situação de classificação aleatória de forma que modelos que se situam abaixo desse limiar são considerados modelos piores que um modelo de escolhas randômicas. Um classificador perfeito obteria Taxa de Verdadeiros Positivos igual a 1 e Taxa de Falsos Positivos igual a zero e se posicionaria no canto superior esquerdo do gráfico (RASHKA e MIRJALILI, 2019).

Ao variar o limiar de decisão do classificador e computando as taxas de Verdadeiros Positivos e de Falsos Positivos para cada caso, traça-se a respectiva curva ROC. A área embaixo da curva ou *Area Under the Curve* (AUC), que pode variar entre 0 e 1 é utilizada para caracterizar o desempenho do modelo de classificação (RASHKA e MIRJALILI, 2019).

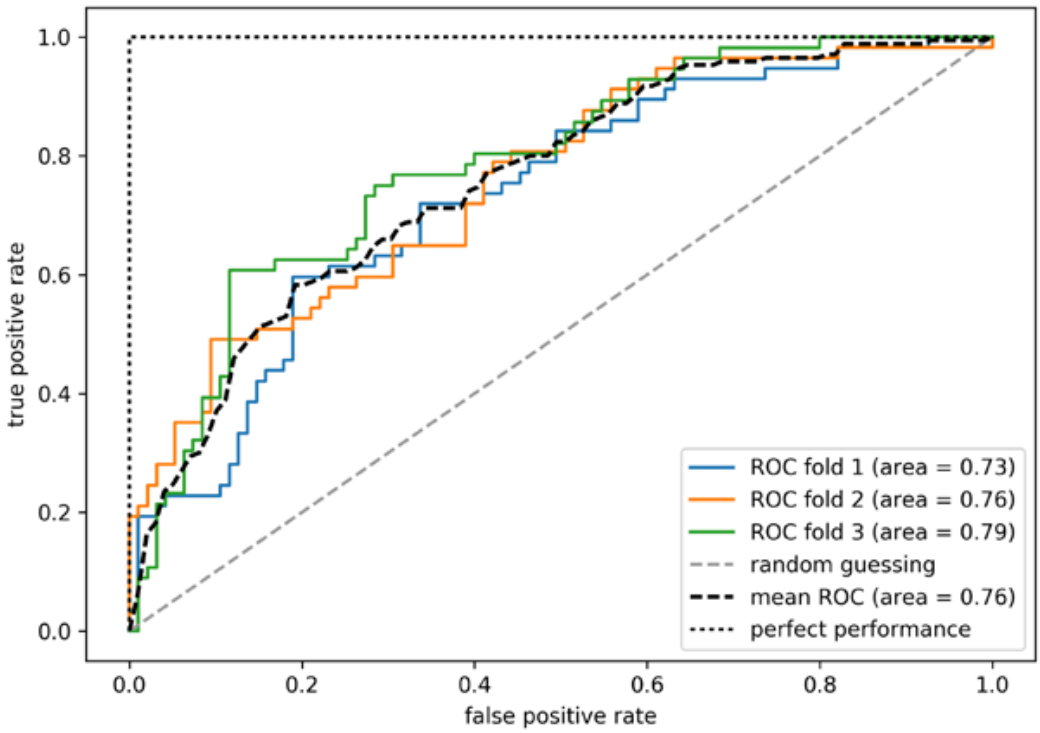

Figura 4 - Exemplo de curva ROC
Fonte: *(RASHKA e MIRJALILI, 2019)*

2.4 Informações Falsas em Redes Sociais

O acesso generalizado à *Internet* e às redes sociais ocasionou uma transformação na forma com que grande parte das notícias são consumidas pela população. Atualmente, observa-

se uma redução no consumo de notícias por meios impressos, como revistas ou jornais, e um grande aumento no consumo de informações por meios digitais, principalmente pela população mais jovem. Entretanto, essa mudança de paradigma possibilitou que pessoas ou organizações que não têm o mesmo conhecimento ou mesmo compromisso ético da maioria das empresas de jornalismo disponibilizassem informações em um meio em que a propagação das informações é muito menos custosa e muito mais rápida do que era quando a principal maneira de divulgação de informações era feita por meios físicos (GIANSIRACUSA, 2021).

Essa facilidade de propagação de informações, a considerável anonimidade que a *Internet* fornece ao usuário mal-intencionado e o uso de robôs de automação culminou em um aumento generalizado na divulgação de notícias falsas ou *Fake News* em meios eletrônicos. Esse fenômeno resulta em diversos problemas para o indivíduo e para a sociedade, pois, além de fazer com algumas pessoas tomem decisões baseadas em informações falsas, leva deão descrédito todas as notícias consumidas, uma vez que o indivíduo tende a não conseguir mais discernir o que é verdadeiro ou falso (GIANSIRACUSA, 2021).

Shu et al. (2020) exemplifica possíveis intenções na divulgação de informações falsas: persuadir pessoas a apoiar indivíduos, grupos, ideias ou ações futuras; persuadir pessoas a se oporem a indivíduos, grupos, ideias ou ações futuras; produzir reações emocionais (medo, ódio ou alegria) em indivíduos ou grupos; prevenir constrangimento ou persecução penal ao ser acreditado; exagerar a seriedade de algo dito ou feito; criar confusão a respeito de incidentes passados ou atividades.

A propagação de notícias falsas nas redes sociais se mostrou motivo de grandes preocupações para os governantes desde as eleições presidenciais dos Estados Unidos em 2016. Mais recentemente, na pandemia da COVID-19, a propagação de informações falsas a respeito do tema era tão grande que foi cunhado o termo “infodemia” (do inglês *infodemic*) para descrever o fenômeno. Relatos de indivíduos tomando decisões que contrariam completamente a ciência e o bom senso por influência de informações falsas postadas em redes sociais não são incomuns. Diante desse cenário há esforços de pesquisadores e provedores de redes sociais para entender mais profundamente a propagação de informações falsas e também para desenvolver tecnologias para a sua detecção mais rápida e eficaz.

A análise de alguns aspectos do contexto das redes sociais ajuda na detecção de postagem de informações falsas como exemplo, pode-se citar, o tipo de usuário que fez a postagem, características da postagem e a rede de contato dos usuários que postaram, repostaram ou reagiram à publicação. Primeiramente, constata-se que informações falsas normalmente são criadas e espalhadas por usuários não humanos, como robôs ou *cyborgs*, de modo que a

verificação desse tipo de usuário pode fornecer informações valiosas para a detecção das *Fake News*. Em segundo lugar, a verificação da postagem e de suas reações também pode auxiliar a detecção, uma vez que as pessoas expressam suas opiniões e emoções por meio das redes sociais, tais como opiniões céticas ou respostas sensacionalistas. Por último, verifica-se pela teoria da homofilia que usuários com interesses e opiniões semelhantes tendem a se seguir nas redes sociais e esse fato, aliado com os algoritmos das redes sociais, gera um efeito de "câmara de eco", que também fornece importantes informações a respeito da postagem (SHU, WANG, *et al.*, 2020).

No entanto, para que a detecção ocorra com o mínimo de interação humana e em tempo hábil, tecnologias de aprendizado de máquina e de Processamento de Linguagem Natural são utilizadas como poderosas ferramentas para o alcance dessa finalidade. No entanto, para realizar o treinamento desses algoritmos é de muita valia a atividade de entidades comprometidas em verificar, de forma manual, a veracidade das informações publicadas nas redes sociais, uma vez que o *dataset* de treinamento deve conter informações já verificadas e classificadas como verdadeiras ou falsas.

2.5 Facebook e propagação de Informações Falsas

O Facebook é uma rede social online, fundada em 2004, que pertence à empresa Meta Platforms. De acordo com o que foi divulgado pelo próprio Facebook nos resultados do terceiro quadrimestre de 2021, a média de usuários mensais ativos chegou a 2,81 bilhões no mês de setembro de 2021 (PARK, 2021).

O Facebook pode ser acessado por quaisquer aparelhos com conexão com a internet, tais como *smartphones*, computadores e *tablets*. Após se registrar e criar um perfil, os usuários podem se conectar com outros usuários registrados e compartilhar postagens por intermédio de textos, fotos, documentos e vídeos. Além disso, o Facebook permite que o usuário configure o controle de acesso de seu perfil e de suas postagens, podendo definir quem poderá visualizar as suas postagens.

Dada a enorme quantidade de usuários do Facebook e a grande flexibilidade no formato de suas postagens, o Facebook se tornou um dos principais meios eletrônicos de divulgação de notícias, tanto verdadeiras como falsas. Com a pandemia da COVID-19, as preocupações com a divulgação de informações falsas nas redes sociais aumentaram e o Facebook, em 2021, foi considerado o principal meio de propagação desse tipo de informação no mundo (NEWMAN, FLETCHER, *et al.*, 2021).

3 TRABALHOS RELACIONADOS

3.1 Considerações Iniciais

Este capítulo compila os principais trabalhos relacionados à detecção de postagens com informações falsas a respeito da pandemia da COVID-19 no Facebook. Conforme já mencionado, este trabalho de conclusão de curso consiste em uma complementação do tema desenvolvido na pesquisa de Mestrado de Mateus Oliveira Cabral, orientado pelo Prof. Dr. Ricardo Rodrigues Ciferri, e, por isso, foi possível aproveitar todos os trabalhos lá mencionados, uma vez que o que diferencia o presente trabalho e a tese do Mateus Oliveira Cabral é que o primeiro foca na rede social Facebook e o segundo na rede social Instagram. Ressalta-se que foram acrescentados alguns novos trabalhos que surgiram entre a data do término da pesquisa de (CABRAL, 2021) do primeiro trabalho até a data atual.

Assim como apresentado na tese do Mateus Oliveira Cabral, esta seção também foi dividida em três tópicos. O primeiro apresenta os principais trabalhos relacionados a redes sociais e informações falsas, o segundo apresenta trabalhos sobre a COVID-19 e informações falsas e a terceira apresenta trabalhos relacionados sobre conjuntos de dados (*datasets*) relacionados a conteúdos falsos de diversos domínios.

3.2 Redes Sociais e Informações Falsas

A Tabela 1 apresenta a lista dos trabalhos relacionados a respeito de Redes Sociais e Informações Falsas. Em seguida é apresentado um breve resumo dos trabalhos que possuem mais relação com o tema detecção de postagens com informações falsas sobre a pandemia da COVID-19 na rede social Facebook.

