[1]

Detecção de fake news utilizando inteligência artificial

Felipe Martins Francisco - 163714
Pedro Henrique Fava da Silva - 163969

I. Resumo

Nosso trabalho tem o objetivo de treinar um algoritmo BERT para identificar se uma notícia é fake news ou real, utilizando o aprendizado não-supervisionado e supervisionado. Utiliza-se de uma extensão do Chrome em que essa retornará se a notícia selecionada pelo usuário é fake ou não.

Palavras-Chaves: BERT, SVM, API, Extensão Chrome.

II. Introdução e motivação

A. Introdução

Na era digital, a disseminação de informações ocorre em uma escala sem precedentes, frequentemente sem qualquer tipo de verificação ou filtro. Esse fluxo constante de dados apresenta um desafio significativo: como distinguir informações verídicas de desinformação. Reconhecendo a necessidade de uma solução acessível e eficaz, propomos o desenvolvimento de uma extensão para o navegador Chrome, projetada para identificar e sinalizar fake news de forma rápida e prática (utilizando da inteligência artificial).

B. Motivação

Este relatório tem como objetivo explorar e avaliar o papel da inteligência artificial (IA) no combate às fake news, abordando tanto métodos clássicos quanto a implementação de algoritmos pré-treinados. A urgência deste estudo está fundamentada na crescente sofisticação das técnicas de criação e disseminação de desinformação, que se espalham rapidamente, influenciando negativamente a sociedade, especialmente em momentos críticos, como eleições. Diante disso, torna-se imperativo desenvolver soluções tecnológicas que não apenas detectem, mas também acompanhem e se adaptem à velocidade das fake news. Este estudo busca, portanto, identificar e analisar as abordagens mais eficazes no uso da IA para mitigar os efeitos nocivos da desinformação, contribuindo para uma sociedade mais informada e resiliente.

III. Reconhecimento de Padrões

A. Estudo fake news

Para iniciar nossa análise, buscamos compreender as principais diferenças entre uma fake news e uma notícia verdadeira. Nessa investigação, consultamos pesquisas relevantes e diretrizes governamentais. Segundo o Tribunal Regional Eleitoral de São Paulo (TRE-SP) [3] , uma fake news frequentemente apresenta características como títulos sensacionalistas ou bombásticos e erros ortográficos ou gramaticais. Além disso, estudos anteriores que abordam o tema têm utilizado o conteúdo textual, especialmente os títulos, para determinar a veracidade das notícias. Com base nesse fundamento, nossa abordagem também focará na análise da informação principal da notícia para avaliar sua autenticidade.

B. Trabalhos relacionados

Para fundamentar o presente estudo, realizamos uma pesquisa sobre a relação entre fake news e machine learning. Entre os artigos relevantes encontrados, destacam-se “Liar, Liar Pants on Fire” de William Yang e “Comparação entre Modelos com Diferentes Abordagens para a Classificação de Fake News” de Lucas Cordeiro. Ambos enfatizam a importância da análise textual na detecção de fake news e destacam a eficiência dos algoritmos SVM e BERT nesse contexto, os quais serão testados ao longo deste trabalho.

O artigo “Liar, Liar Pants on Fire” [2] explora o uso do algoritmo LIAR, demonstrando sua eficácia na resolução de problemas político-sociais ao combinar modelos de classificação clássicos com o uso de metadados. Este estudo apresenta uma abordagem inovadora, mostrando o potencial dessa metodologia na detecção de desinformação.

Por sua vez, o artigo “Comparação entre Modelos com Diferentes Abordagens para a Classificação de Fake News” [1] avalia a eficácia de vários algoritmos, destacando a superioridade do BERT para essa tarefa. O estudo reforça a capacidade do BERT em lidar com a complexidade dos textos, tornando-o uma ferramenta poderosa para a classificação de fake news.

[1]

IV. Objetivos

O objetivo principal deste relatório é apresentar um algoritmo inovador capaz de identificar a veracidade de notícias. Este algoritmo será implementado em uma extensão de navegador, permitindo que os usuários selecionem qualquer texto online para verificar sua autenticidade. Os objetivos específicos incluem:

1. Desenvolver um algoritmo eficiente;
2. Integrar o algoritmo em uma extensão de navegador;
3. Avaliar a precisão e a eficácia do algoritmo através de comparações entre algoritmos existentes.
4. Fornecer uma interface amigável que permita aos usuários selecionar e verificar textos de forma intuitiva e rápida.
5. Promover a conscientização sobre a importância da verificação de fatos e combater a disseminação de desinformação online.

V. Metodología experimental

a. Banco de dados

Para o treinamento da nossa inteligência artificial, utilizamos um conjunto de dados composto por um subconjunto de notícias e desinformações (fake news). O banco de dados consiste em 100 linhas, das quais 50 correspondem a notícias verificadas e 50 a fake news. Essas informações foram extraídas de diversas plataformas, incluindo YouTube, Facebook, Telegram, WhatsApp, X e Instagram, garantindo uma diversidade de fontes e contextos.

O banco de dados contém duas colunas principais:

- Corpo da Notícia: Texto completo da notícia ou desinformação.
- Classificação: Indicação de se a entrada é uma fake news ou não.

Este conjunto de dados foi projetado para fornecer uma base sólida para o treinamento e avaliação de modelos de detecção de notícias falsas, promovendo a precisão e a eficácia no processo de identificação de desinformação.

b. Pré-processamento e retirada de padrões (clássicos)

Inicialmente, avaliamos o desempenho de nosso projeto utilizando modelos clássicos, aplicando uma série de tratamentos de dados ao corpo das notícias para otimizar a precisão das previsões. Os seguintes procedimentos foram realizados:

Tratamentos de Dados:

1. Remoção de Domínios e URLs: Eliminamos elementos como "http" e "www" presentes no texto copiado.
2. Remoção de Caracteres Especiais: Foram removidos caracteres especiais, como pontuação, emojis e símbolos.
3. Eliminação de Números: Números foram retirados do texto.
4. Tokenização: O texto foi segmentado em palavras individuais (tokens).
5. Stemização: Aplicamos a stemização para reduzir palavras à sua raiz comum.

Extração de Informações Cruciais:

Com base em pesquisas que indicam que fake news frequentemente apresentam características chamativas e capciosas, extraímos parâmetros específicos para identificar essas características. Esses parâmetros incluem:

1. Contagem de Emojis: Quantificamos o número de emojis presentes no texto.
2. Porcentagem de Letras Maiúsculas: Calculamos a proporção de letras maiúsculas no texto.
3. Contagem de Erros Ortográficos: Utilizamos o SpellChecker para identificar e contar erros ortográficos no corpo da notícia.
4. Contagem de Palavras Frequentes: Identificamos palavras que aparecem pelo menos 5 vezes em todas as fake news e criamos uma lista com essas palavras. Contamos a frequência de cada palavra em cada linha do texto para análise posterior.

Esses procedimentos de pré-processamento e análise de padrões foram implementados para melhorar a eficácia dos modelos de detecção, assegurando uma melhor capacidade de distinguir entre notícias verdadeiras e fake news.

c. Entendimento e análise

Nossa principal tarefa foi classificar os dados para identificar de forma clara quando uma notícia era falsa ou verdadeira, utilizando parâmetros extraídos do corpo do texto. Durante a avaliação, observou-se que os parâmetros de erros ortográficos e contagem de emojis não apresentavam uma correlação significativa com a categorização de fake news. Essa falta de relação clara foi ilustrada pelos gráficos apresentados a seguir.