Tabela 1: Tabela contendo os principais trabalhos a respeito de Redes Sociais e Conteúdos Falsos

Referencia	Título	Foco
(SINGHAL, CHAKRABORTY, *et al.*, 2019)	SpotFake: A multi-modal framework for fake news detection	Notícia
(SHU, BERNARD e LIU, 2019)	Studying Fake News via Network Analysis: Detection and Mitigation	Notícia

Referencia	Título	Foco
(ZHOU e ZAFARANI, 2019)	Network-based Fake News Detection: A Pattern-driven Approach	Notícia
(MONTI, FRASCA, *et al.*, 2019)	Fake News Detection on Social Media using Geometric Deep Learning	Notícia
(RECUERO e GRUZD, 2019)	Cascatas de Fake News Políticas: um estudo de caso no Twitter	Notícia
(JAIN e KASBE, 2018)	Fake News Detection	Notícia
(REIS, CORREIA, *et al.*, 2019)	Explainable machine learning for fake news detection	Notícia
(SAMPLE, JUSTICE e DARRAJ, 2018)	Fake News: A Method to Measure Distance from Fact	Notícia
(NYOW e CHUA, 2019)	Detecting Fake News with Tweets' Properties	*Tweets*, *Sites*
(REIS, CORREIA, *et al.*, 2019)	Supervised Learning for Fake News Detection	Notícia
(FAUSTINI e COVÕES, 2019)	Fake news detection using one-class classification	Notícia, *Tweets*, WhatsApp
(TRAYLOR, STRAUB, *et al.*, 2019)	Classifying Fake News Articles Using Natural Language Processing to Identify In-Article Attribution as a Supervised Learning Estimador	Documentos
(NAJAR, ZAMZAMI e BOUGUILA, 2019)	Fake news detection using bayesian inference	*Sites*
(ATODIRESEI, TăNăSELEA e IFTENE, 2018)	Identifying Fake News and Fake Users on Twitter	*Tweets*
(BAGADE, PALE, *et al.*, 2020)	The Kauwa-Kaate fake news detection system: Demo	*Tweets*

Referencia	Título	Foco
(DONG, VICTOR e QIAN, 2020)	Two-path Deep Semi-supervised Learning for Timely Fake News Detection	Notícia
(TSCHIATSCHEK, SINGLA, *et al.*, 2018)	Fake News Detection in Social Networks via Crowd Signals	Notícia
(BOVET e MAKSE, 2019)	Influence of fake news in Twitter during the 2016 US presidential election	*Tweets*
(PREVITI, RODRIGUEZ-FERNANDEZ, *et al.*, 2020)	Fake news detection using time series and user features classification.	*Tweets*
(AHMED, HINKELMANN e CORRADINI, 2019)	Combining machine learning with knowledge engineering to detect fake news in social networks - A survey	Notícia
(CONROY, RUBIN e CHEN, 2015)	Automatic deception detection: Methods for finding fake news	Notícia
(PÉREZ-ROSAS, KLEINBERG, *et al.*, 2017)	Automatic detection of fake news.	Notícia
(TANDOC, LIM e LING, 2018)	Defining “Fake News”: A typology of scholarly definitions	Notícia
(VOSOUGHI, ROY e ARAL, 2018)	The spread of true and false News online	Twitter
(SILVA, VIEIRA, *et al.*, 2019)	Can Machines Learn to Detect Fake News? A Survey Focused on Social Media	Facebook, Twitter e Google+
(PARIKH e ATREY, 2018)	Media-Rich Fake News Detection: A Survey	Notícia
(PIERRI e CERI, 2019)	False news on social media: A data driven survey	Notícia

Referencia	Título	Foco
(MAHID, MANICKAM e KARUPPAYAH, 2018)	Fake news on social media: Brief review on detection techniques	Notícia
(ELHADAD, LI e GEBALI, 2019)	Fake News Detection on Social Media: A Systematic Survey	Twitter
(MLADENOVA e VALOVA, 2021)	Analysis of the KNN Classifier Distance Metrics for Bulgarian Fake News Detection	Facebook
(SHRIVASTAVA, SINGH, *et al.*, 2022)	A Research on Fake News Detection Using Machine Learning Algorithm	Redes Sociais e Notícias
(OLALEYE, UGEGE, *et al.*, 2021)	Veracity Assessment of Multimedia Facebook Posts for Infodemic Symptom Detection using Bi-modal Unsupervised Machine Learning Approach	Facebook

Por meio de um método sistemático de revisão na literatura, (SILVA, VIEIRA, *et al.*, 2019) procurou nas livrarias eletrônicas clássicas para encontrar os artigos mais recentes a respeito de detecção de *fake news* nas redes sociais com o objetivo de mapear o estado da arte a respeito do assunto. No trabalho os autores definem o que são as *fake News* e encontra as técnicas mais úteis de aprendizado de máquina para a detecção de notícias falsas. No trabalho os autores concluem que o método mais utilizado para essa finalidade não é apenas uma técnica clássica de aprendizado de máquina, mas uma reunião de várias técnicas clássicas coordenadas por uma rede neural artificial.

Utilizando o algoritmo de classificação *K Nearest Neighbor* (KNN) de aprendizado de máquina, o trabalho de (MLADENOVA e VALOVA, 2021) foca na classificação de notícias falsas e *click-baits* nas páginas de Facebook Búlgaras.

No trabalho de (SHRIVASTAVA, SINGH, *et al.*, 2022) os autores implementam modelos utilizando as abordagens de Saco de Palavras e vetorização TF-IDF. Em seguida, utilizaram os seguintes classificadores para categorizar notícias: classificador de regressão logística; classificador de naive Bayes, classificador de florestas aleatórias; e o classificador passivo agressivo. A pesquisa foi conduzida em dois *datasets* distintos nos quais o modelo de saco de palavras, combinado com um classificador de regressão logística, alcançou um F1-Score médio de 92,16% e o modelo de vetorização TF-IDF combinado com um classificador

de regressão logística alcançou um F1-Score médio de 93,47%. Além disso, concluíram que o classificador passivo agressivo funciona bem com grandes volumes de dados e aliado com a vetorização TF-IDF.

O estudo de (OLALEYE, UGEGE, *et al.*, 2021) tenta melhorar resultados obtidos em estudos anteriores para detecção de informações falsas em redes sociais utilizando a metodologia de aprendizado de máquina *bi-modal*. Para isso se utilizou de um *corpus* de multimídia coletado do Facebook, um processador de Linguagem Natural não supervisionado, *Inception v3 model* acoplado com uma rede hierárquica de agrupamento para realizar dupla análise de sentimento em imagens e em textos.

3.3 COVID-19 e detecção de Conteúdos Falsos

A Tabela 2 apresenta a lista dos trabalhos relacionados a respeito de COVID-19 e Informações Falsas. Em seguida é apresentado um breve resumo dos trabalhos que possuem mais relação com o tema detecção de postagens com informações falsas sobre a pandemia da COVID-19 na rede social Facebook.

Tabela 2: Tabela contendo os principais trabalhos a respeito de COVID-19 e Conteúdos Falsos

Referencia	Título	Foco
(BANG, ISHII, *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, CHEN, *et al.*, 2021)	Transformer-based Language Model Finetuning Methods for COVID-19 Fake News Detection	Twitter, Facebook e Instagram
(WANI, JOSHI, *et al.*, 2021)	Evaluating Deep Learning Approaches for Covid19 Fake News Detection	Twitter, Facebook e Instagram
(KOLOSKI, STEPIšNIK-	Identification of COVID-19 related Fake News via Neural Stacking	Twitter, Facebook e Instagram

Referencia	Título	Foco
PERDIH, *et al.*, 2021)		
(DAS, BASAK e DUTTA, 2021)	A Heuristic-driven Ensemble Framework for COVID-19 Fake News Detection	Twitter, Facebook e Instagram
(LI, XIA, *et al.*, 2021)	Exploring Text-transformers in AAAI 2021 Shared Task: COVID-19 Fake News Detection in English	Notícia
(GUPTA, SUKUMARAN, *et al.*, 2021)	Hostility Detection and Covid-19 Fake News Detection in Social Media	Notícia
(GALHARDI, FREIRE, *et al.*, 2020)	Fact or fake? an analysis of disinformation regarding the covid-19 pandemic in Brazil.	WhatsApp e Facebook
(LINDEN, ROOZENBEEK e COMPTON, 2021)	Inoculating against fake news about covid-19.	Notícia
(APUKE e OMAR, 2021)	Fake news and covid-19: modelling the predictors of fake news sharing among social media users.	Notícia
(JÚNIOR, RAASCH, *et al.*, 2020)	Da desinformação ao caos: uma análise das fake news frente à pandemia do coronavírus (covid-19) no brasil.	Notícia
(NETO, GOMES, *et al.*, 2020)	Fake news no cenário da pandemia de covid-19.	Notícia
(FERREIRA, LIMA e SOUZA, 2020)	Desinformação, infodemia e caos social: impactos negativos das fake news no cenário da covid-19.	Teórico
(SINGH, BANSAL, *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
	Overview	
(INUWA-DUTSE, 2021)	Towards combating pandemic-related misinformation in social media.	Twitter

Referencia	Título	Foco
(MALLA e ALPHONSE, 2022)	Fake or real news about COVID-19? Pretrained transformer model to detect potential misleading news	Twitter, Facebook e Instagram
(YAFOOZ, EMARA e LAHBY, 2022)	Detecting Fake News on COVID-19 Vaccine from YouTube Videos Using Advanced Machine Learning Approaches	Youtube
(KOLLURI e MURPHY, 2021)	CoVerifi: A COVID-19 news verification system	Notícias

Observando as técnicas automatizadas de detecção de notícias falsas por uma perspectiva de mineração de dados, (WANI, JOSHI, *et al.*, 2021) analisaram diferentes algoritmos de classificação de texto utilizando o *dataset* "Contraint@AAAI 2021 COVID-19 Fake News detection*"*. Os algoritmos de classificação utilizados foram baseados em redes neurais convolucionais (*Convolutional Neural* Networks - CNN), memória de longo e curto prazo (*Long Short Term Memory* – LSTM) e *Bidirectional Encoder Representations from Transformers* (BERT). A melhor acurácia obtida no trabalho para detecção de notícias falsas sobre a COVID-19 foi de 98,41%.

Visando alcançar um modelo robusto de detecção de notícias falsas sobre a COVID-19, (BANG, ISHII, *et al.*, 2021) utilizaram duas abordagens separadas, sendo a primeira formada por modelos baseados em *Transformers* (que são modelos de aprendizado profundo) e a segunda composta pela remoção de instâncias prejudiciais de treinamento por meio de cálculos de influência. Com a primeira abordagem os autores alcançaram 98,13% no W-F1 e 54,33% na segunda abordagem.

No trabalho de (MALLA e ALPHONSE, 2022), visando conscientizar a sociedade a respeito da importância de informações legítimas e precisas e de prevenir a propagação de informação falsa, investigou dados de diversas plataformas de redes sociais como Twitter, Facebook e Instagram relacionados com a COVID-19. O objetivo do artigo é categorizar *Tweets* como verdadeiros ou falsos e, para isso, testaram vários modelos de aprendizado profundo no *dataset* de notícias falsas da COVID-19 e perceberam que os modelos CT-BERT e RoBERTa superaram o desempenho de outros modelos de aprendizado profundo como BERTweet, AlBERT, e DistlBERT e alcançaram acurácia de 98,88% e F1-Score de 98,31%.

Os autores (YAFOOZ, EMARA e LAHBY, 2022) propuseram um modelo para detectar as notícias falsas a respeito da vacina da COVID-19 em vídeos do YouTube usando uma

abordagem de análise de sentimentos através de métodos de aprendizado de máquina e aprendizado profundo focando na linguagem arábica do Oriente Médio. Dois experimentos foram conduzidos utilizando classificadores de aprendizado de máquina e modelos de aprendizado profundo. O desempenho obtido foi 94% de acurácia. Na abordagem que utilizou aprendizado profundo essa medida alcançou 99%.

No trabalho de (KOLLURI e MURPHY, 2021) é apresentado uma aplicação da *Web*, batizada de CoVerifi, que combina o poder do aprendizado de máquina com o *feedback* humano para verificar a credibilidade de notícias a respeito da COVID-19. A aplicação permite que usuários opinem a respeito do conteúdo das notícias e, de acordo com os autores, o CoVerif vai permitir a disponibilização de dados públicos a respeito da pandemia.

3.4 Conjunto de Dados para Conteúdos Falsos

Nesta seção são apresentados trabalhos que reuniram dados a respeito de conteúdos falsos. Esses trabalhos estão listados na Tabela 3.

Tabela 3: Trabalhos contendo conjuntos de dados a respeito de conteúdos falsos

Referencia	Título	Foco	Idioma	Tamanho
(PATWA, SHARMA, *et al.*, 2021)	Fighting an Infodemic: COVID-19 Fake News Dataset	Twitter	Inglês	5.600 verdadeiros, 5.100 falsos
(TACCHINI, BALLARIN, *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

Referencia	Título	Foco	Idioma	Tamanho
(SHU, MAHUDESWARAN, *et al.*, 2019)	FakeNewsNet: A Data Repository with News Content, Social Context and Spatialtemporal Information for Studying Fake News on Social Media	Notícias	Inglês	2.898 (PoliticFact), 62.198 (Gossip-Cop)
(MONTEIRO, SANTOS, *et al.*, 2018)	Contributions to the study of fake news in portuguese: Newcorpus and automatic detection results	Notícias	Português	3.600 verdadeiras, 3600 falsas
(ZAREI, FARAHBAKHSH, *et al.*, 2020)	A First Instagram Dataset on COVID-19	Instagram	Inglês e outros idiomas	5.300 postagens
(KIM, AUM, *et al.*, 2021)	FibVID: Comprehensive fake news diffusion dataset during the COVID-19 period	Twitter	Diversos	1.353 alegações de notícias e 221.253 *tweets* relevantes

Os autores (PATWA, SHARMA, *et al.*, 2021) reuniram um *dataset* com 10.700 postagens em redes sociais e artigos de notícias verdadeiras e falsas a respeito da pandemia da COVID-19. Em seu trabalho realizaram uma classificação binária (verdadeiro ou falso) e compararam o *dataset* resultante com quatro abordagem de aprendizado de máquina: arvores de decisão, regressão logística, *Gradient Boost* e SVM obtendo performance de 93,32% no F1-Score com SVM nos dados de teste.

Para contribuir com a detecção de notícias falsas em redes sociais, (TACCHINI, BALLARIN, *et al.*, 2017) mostraram que postagens no Facebook podem ser classificadas com elevada acurácia entre boatos e não boatos com base nos usuários que as “curtiram”. Os autores apresentaram duas técnicas de classificação, uma baseada em regressão logística e a outra em uma adaptação de algoritmos booleanos de *crowdsourcing*. O *dataset* reúne 15.500 postagens

feitas no Facebook e 909.236 usuários e obtiveram acurácia de classificacão de 99% mesmo quando os dados de treinamento continham menos de 1% da quantidade de postagens.

O conjunto de dados de (SANTIA e WILLIAMS, 2018) contém 2.282 artigos de notícias postadas no Facebook durante o mês de setembro de 2016. Esses artigos foram categorizados de acordo com a veracidade da notícia postada pelo BuzzFeed. Posteriormente a categorização foi refinada em quatro grupos: *predominantemente verdadeira*; *predominantemente falsa*; *mistura de verdadeira ou falsa*; e *sem contexto factual*. Além disso, o trabalho integra dados do Facebook referentes a comentários e reações publicitárias disponíveis na plataforma do *Graph API* e fornece ferramentas para acessar artigos de notícias na *web*. O *dataset* resultante disponibilizou mais de 1,6 milhão de itens de texto e foi o mais extensivo criado até a época do trabalho e permite a utilizações de aplicações de aprendizado de máquina para detecção de notícias falsas.

No trabalho de (ZAREI, FARAHBAKHSH, *et al.*, 2020), os autores construíram um *dataset* contendo 5.300 postagens feitas no Instagram nas quais foi coletado dados textuais que foram categorizados em duas classes: notícias verdadeiras e notícias falsas. Os dados foram coletados a partir da data de 30 de março de 2020 até a data da publicação do trabalho.

O *dataset* denominado FibVID (*Fake news* information-broadcasting dataset of COVID-19), produzido por (KIM, AUM, *et al.*, 2021), endereça tanto notícias a respeito da COVID-19 quanto as que não tem relação com assunto por três ângulos: indicador da veracidade (verdadeira ou falsa) de cada item de notícia; reunião de alegações espúrias de *tweets*; e por fim, disponibiliza informações básicas para o usuário, incluindo termos e características de notícias consideradas “descaradamente” falsas, para que este entenda a categorização da notícia. O *dataset* é composto por 1.353 alegações de notícias e 221.253 *tweets* relevantes escritos por 144.741 usuários.

3.5 Considerações Finais

Esse capítulo identificou os trabalhos mais relevantes a respeito do tema deste trabalho de conclusão de curso, a detecção de postagens com informações falsas no Facebook. Esses trabalhos são importantes para realizar comparações com os resultados que foram obtidos neste trabalho. Finalmente, foi possível entender limitações de cada abordagem das linhas de pesquisas listadas.

4 DETECÇÃO DE POSTAGENS FALSAS SOBRE COVID-19 NO FACEBOOK

4.1 Considerações Iniciais

O objetivo deste trabalho de conclusão de curso foi investigar a detecção de postagens com informações falsas no idioma Português sobre a pandemia do Covid-19 especificamente aplicada na rede social Facebook.

Nesse sentido, este Capítulo primeiramente detalha a proposta do projeto deste trabalho de conclusão de curso. Na seção 4.2 é descrita a metodologia que foi utilizada no desenvolvimento das atividades deste projeto de pesquisa, em conformidade com os objetivos propostos na seção 1.4 , descrevendo a arquitetura de *framework* que foi implementada e quais recursos foram utilizados para realizar o Processamento de Linguagem Natural dos textos das postagens. As postagens usadas nos experimentos abrangem o período de janeiro de 2020 a junho de 2022 e na seção 4.3 é apresentada as características das postagens do Facebook e que tipos de dados foram coletados das mesmas. Como a análise de texto de imagens não faz parte do presente trabalho de conclusão de curso, as imagens das postagens foram desconsideradas.

Foram extraídos do Facebook um total de 401 legendas de postagens em português com as hashtag “covid-19” e “coronavírus” entre janeiro de 2020 a junho de 2022, e separadamente foram extraídos e filtrados o texto de cada postagem das páginas jornalísticas e de verificadores de conteúdo que continham informações a respeito do coronavírus nas seguintes quantidades: Agência Lupa 143 postagens, Aos Fatos 58 postagens, CNN Brasil 30 postagens, Fato ou Fake 64 postagens, Folha de São Paulo 16 postagens, Projeto Comprova 76 postagens, demais usuários 14 postagens.

Para a etapa de aprendizado de máquina, 401 legendas de postagens de usuários foram classificadas manualmente como verdadeiras ou falsas por meio de uma checagem manual utilizando-se como referência o texto de postagens feitas no Facebook nas páginas jornalísticas ou nas verificadoras de conteúdo. Em 200 dessas postagens continha informações falsas e foram etiquetadas como “Falsa” e as outras 201 postagens continham informações consideradas verdadeiras pelas páginas jornalísticas e verificadoras de conteúdo e foram etiquetadas como “Verdadeira”.

Por fim, na seção 4.4, são apresentados os resultados obtidos após aplicação dos modelos de aprendizado de máquina.

4.2 Metodologia

Nesta seção é descrita a proposta da arquitetura para implementar um algoritmo para detectar postagens falsas sobre Covid-19 no Facebook e a forma que será trabalhado as principais técnicas de PLN e de Aprendizado de Máquinas que serão utilizadas na implementação do projeto.

4.2.1 Identificação do Problema

A *Internet* possibilitou que o processo de comunicação fosse intensificado de maneira relevante. Com o surgimento das redes sociais, esse processo se tornou ainda mais acentuado, o que fez com que não apenas as pessoas passassem a se comunicar predominantemente por meios eletrônicos como também fez com que elas passassem a consumir informações por esses mesmos meios. Dessa forma, jornais e revistas já migraram para a Internet e para as redes sociais para acompanhar essa evolução tecnológica.

No entanto, não são as entidades jornalísticas as únicas que têm permissão para compartilhar informações em meios eletrônicos. Na *Internet* e nas redes sociais, qualquer usuário pode ser tanto o locutor como o interlocutor e, como, até pouco tempo atrás não havia, praticamente nenhum controle direcionado ao conteúdo das postagens, observou-se nos últimos anos um aumento disseminado na propagação de informações distorcidas ou completamente inverídicas o que resultou em diversas consequências negativas no comportamento de muitos indivíduos, inclusive durante a época da pandemia da COVID-19.

Em resposta a esse fenômeno, atualmente a maioria das redes sociais já possuem mecanismos para tentar evitar a propagação de informações falsas, no entanto, seja pela grande quantidade de dados que são gerados em uma rede social ou seja pela dificuldade de filtragem desses dados, o desempenho desses mecanismos ainda não alcançou um patamar para evitar de forma segura as consequências sociais da propagação de informações falsas.

Por fim, ao realizar a revisão bibliográfica, do presente trabalho de conclusão de curso, que foi sintetizada no Capítulo 3, constatou-se que há poucos trabalhos relacionados à detecção de informações falsas em português na rede social Facebook, uma plataforma muito utilizada para o compartilhamento de notícias, seja por páginas jornalísticas ou por usuários.

4.2.2 Descrição da base de dados

Dada a especificidade do tema abordado neste trabalho de conclusão de curso, não foi encontrada nenhuma base de dados em português para o assunto tratado. Dessa forma, houve a

necessidade da coleta dos dados para posterior processamento e utilização neste trabalho. Para isso, os dados foram coletados manualmente na plataforma do Facebook. Para isso, utilizou-se o mecanismo de busca dessa rede social utilizando-se as palavras chave “covid-19” e “coronavírus” para filtrar as postagens relacionadas ao tema da pandemia da COVID-19. Ao total foram coletadas 401 tuplas de postagens em português sobre esse assunto nessa rede social, todas elas compreendidas no período de janeiro de 2020 até junho de 2022. Como o trabalho focou apenas nos aspectos textuais das postagens, foram coletadas apenas as legendas das mesmas, desconsiderando os possíveis arquivos de mídia e os metadados. Esses dados foram armazenados em um arquivo no formato CSV (Comma Separated Value).

Em seguida realizou-se consulta em páginas de agências de verificação para ser possível a classificação manual das postagens em duas categorias: “Verdadeira” ou “Falsa”. Partiu-se do pressuposto que tais agências de verificação são realmente comprometidas com sua atividade e suas verificações de informações são confiáveis. Utilizou-se as seguintes páginas para essa atividade Aos Fatos, Agência Lupa, Comprova, CNN Brasil, Fato ou Fake.