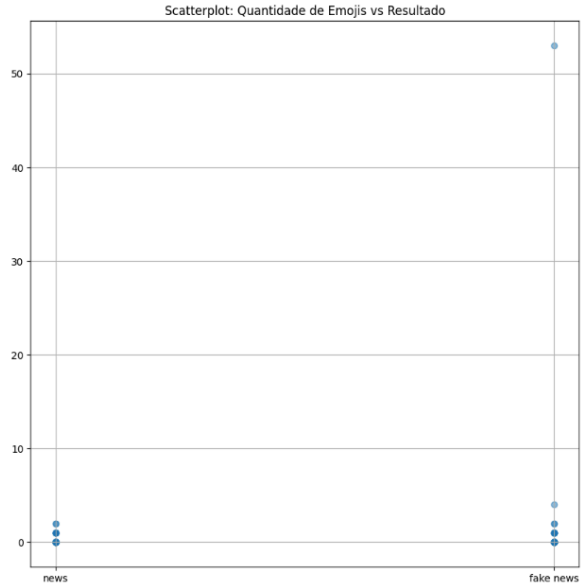
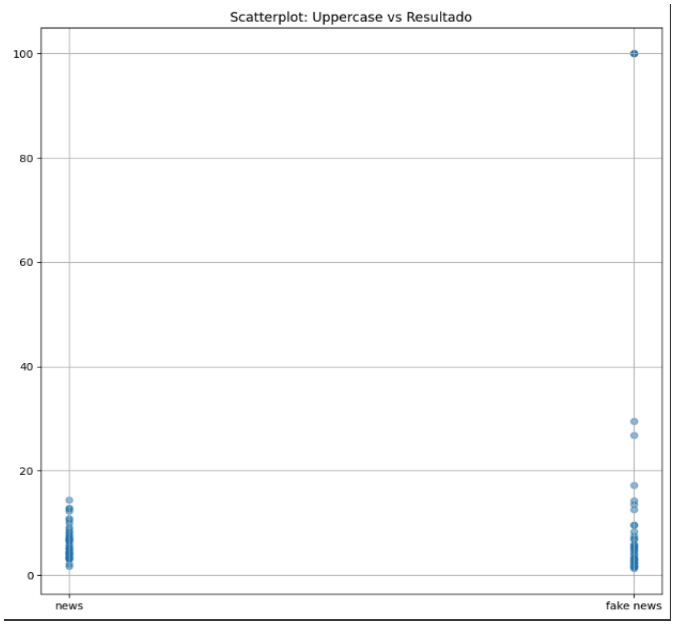


Gráfico que mostra a relação entre quantidade de emojis e fake news.



Relação de porcentagem de letras maiúsculas com fake news.

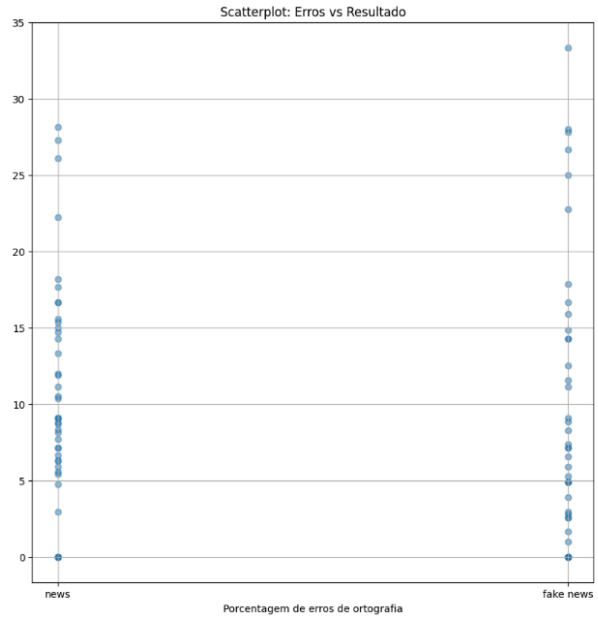
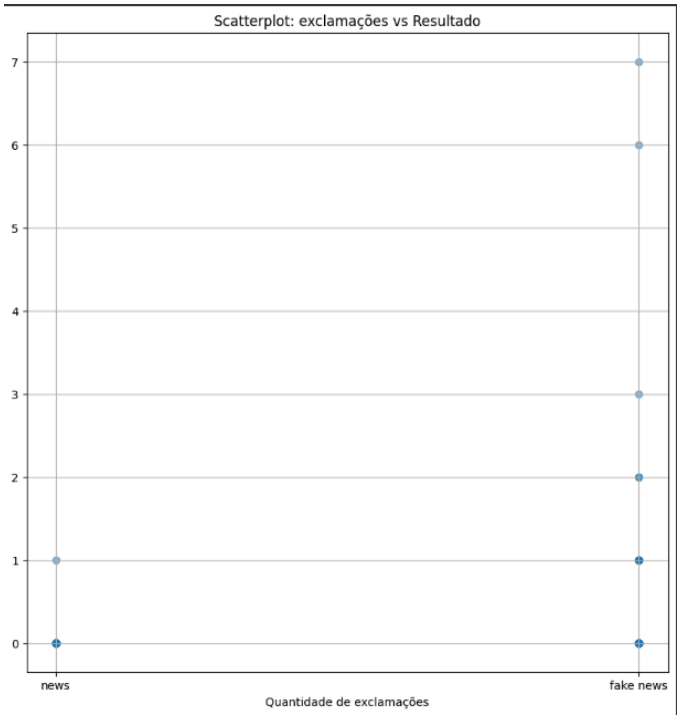