Após a classificação manual foi adicionada a coluna nomeada “Classificação” a cada tupla dos dados das postagens coletados, referente à classificação binária mencionada no parágrafo anterior, o que finalizou a construção da base de dados. Dessa forma a base de dados é formada por 401 tuplas, cada uma contendo apenas 2 colunas: a primeira contendo o texto da legenda coletada; e a segunda a classificação manual realizada entre “Verdadeira” e “Falsa”. A quantidade de tuplas classificadas como “Verdadeira” nessa base de dados é 200 e como “Falsa”é 201.

4.2.3 Pré-processamento

Como a base de dados foi construída manualmente, coletando apenas as informações estritamente necessárias para a finalidade do trabalho, não foi necessária nenhuma seleção de atributo, exclusão de registros ou limpeza de dados. Além disso, não foi necessário nenhum procedimento para balancear a quantidade de dados das duas classes uma vez que haviam 200 tuplas classificadas como “Verdadeira” e 201 classificadas como “Falsa”.

No entanto, para verificar sob quais condições o algoritmo de aprendizado de máquina chega aos melhores resultados, o texto da legenda das postagens foi submetido a alguns pré-processamentos. Foi aplicada a técnica de stemização para extrair apenas o radical de cada palavra das postagens, formando uma nova coluna na base de dados e também se criou uma coluna contendo o texto com todos os caracteres em minúsculos. Na etapa do aprendizado de

máquina o algoritmo foi executado nesses dois formatos e também para o texto sem nenhum pré-processamento.

Além disso, essas três possibilidades foram executadas para três opções de transformadores: o TF-IDF, o *Count*, e *Hashing*. E, para cada opção desses transformadores utilizou-se as técnicas de *Stopwords* e *N-Grams*. Por fim, utilizou-se 3 diferentes opções de vetorização de texto para multiplicar ainda mais as possibilidades de modelos possíveis a serem aplicados na etapa de aprendizado de máquina: a primeira gerando um vetor de características composto apenas de *N-grams* de palavras; a segunda gerando um vetor de características composto apenas de *N-grams* de caracteres; e a terceira gerando um vetor de características composto apenas de *N-grams* de caracteres que apenas estão contidos no texto dentro das fronteiras das palavras, que são definidas por espaço.

Ao final, obteve-se 108 combinações de dados pré-processados a serem aplicados aos algoritmos de aprendizado de máquina.

4.2.4 Processamento de Linguagem Natural

Este módulo tem como objetivo extrair as informações textuais de postagens classificadas como verdadeiras e falsas, com análises morfológicas, sintáticas e semânticas por meio de técnicas como a classificação de reconhecimento de entidades, etiquetagem, tokenização e de *bag of words*. Para isso, foi utilizado o pacote NTLK (*Natural Language Toolkit*), uma biblioteca em Python para análise de textos. As técnicas deste pacote foram implementadas nos textos das legendas de cada postagem armazenada do Facebook e o resultado de algumas funções estatísticas do pacote são apresentados no subtópico 4.3.1.

4.2.5 Aprendizado de Máquina

Já para o aprendizado de máquina utilizou-se 17 classificadores diferentes disponíveis em duas bibliotecas Python: *SkLearn* e *XGBoost*. Da primeira utilizou-se os classificadores SVM (NuSVC e SVC), Naive Bayes (Multinomial, Bernoulli e *Complement*), *Ensemble* (*Random Forest, Bagging, Gradient Boosting, Ada Boost, Extra Tree*), Árvores de decisão (*Extremely Randomized Tree*), Redes Neurais (MLP), K-vizinhos mais próximos (KNN) e XGBoost. Todos esses algoritmos foram aplicados às 108 combinações de dados pré-processados a partir das legendas das postagens coletadas, o que totaliza 1836 modelos de aprendizado de máquina diferentes testados e avaliados.

Para medir o desempenho de cada modelo de classificação utilizou-se as medidas de Acurácia, Precisão, Revocação, F1-Score e área abaixo da curva ROC. A curva ROC também foi traçada e apresentada no subtópico 4.4.

4.2.6 Arquitetura do framework

A Figura 5 evidencia a arquitetura do *framework* que foi implementado.

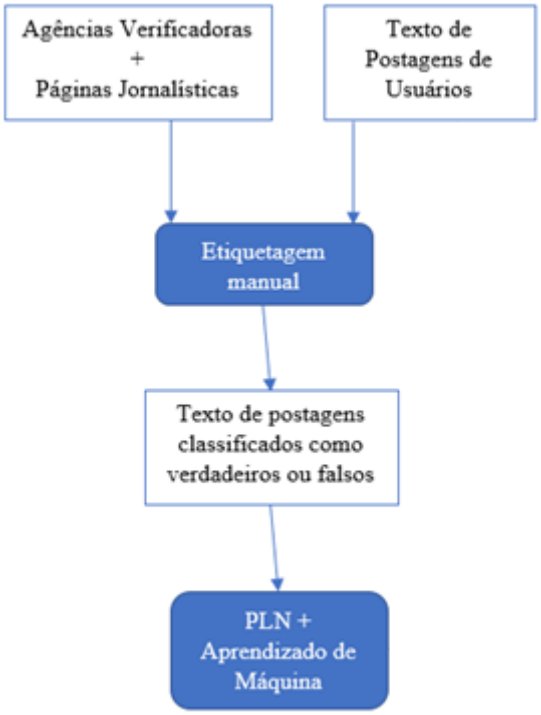

Figura 5 - Arquitetura do framework implementado.

4.3 Características das Postagens

As postagens de usuários na plataforma do Facebook podem ser publicadas de diversas formas. Normalmente são feitas utilizando-se uma ou várias imagens combinadas com uma legenda escrita por um usuário. É possível também a disponibilização de arquivos de vídeo nas postagens ou links para outras páginas da Internet. Além disso, a plataforma permite que outros usuários reajam às postagens curtindo ou publicando comentários.

O modelo de detecção de postagens falsas do presente trabalho se baseou apenas no texto das legendas das postagens. Dessa forma, não foi levado em consideração a análise das imagens, vídeos e dos metadados contidos nas mesmas.

No subtópico seguinte é descrito como os textos das legendas das postagens foram utilizados nos modelos e o resultado obtido por cada um.

4.3.1 Texto das Postagens

Os textos das legendas que foram coletadas das postagens do Facebook foram armazenados em uma coluna de um *DataFrame*. Em seguida, com ajuda das informações fornecidas pelas páginas de verificação de notícias, os textos das postagens foram classificados em “Verdadeiros” ou “Falsos”. Essa classificação foi armazenada em outra coluna do *DataFrame*. Essas são as duas únicas informações que foram utilizadas para levantar as estatísticas de linguagem natural e para o modelo de aprendizado de máquina de classificação de notícias.

Utilizando o pacote *NLTK* foi possível a exploração de algumas características da base de dados textuais coletada. Na Figura 6 é evidenciado as 30 palavras mais presentes nos textos das legendas das 401 postagens coletadas. Já a Figura 7 mostra o resultado da geração de uma nuvem de palavras dos textos das legendas de todas as postagens coletadas. O mesmo procedimento foi feito para o texto das postagens classificados como “Verdadeiras”, evidenciado na Figura 8 e na Figura 9, e para o texto das postagens classificados como “Falsas”, evidenciado na Figura 10 e na Figura 11.

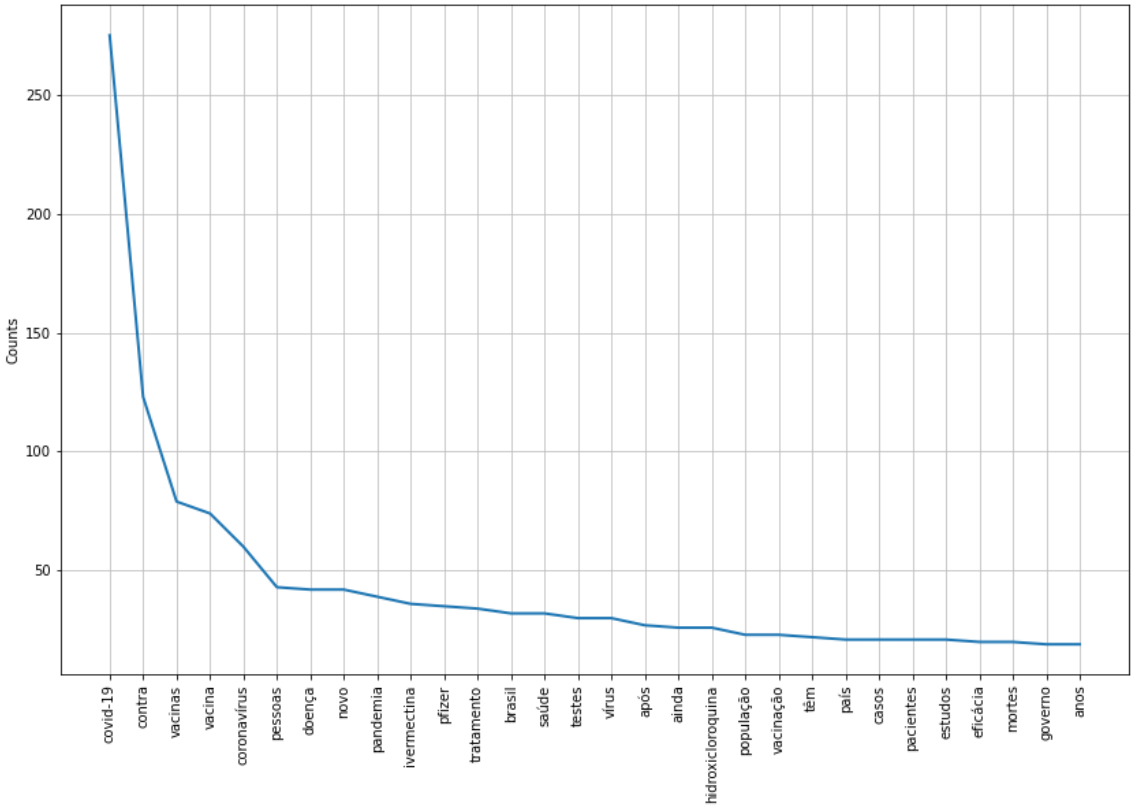

Figura 6 - Gráfico de frequência das 30 palavras mais presentes no texto das legendas de todas as postagens coletadas.

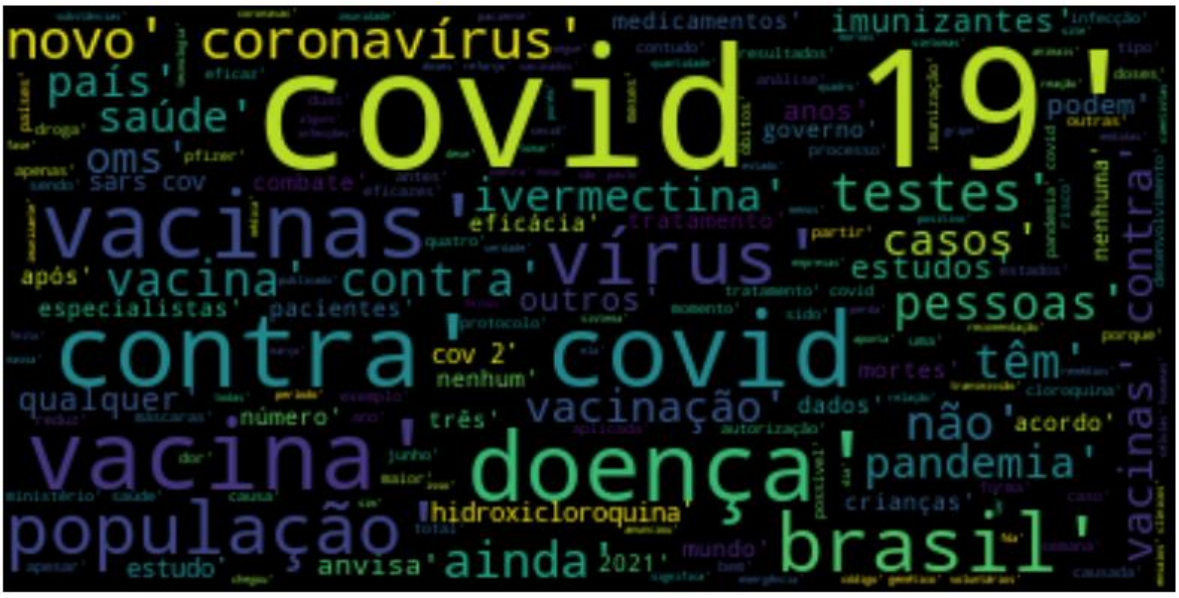

Figura 7 - Nuvem de palavras dos textos de todas as postagens coletadas.

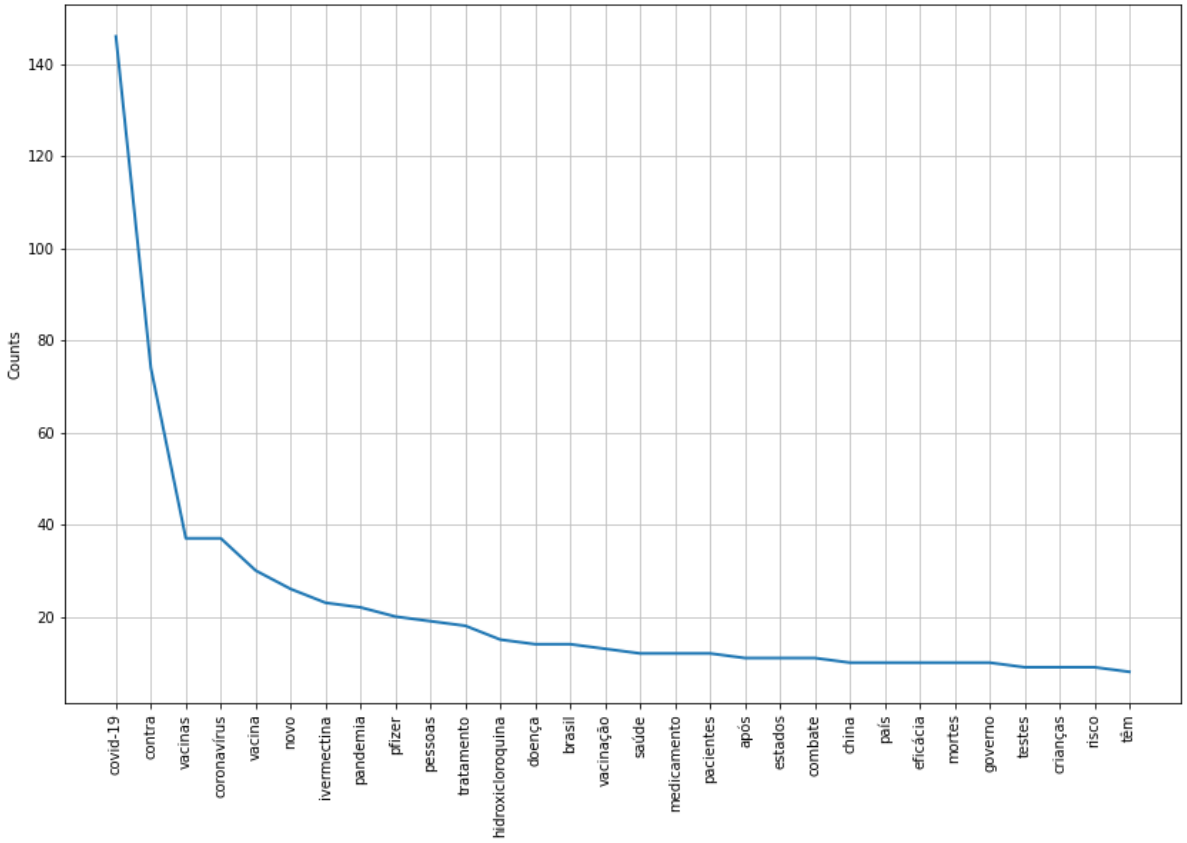

Figura 8 - Gráfico de frequência das 30 palavras mais presentes no texto das legendas das postagens classificadas como "Verdadeira".

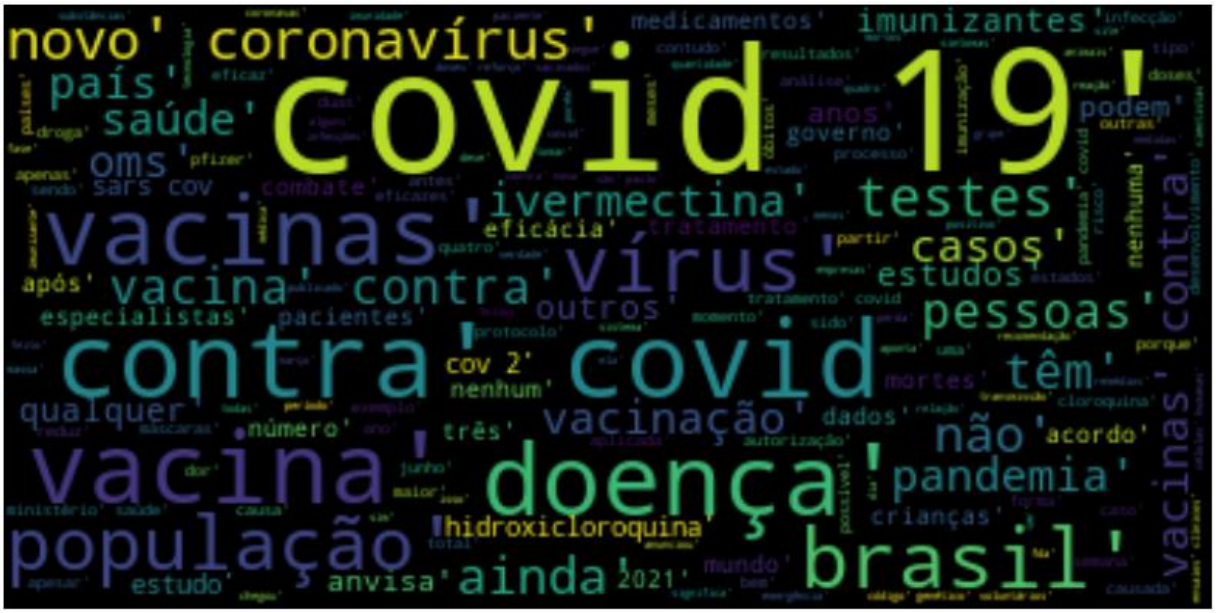

Figura 9 - Nuvem de palavras dos textos das postagens classificadas como "Verdadeira".

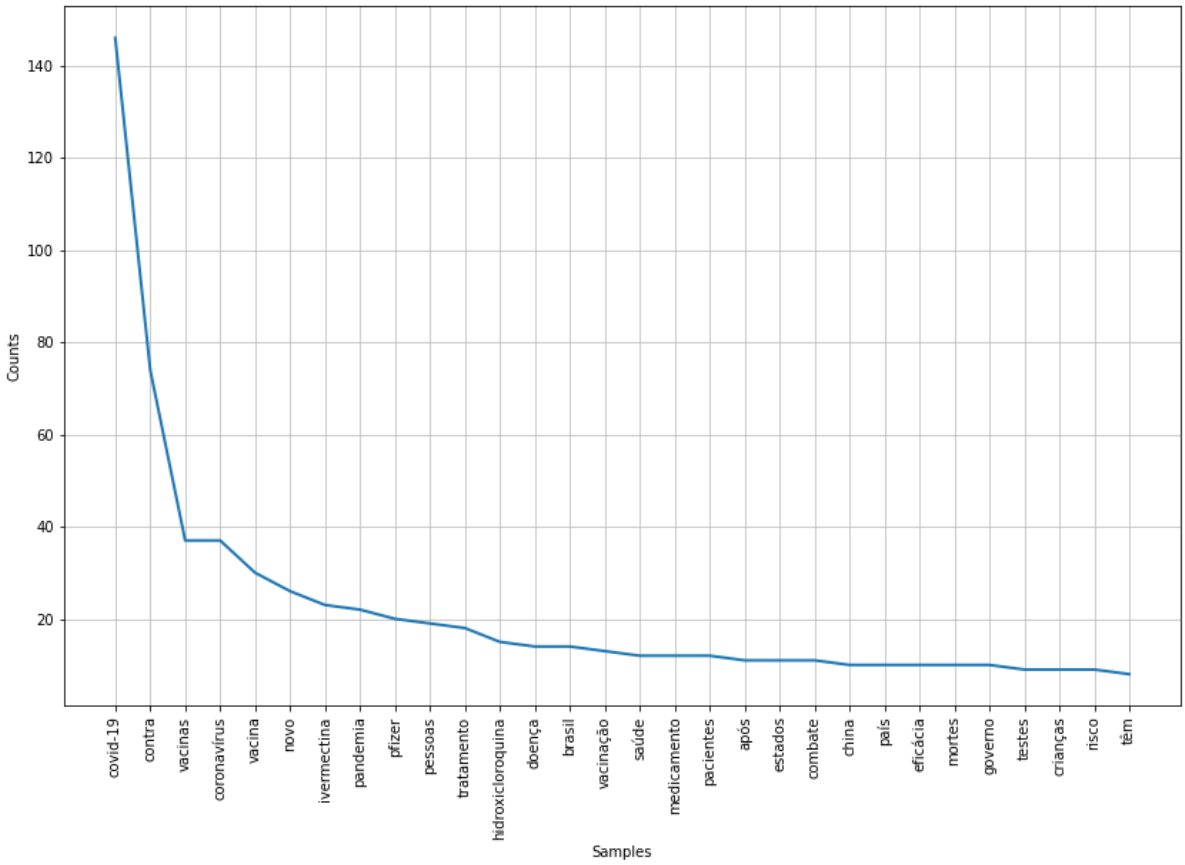

Figura 10 - Gráfico de frequência das 30 palavras mais presentes no texto das legendas das postagens classificadas como “Falsa”.