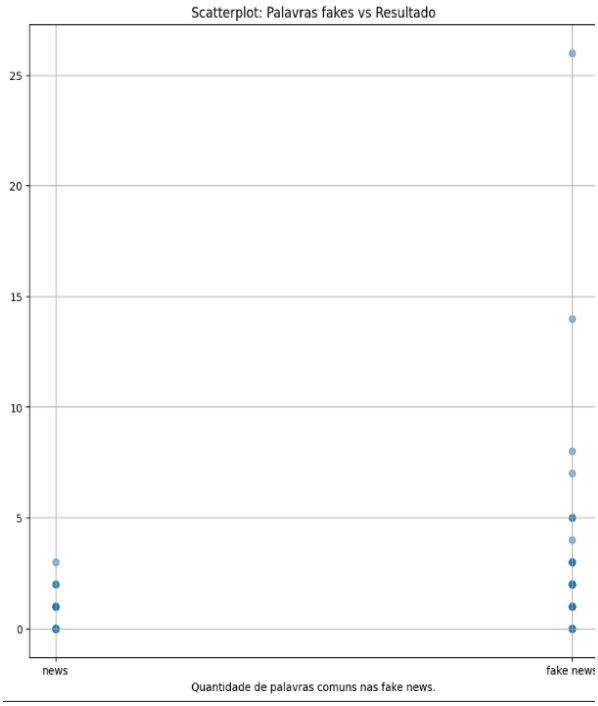
Gráfico que mostra a relação entre fake news e quantidade de erros de ortografia

Dessa forma, foram retirados esses dois parâmetros do nosso projeto, focando nos outros que apresentaram uma boa correlação.

Os demais parâmetros tiveram uma boa correlação e foram mantidos. Seguem os parâmetros e seus respectivos gráficos:



Relação de quantidade de exclamações no texto e fake news.



Relação da quantidade de palavras do texto que aparecem na lista e fake news.

Assim, esses parâmetros foram escolhidos inicialmente para o treinamento do nosso modelo.

d. *Treinamento (clássicos)*

1. Metodologia

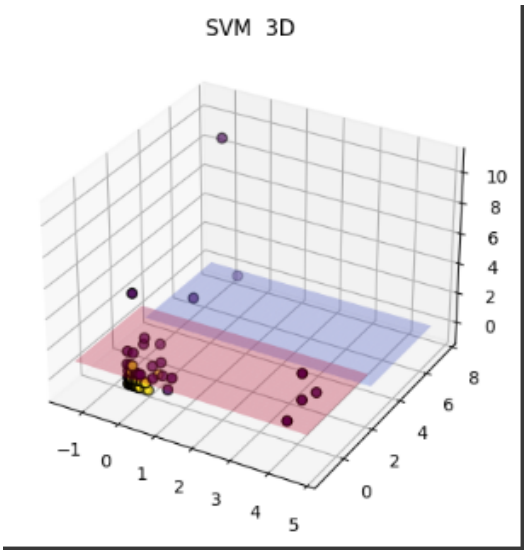
Para realizar o treinamento dos modelos, seguimos os seguintes passos:

1. Reclassificamos a coluna 'RESULTADO', atribuindo o valor 1 para 'fake news' e 0 para 'news'.
2. Dividimos os dados em conjuntos de treino (80%) e teste (20%).
3. Aplicamos o StandardScaler para normalizar os dados, evitando problemas durante o treinamento.
4. Testamos vários modelos de machine learning utilizando GridSearch para otimização de hiperparâmetros.
5. Implementamos validação cruzada com 10 K-Folds, realizando até 30 iterações, alterando o fator randômico do K-Fold, resultando em 300 cenários distintos.
6. Treinamos os seguintes modelos: DecisionTree, RandomForest, KNN e SVM.