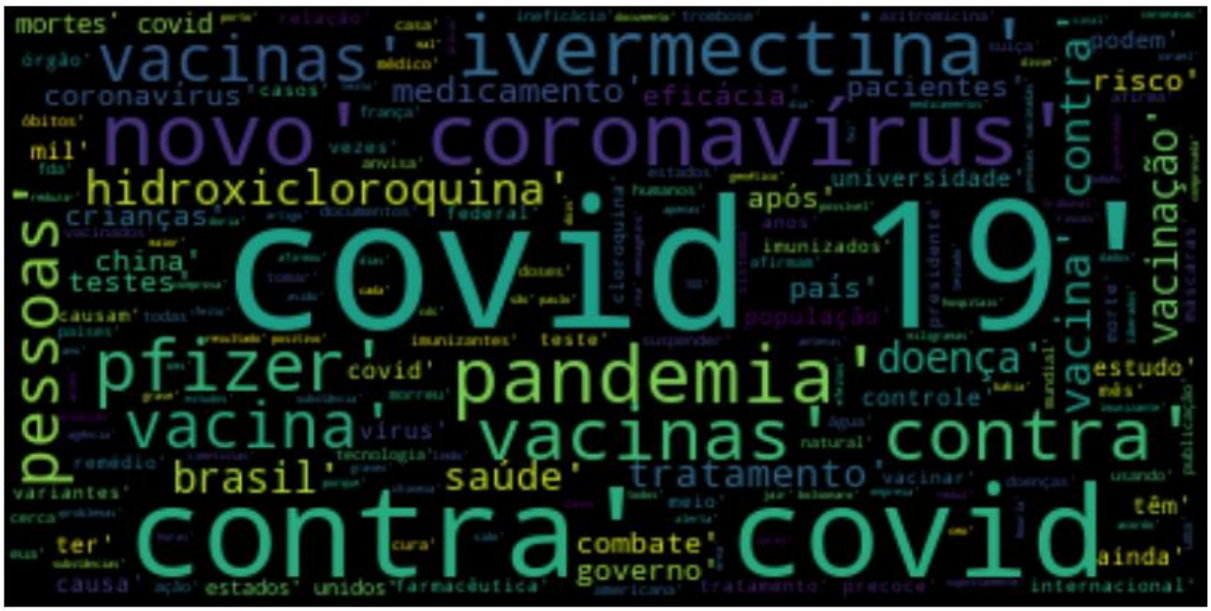

Figura 11 - Nuvem de palavras dos textos das postagens classificadas como "Falsa".

4.4 Resultados Obtidos

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

Os modelos foram executados em uma máquina ACER Aspire Nitro AN515-54-528V com processador Intel Core i5-9300H de 9ª geração, 2,4 GHz de frequência básica, 8 núcleos e placa de vídeo NVIDIA GeForce GTX 1650. O Sistema Operacional utilizado foi Windows 10 Pro, 64 bits, com um HD SSD de 1 TB e memória RAM DDR4 de 32 GB e frequência de 2667 MHz.

Foram utilizados 16 classificadores da biblioteca *Sklearn* e o classificador *XGBoost*, totalizando 17 classificadores. Todos são utilizados para aprendizado supervisionado e a fonte de dados, tanto para o treino quanto para o teste, foi rotulada como "Verdadeira" ou "Falsa", conforme mencionado anteriormente. Ao final avaliou-se a média de desempenho desses classificadores.

Uma das dificuldades encontradas ao longo deste trabalho foi a ausência de uma base que contivesse os dados já reunidos e rotulados para a direta aplicação dos modelos de aprendizado de máquina. Dessa forma, foi necessário a classificação manual dos dados textuais coletados das legendas das postagens para utilizar como base para o aprendizado de máquina.

Para uma análise geral de desempenho dos classificadores foi feita uma média das principais métricas de todos os modelos obtidos com o pré-processamento dos dados para cada um dos 17 classificadores. São essas métricas a precisão, a revocação, o F1-Score, a acurácia e a área abaixo da curva ROC (ROC-AUC). A **Erro! Fonte de referência não encontrada.** evidencia o resultado obtido para cada classificador ordenados de acordo com a acurácia obtida e a Figura 12 mostra as mesmas medidas, porém em forma de gráfico e separando o resultado obtido para os casos de classificação "Verdadeira" do casos de classificação "Falsa".

Tabela 4 - Resultado das médias de cada métrica para cada classificador.

classificador	Precisão 1	Revocação 1	F1-Score 1	Precisão 2	Revocação 2	F1-Score 2	Acurácia	ROC-AUC
RandomForestClassifier	0.687065	0.714537	0.697556	0.709574	0.678778	0.690528	0.697676	0.761085
SVC	0.688620	0.712463	0.694231	0.712991	0.679444	0.689278	0.696870	0.761605
ExtraTreesClassifier	0.695463	0.687120	0.687500	0.699593	0.703361	0.697602	0.696724	0.762577
BaggingClassifier	0.679204	0.721620	0.693926	0.714750	0.660065	0.679694	0.692602	0.742127
LogisticRegression	0.688519	0.691639	0.685676	0.695148	0.688380	0.687065	0.690294	0.748770
NuSVC	0.679120	0.700602	0.686185	0.699093	0.671630	0.680944	0.688402	0.750625
MLPClassifier	0.682630	0.692796	0.684046	0.693148	0.678620	0.681889	0.687354	0.740170
BernoulliNB	0.654324	0.773306	0.705463	0.737139	0.595130	0.652898	0.685794	0.737837
GradientBoostingClassifier	0.668537	0.699806	0.679815	0.691463	0.654269	0.668306	0.678629	0.737001
XGBClassifier	0.661648	0.697417	0.675509	0.685741	0.645287	0.661324	0.672758	0.732155
ComplementNB	0.679731	0.681278	0.658796	0.685287	0.667620	0.661037	0.669646	0.747205
MultinomialNB	0.676426	0.669417	0.644907	0.674630	0.669324	0.654407	0.662087	0.747202
SGDClassifier	0.668750	0.662171	0.647713	0.669699	0.654380	0.642852	0.659970	0.690884
AdaBoostClassifier	0.638315	0.657361	0.643296	0.651389	0.627620	0.634324	0.644115	0.678548
KNeighborsClassifier	0.626463	0.618704	0.605389	0.632222	0.612870	0.602833	0.617462	0.656325
DecisionTreeClassifier	0.610407	0.620269	0.610472	0.617278	0.604500	0.605963	0.613159	0.612790
ExtraTreeClassifier	0.599907	0.609037	0.599704	0.605519	0.593556	0.593981	0.601782	0.601600

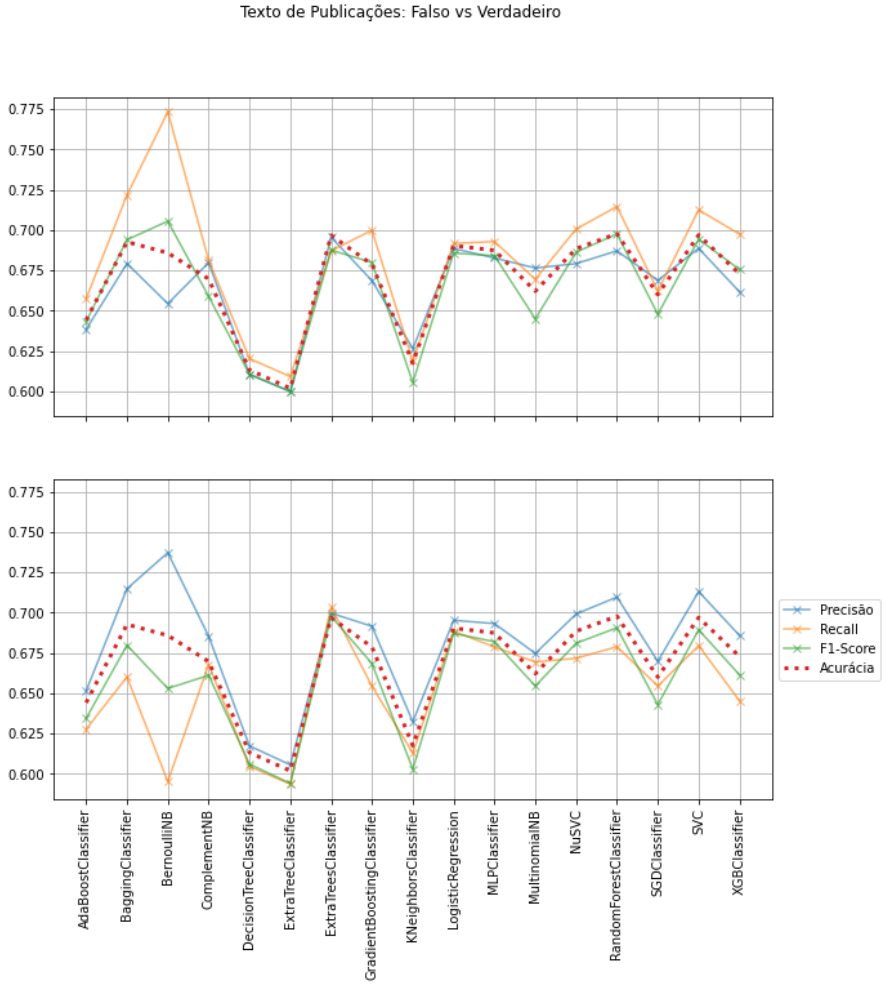

Figura 12 - Média das métricas dos classificadores.

Nota-se que as médias de acurácia e F1-Score, que são usados para medir a exatidão da etapa de classificação, variaram de 59% a 70%, com destaque para os classificadores RandomForestClassifier, SVC e ExtraTreeClassifier que obteram resultados próximos a 70% em todos as suas métricas de desempenho.

Além de avaliar a média dos resultados obtidos por cada algoritmo de classificação, foi avaliado também o desempenho do melhor experimento de cada um dos classificadores. Os resultados obtidos foram evidenciados na **Erro! Fonte de referência não encontrada.** juntamente com suas condições de pré-processamento.

Tabela 5 - Resultado das métricas do melhor experimento de cada classificador.

classificador	efeito	extrator	vetorizador	nivel	pré-processamento	tipo	Precisão 1	Revocação 1	F1-Score 1	Precisão 2	Revocação 2	F1-Score 2	Acurácia
NuSVC	steaming	tfidf_lword	TF-IDF	Word	Sem pré-processamento	texto steam	0.765	0.7840	0.7720	0.776	0.7560	0.763	0.770732
BaggingClassifier	steaming	tfidf_tcharwb	TF-IDF	Char WB	Stopword + N-Gram	texto	0.737	0.8040	0.7670	0.794	0.7210	0.755	0.765610
LogisticRegression	steaming	count_lword	Count	Word	Sem pré-processamento	texto	0.744	0.8070	0.7730	0.788	0.7160	0.747	0.763110
ExtraTreesClassifier	steaming	tfidf_lword	TF-IDF	Word	Sem pré-processamento	texto	0.754	0.7680	0.7590	0.773	0.7490	0.756	0.760610
SVC	steaming	count_ngword	Count	Word	N-Gram	texto	0.716	0.8520	0.7780	0.818	0.6610	0.728	0.758232
RandomForestClassifier	steaming	tfidf_lword	TF-IDF	Word	Sem pré-processamento	texto	0.731	0.8010	0.7620	0.784	0.7080	0.741	0.753415
GradientBoostingClassifier	steaming	count_lword	Count	Word	Sem pré-processamento	texto	0.722	0.8120	0.7610	0.795	0.6850	0.733	0.750610
BernoulliNB	minusculo	tfidf_ngword	TF-IDF	Word	N-Gram	texto	0.685	0.8760	0.7700	0.840	0.6070	0.704	0.743293
MultinomialNB	steaming	count_ngchar	Count	Char	N-Gram	texto steam	0.731	0.7560	0.7410	0.754	0.7290	0.739	0.743049
MLPClassifier	steaming	count_lword	Count	Word	Sem pré-processamento	texto	0.720	0.7910	0.7510	0.770	0.6900	0.722	0.740854
ComplementNB	steaming	count_lword	Count	Word	Sem pré-processamento	texto	0.717	0.8000	0.7520	0.776	0.6820	0.722	0.740610
SGDClassifier	minusculo	tfidf_tchar	TF-IDF	Char	Stopword + N-Gram	texto	0.736	0.7355	0.7325	0.748	0.7415	0.741	0.740610
XGBClassifier	normal	count_lword	Count	Word	Sem pré-processamento	texto	0.711	0.7830	0.7450	0.767	0.6870	0.724	0.738415
KNeighborsClassifier	normal	tfidf_tchar	TF-IDF	Char	Stopword + N-Gram	texto	0.746	0.5670	0.6440	0.656	0.8150	0.727	0.693476
AdaBoostClassifier	minusculo	tfidf_tcharwb	TF-IDF	Char WB	Stopword + N-Gram	texto	0.692	0.6970	0.6900	0.700	0.6940	0.693	0.693171
ExtraTreeClassifier	minusculo	tfidf_swword	TF-IDF	Word	Stopword	texto	0.673	0.7120	0.6870	0.701	0.6550	0.671	0.683171
DecisionTreeClassifier	normal	hash_ngcharwb	Hashing	Char WB	N-Gram	texto	0.673	0.6870	0.6730	0.682	0.6630	0.667	0.676037

Constata-se que se for selecionada o melhor experimento de cada classificador o resultado para Acurácia, F1-Score e Precisão variaram no intervalo de 67% a 78%. O melhor experimento realizado foi resultante do classificador NuSVC utilizando-se o texto stematizado, com o vetorizador TF-IDF utilizando a frequência de palavras e o texto sem pré-processamento. Neste experimento foi obtida Acurácia de 77%.

Por fim, na Figura 13 é evidenciada a curva ROC da seleção do melhor experimento de cada classificador. Conforme explicado na seção 2.3.3.6, quanto maior a área abaixo da curva traçada, maior a Taxa de Verdadeiros Positivos e uma área igual a 1, representaria um modelo que acertou todas as classificações.

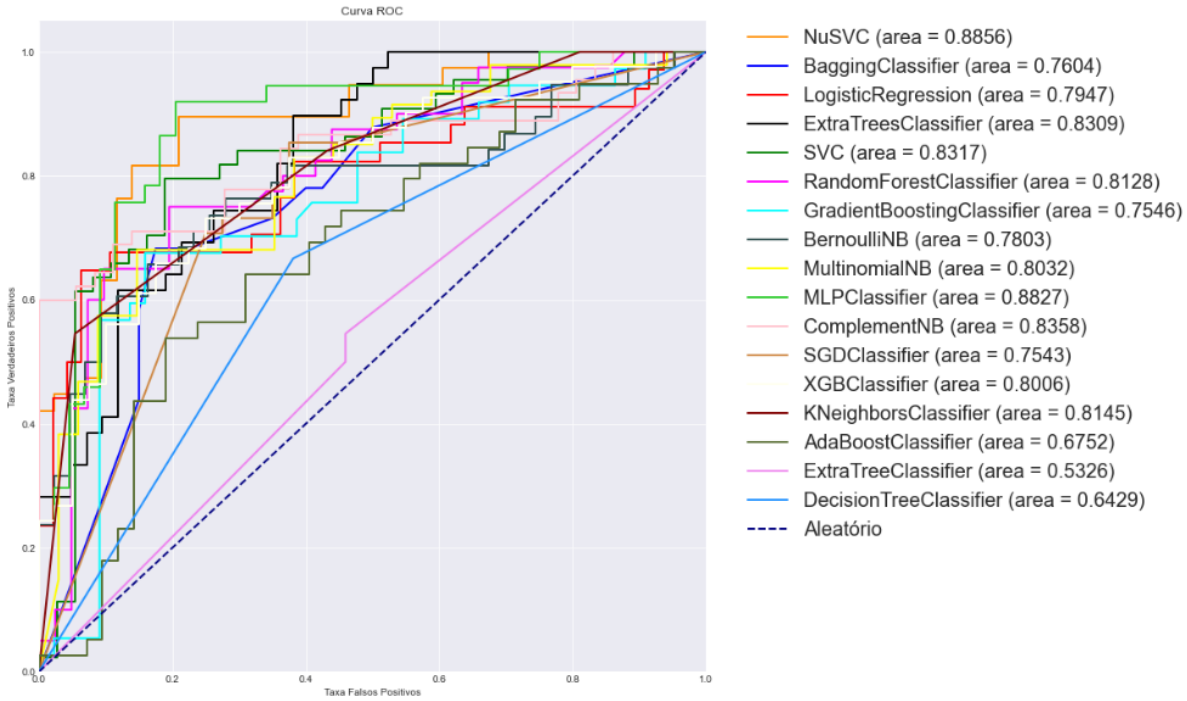

Figura 13 - Curva ROC da seleção do melhor experimento de cada classificador

Pela análise das curvas traçadas conclui-se que todos os classificadores possuíram desempenho superior a um classificador aleatório, (linha tracejada em azul escuro da Figura 13). Além disso notou-se que, apesar de alguns classificadores mostrarem desempenho ligeiramente superior em comparação com os demais, 14 desses alcançaram área abaixo da curva ROC acima de 0,75 e abaixo de 0,89. Apenas os classificadores ExtraTreeClassifier (0,53), DecisionTreeClassifier (0,64) e AdaBoostClassifier (0,67) alcançaram valores abaixo de 0,70.

5 CONCLUSÃO

5.1 Considerações Iniciais

Neste capítulo final são apresentadas as conclusões deste Trabalho de Conclusão de Curso e está dividido em três seções: na primeira são discutidos o que pode ser concluído com os resultados obtidos após a realização dos experimentos; na segunda são listadas as principais contribuições que a realização deste trabalho deu para a área de detecção de notícias falsas sobre a COVID-19 na rede social Facebook; e, por fim, na terceira seção são enumeradas diversas sugestões para a complementação do presente trabalho em projetos futuros.

5.2 Resultado dos experimentos

Ao interpretar os resultados evidenciados na seção anterior, podem ser feitas algumas afirmações. Primeiramente, como os resultados obtidos nas métricas de avaliação desses modelos não chegaram a valores expressivos, não é possível utilizar os algoritmos, da maneira que estão, para realização de classificação de notícias de maneira confiável no Facebook. No entanto, as métricas obtidas mostram que houve aprendizado por todos os algoritmos utilizados e a maioria deles apresentou resultados consideravelmente superiores à uma classificação aleatória. Dessa forma pode-se utilizar os algoritmos implementados neste trabalho em conjunto com outros algoritmos para construir um modelo com precisão aceitável na classificação de informações. Exemplo disso seria utilizar não apenas as legendas das postagens, mas também o texto das imagens e os metadados para alimentar o sistema de aprendizado de máquina, chegando-se assim a resultados superiores.

Outra constatação foi o fato de que a maneira que o texto foi pré-processado (seja com o texto sem pré-processamento ou com o texto pré-processado com *stopword*, *N-Grams* ou ambos) não surtiu diferenças relevantes para as métricas de classificação. O mesmo pode-se afirmar para os diferentes vetorizadores, extratores ou efeitos utilizados que apresentaram resultados semelhantes.

Por fim, após a realização dos experimentos nota-se que os algoritmos ExtraTreeClassifier , DecisionTreeClassifier e AdaBoostClassifier foram o que apresentaram os piores resultados e talvez sejam os menos apropriados para a realização de classificação binária de textos de legendas de postagens tendo o primeiro deles alcançado resultados próximos a um classificador aleatório.

5.3 Contribuições

Dada a necessidade de dados para os algoritmos de aprendizado de máquina, o primeiro passo para a implementação de um sistema de classificação de notícias é a construção de uma base de dados para alimentar esses algoritmos. Dessa forma, foram coletados a legenda de 401 postagens feitas no Facebook para essa finalidade. Além disso, essas postagens foram classificadas manualmente, com ajuda de informações disponibilizadas por agências de verificação de notícias, como verdadeiras ou falsas. Essa base de dados foi armazenada em um arquivo no formato *csv* e pode ser utilizada em outros trabalhos em português relacionados a notícias falsas, a COVID-19 ou ao Facebook.

Com a base de dados pronta, a etapa seguinte foi adaptar o código do Mateus Cabral, que foi feito para o Instagram, para o Facebook. Os códigos foram feitos na linguagem Python, foram divididos em etapas e salvos em arquivos no formato *ipynb*. Todos os arquivos estão comentados e evidenciam as bibliotecas utilizadas.

Além das acima citadas, este Trabalho de Conclusão de Curso também realizou as seguintes contribuições.

- Construção de modelos de classificação de informações em Português com enfoque para a rede social Facebook;
- Identificação de diversas informações inverídicas que foram espalhadas pela rede social Facebook durante o período da pandemia da COVID-19 até o momento.

5.4 Trabalhos Futuros

A seguir são enumeradas sugestões para trabalhos futuros relacionados ao tema do presente Trabalho de Conclusão de Curso:

Utilização do texto das imagens das postagens como dado contributivo para o aprendizado de máquina. A implementação de uma ferramenta de *Optical Character Recognition* (OCR) para capturar o texto presente nas imagens de uma postagem feita no Facebook poderia aumentar a acurácia do modelo, uma vez que, muitas vezes, o usuário que realiza a postagem assume que o usuário que irá ler a mesma vai abrir a imagem e ler o que estiver escrito na mesma. Por esse motivo, em muitas publicações, a informação que é disponibilizada na imagem não é repetida na legenda, o que resulta em enfraquecimento semântico do texto da legenda;

Utilização de metadados das postagens como dados contributivos para o aprendizado de máquina. Utilizar os metadados como data e hora da postagem, número de curtidas, comentários

e compartilhamentos como parâmetros, em conjunto com o texto da postagem, pode contribuir para melhorar a eficácia do modelo de aprendizado de máquina;

Análise de sentimentos dentro do Facebook levando em consideração o comportamento em termos de volume de postagem em uma linha do tempo, ou como foi a variação da quantidade das postagens dentro de um determinado período de investigação;

Formação das redes de compartilhamento de postagens e analise de sentimentos dentro do Facebook, por exemplo, como um perfil que segue outro perfil se comporta em diferentes postagens. Analisando também os comentários de cada uma das postagens. Portanto haveria a necessidade de extrair os comentários das postagens para analisar e classificar os comentários como positivos ou negativos das páginas do Facebook dos principais jornais do país.

6 REFERÊNCIAS BIBLIOGRÁFICAS

AHMED, S.; HINKELMANN, K.; CORRADINI, F. Combining Machine Learning with Knowledge Engineering to detect Fake News in Social Networks-a survey, 2019.

ALEXA Top Sites. **Alexa**, 2022. Disponivel em: <https://www.alexa.com/topsites>. Acesso em: 26 Janeiro 2022.

APUKE, O. D.; OMAR, B. Fake news and COVID-19: modelling the predictors of fake news sharing among social media users. **Telematics and Informatics**, 2021.

ATODIRESEI, C.-S.; TăNăSELEA, A.; IFTENE, A. Identifying Fake News and Fake Users on Twitter, 2018. 451–461.

BAGADE, A. et al. The Kauwa-Kaate Fake News Detection System: Demo. **Proceedings of the 7th ACM IKDD CoDS and 25th COMAD**, 2020. 302–306.

BANG, Y. et al. Model Generalization on COVID-19 Fake News Detection, 2021.

BOVET, A.; MAKSE, H. A. Influence of fake news in Twitter during the 2016 US presidential election, 2019. 1–14.

CABRAL, M. O. **Detecção de Postagens com Informações Falsas Sobre a Pandemia do COVID-19 na Rede Social Instagram**. São Carlos: [s.n.], 2021.

CHEN, B. et al. Transformer-based language model fine-tuning methods for covid-19 fake news detection, 2021.

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**, 2015. 1–4.

DAS, S. D.; BASAK, A.; DUTTA, S. A Heuristic-driven Ensemble Framework for COVID-19 Fake News Detection, 2021.

DIVYA, T. V.; BANIK, B. G. A Walk Through Various Paradigms for Fake News Detection on Social Media. **Proceedings of International Conference on Computational Intelligence and Data Engineering**, Dezembro 2020. 173-183.

DONG, X.; VICTOR, U.; QIAN, L. Two-Path Deep Semisupervised Learning for Timely Fake News Detection. **IEEE Transactions on Computational Social Systems (Volume: 7, Issue: 6, Dec. 2020)**, 2020. 1386-1398.

ELHADAD, M. K.; LI, K. F.; GEBALI, F. Fake News Detection on Social Media: A Systematic Survey. **2019 IEEE Pacific Rim Conference on Communications, Computers and Signal Processing (PACRIM)**, 2019.

FAUSTINI, P.; COVÕES, T. Fake News Detection Using One-Class Classification. **2019 8th Brazilian Conference on Intelligent Systems (BRACIS)**, 2019. 592–597.

FENNER, M. E. **Machine Learning With Python For Everyone**. [S.l.]: Addison-Wesley, 2019.

FERREIRA, J. R. S.; LIMA, P. R. S.; SOUZA, E. D. D. Desinformação, infodemia e caos social: impactos negativos das fake news no cenário da COVID-19. **Em Questão**, 2020. 30–53.

GALHARDI, C. P. et al. Fact or Fake? An analysis of disinformation regarding the Covid-19 pandemic in Brazil, 2020.

GIANSIRACUSA, N. **How Algorithms Create and Prevent Fake News:** Exploring the Impacts of Social Media, Deepfakes, GPT-3, and More. Acton, MA, USA: Apress, 2021.

GUNDAPU, S.; MAMIDI, R. Transformer based Automatic COVID-19 Fake News Detection System, 2021.

GUPTA, A. et al. Hostility Detection and Covid-19 Fake News Detection in Social Media, 2021.

INUWA-DUTSE, I. Towards combating pandemic-related misinformation in social media, 2021.

JAIN, A.; KASBE, A. Fake news detection, 2018. 1-5.

JÚNIOR, J. H. D. S. et al. Da desinformação ao caos: uma análiseuma análise das fake news frente à pandemia do coronavírus (covid-19) no Brasil. **Cadernos de Prospecção**, 2020. 331 – 346.

KIM, J. et al. FibVID: Comprehensive fake news diffusion dataset during the. **Telematics and Informatics**, Novembro 2021.

KOLLURI, N. L.; MURPHY, D. CoVerifi: A COVID-19 news verification system. **Online Social Networks and Media**, Março 2021.

KOLOSKI, B. et al. Identification of covid-19 related fake news via neural stacking, 2021.

LI, X. et al. Exploring Text-transformers in AAAI 2021 Shared Task: COVID-19 Fake News Detection in English, 2021.

LINDEN, S. V. D.; ROOZENBEEK, J.; COMPTON, J. Inoculating against fake news about covid-19, 2021.

MAHID, Z. I.; MANICKAM, S.; KARUPPAYAH, S. Fake news on social media: Brief reviewon detection techniques. **2018 Fourth International Conference on Advances in Computing, Communication & Automation (ICACCA)**, 2018.

MALLA, S.; ALPHONSE, P. J. A. Fake or real news about COVID-19? Pretrained transformer model to detect potential misleading news. **The European Physical Journal Special Topics**, 13 Janeiro 2022.

MLADENOVA, T.; VALOVA, I. Analysis of the KNN Classifier Distance Metrics for Bulgarian Fake News Detection. **2021 3rd International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)**, Junho 2021.

MONTEIRO, R. A. et al. Contributions to the Study of Fake News in Portuguese: New Corpus and Automatic Detection Results. **Computational Processing of the Portuguese Language**, 26 Agosto 2018. 324-334.

MONTI, F. et al. **Fake News Detection on Social Media using Geometric Deep Learning**. [S.l.]: [s.n.], 2019.

MORI, S.; NISHIDA, H.; YAMADA, H. **Optical Character Recognition**. [S.l.]: Wiley-Interscience, 1999.

NAJAR, F.; ZAMZAMI, N.; BOUGUILA, N. Fake news detection using bayesian inference. **2019 IEEE 20th International Conference on Information Reuse and Integration for Data**, 2019. 389–394.

NETO, M. et al. Fake news no cenário da pandemia de covid-19. **Cogitare Enfermagem**, 2020.

NEWMAN, N. et al. **Reuters Institute Digital News Report 202**. Reuters Institute. [S.l.]. 2021.

NYOW, N. X.; CHUA, H. N. Detecting fake news with tweets' properties. **2019 IEEE Conference on Application, Information and Network Security (AINS)**, 2019. 24–29.

OLALEYE, T. et al. Veracity Assessment of Multimedia Facebook Posts for Infodemic Symptom Detection using Bi-modal Unsupervised Machine Learning Approach. **International Journal for Research in Applied Science & Engineering Technology (IJRASET)**, Dezembro 2021.

PARIKH, S. B.; ATREY, P. K. Media-Rich Fake News Detection: A Survey. **2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)**, 2018.

PARK, M. **Facebook Reports Third Quarter 2021 Results**. Facebook. [S.l.]. 2021.

PATWA, P. et al. Fighting an Infodemic: COVID-19 Fake News Dataset, 2021.

PÉREZ-ROSAS, V. et al. Automatic Detection of Fake News, 2017. 3391–3401.

PIERRI, F.; CERI, S. False news on social media: A data-driven survey. **SIGMOD Record,** 2019. 18–32.

PREVITI, M. et al. Fake news detection using time series and user features classification. In: CASTILHO, P. A.; LAREDO, J. L. J.; VEGA, F. F. D. **Applications of Evolutionary Computation**. [S.l.]: Springer International Publishing, 2020. p. 339–353.

RASHKA, S.; MIRJALILI, V. **Python Machine Learning:** Machine Learning and Deep Learning With Python, Schikit-Learn and TensorFlow 2. Third Edition. ed. Birmingham - Mumbai: Packt Publishing, 2019.

RECUERO, R.; GRUZD, A. Cascatas de Fake News Políticas: um estudo de caso no Twitter, São Paulo, 2019. 31-47.

REIS, J. C. S. et al. Explainable Machine Learning for Fake News Detection, 2019. 17-26.

REIS, J. C. S. et al. Supervised learning for fake news detection, 2019. 76–81.

ROVETTA, A.; BHAGAVATHULA, A. S. Global Infodemiology of COVID-19: Analysis of Google Web Searches and Instagram Hashtags. **Journal of Medical Internet Research**, 2020.

SAMPLE, C.; JUSTICE, C.; DARRAJ, E. Fake news: A method to measure distance from fact. **2018 IEEE International Conference on Big Data (Big Data)**, 2018. p. 4443–4452.

SANTIA, G. C.; WILLIAMS, J. R. BuzzFace: A News Veracity Dataset with Facebook User Commentary and Egos, 15 Junho 2018.

SHRIVASTAVA, S. et al. A Research on Fake News Detection Using Machine Learning Algorithm. In: A.K., S.; MUNDRA A., D. R.; S., B. **Smart Systems:** Innovations in Computing. Singapura: Springer, v. 235, 2022. p. 273-287.

SHU, K. et al. FakeNewsNet: A Data Repository with News Content, Social Context and Spatialtemporal Information for Studying Fake News on Social Media, 27 Março 2019.

SHU, K. et al. FakeNewsNet: A Data Repository with News Content, Social Context and Spatialtemporal Information for Studying Fake News on Social Media, 27 Março 2019.

SHU, K. et al. **Disinformation, Misinformation, and Fake News in Social Media:** Emerging Research Challenges and Opportunities. [S.l.]: Springer, 2020.

SHU, K.; BERNARD, H. R.; LIU, H. **Studying Fake News via Network Analysis:** Detection and Mitigation. [S.l.]: Springer International Publishing, 2019.

SILVA, C. D. D. et al. Can Machines Learn to Detect Fake News? A Survey Focused on Social Media, 2019.

SINGH, L. et al. A first look at COVID-19 information and misinformation sharing on Twitter, 2020.

SINGHAL, S. et al. **SpotFake:** A Multi-modal Framework for Fake News Detection. [S.l.]: [s.n.], 2019.

SULLIVAN, W. **Machine Learning Beginers Guide Algorithms:** Supervised & Unsupervised Learning, Decision Tree & Random Forest Introduction. [S.l.]: [s.n.], 2017.

TACCHINI, E. et al. Some Like it Hoax: Automated Fake News Detection in Social Networks, 25 Abril 2017.

TANDOC, E. C.; LIM, Z. W.; LING, R. Defining “fake news”: A typology of scholarly definitions. **Digital Journalism**, 2018. 137–153.

TRAYLOR, T. et al. Classifying Fake News Articles Using Natural Language Processing to Identify In-Article Attribution as a Supervised Learning Estimator. **2019 IEEE 13th International Conference on Semantic Computing (ICSC)**, 2019. 445–449.

TSCHIATSCHEK, S. et al. Fake News Detection in Social Networks via Crowd Signals. **Companion Proceedings of the**, 2018. 517–524.

VAJJALA, S. et al. **Pratical Natural Language Processing:** Comprehensive Guide to Building Real-World NLP Systems. [S.l.]: O'Reilly Media, 2020.

VASILIEV, Y. **Natural Language Processing With Python and SpaCy:** A Pratical Introduction. San Francisco: [s.n.], 2020.

VOSOUGHI, S.; ROY, D.; ARAL, S. The spread of true and false News online, 2018. 1146–1151.

WANG, W. Y. "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection, 2017.

WANI, A. et al. Evaluating deep learning approaches for covid19 fake news detection, 2021.

YAFOOZ, W. M. S.; EMARA, A.-H. M.; LAHBY, M. Detecting Fake News on COVID-19 Vaccine from YouTube Videos Using Advanced Machine Learning Approaches. In: LAHBY M., P. A. K. . M. Y. . Y. W. M. S. **Combating Fake News with Computational Intelligence Techniques**. [S.l.]: Springer, 2022. p. 421-435.

ZAREI, K. et al. A First Instagram Dataset on COVID-19, 25 Abril 2020.

ZHOU, X.; ZAFARANI, R. **Network-based Fake News Detection:** A Pattern-driven Approach. [S.l.]: [s.n.], 2019.