2. Resultados

1. DecisionTree: Acurácia média de 63%, com máximo de 69,86% e desvio padrão de 3,1%.
2. RandomForest: Acurácia média de 71,45%, com máximo de 77,6% e desvio padrão de 3%.
3. KNN: Acurácia média de 72,7%, com máximo de 76,6% e desvio padrão de 2%.
4. SVM: Acurácia média de 75%, com máximo de 80% e desvio padrão de 2%.

Com base nos resultados, utilizando kernel linear, o modelo SVM apresentou o melhor desempenho entre os modelos clássicos testados. A configuração final do modelo SVM pode ser descrita conforme segue:



A relatório de classificação desse SVM pode ser analisada em seguida:

	precision	recall	f1-score
Classe fake	0.87	0.65	0.74
Classe news	0.73	0.90	0.81

1. A precisão para a classe fake é de 87%, indicando que quando o modelo indica uma fake news, muito provavelmente está correta.
2. A precisão para a classe news é de 73%, indicando que não é tão eficaz nessa classe quanto na outra.
3. A revocação para a classe fake é de 65%. Há um número considerável de notícias falsas que não foram identificadas pelo modelo.
4. A revocação para a classe news é de 90%, indicando que o modelo conseguiu identificar 90% das notícias verdadeiras corretamente. Isso sugere que o modelo é muito eficaz na detecção de notícias verdadeiras.

e. BERT

Para aumentar a porcentagem de acerto, foi utilizado o algoritmo *BERT*. O *BERT* foi escolhido por ser um modelo pré-treinado poderoso capaz de entender as palavras dentro de um texto, entendendo o contexto em que elas aparecem.

Dessa forma, podem retirar informações cruciais dos nossos textos retirados de notícia e ajudar-nos a treinar um modelo mais eficaz.

1. Somente BERT

   O primeiro teste foi feito entregando nosso banco de dados para o *BERT*. Dessa forma, ele foi capaz de extrair informações do nosso banco e treinou um modelo capaz de prever se uma notícia é fake ou não.

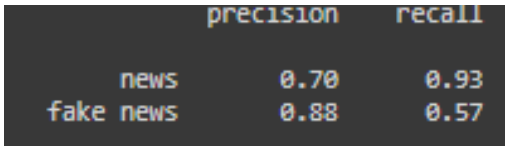
   Para isso:

   1. Foi feito a tokenização e preparação dos dados, utilizando a classe *FakeNewsDetection* e o método *__getitem__;*
   2. Chamada do modelo *BertForSequenceClassification,* configuração dos parâmetros e os parâmetros de resultados.

Após ter os parâmetros retirados dos nossos dados, alimentamos o modelo pré-classificado BERT para que ele se adapte aos nossos dados, utilizando um modelo próprio para classificar duas classes possíveis. Além disso, foi usado as seguintes técnicas para aumentar a acurácia:

1. Uso de GPU: o modelo é movido para a GPU disponível para acelerar o treinamento,
2. Loteamento personalizado: utiliza uma função para assegurar o mesmo tamanho entre lotes, para que o BERT seja treinado corretamente.
3. AdamW: otimização que ajuda o overfitting e melhora o modelo.
4. Validação para cada época: em diferentes épocas, resgatamos a época com maior acurácia.

O treino foi feito em três épocas, onde o melhor resultado de acurácia foi de 82%.



	precision	recall
news	0.70	0.93
fake news	0.88	0.57

Para a classe 'news', 70% das instâncias classificadas como 'news' realmente pertenciam a essa classe. O recall de 93% reflete a capacidade do modelo de identificar corretamente a maioria das instâncias de 'news', com apenas 7% sendo incorretamente classificadas como 'fake news'.

Por outro lado, o modelo demonstrou uma precisão de 88%, sugerindo que 88% das previsões para 'fake news' estavam corretas, com 12% sendo falsos positivos. No entanto, o recall de 57% evidencia uma limitação significativa, onde 43% das instâncias de 'fake news' foram erroneamente classificadas como 'news'.

Esse baixo recall para 'fake news' indica uma área crítica para aprimoramento do modelo, dada a importância de detectar essas instâncias com maior acurácia.

2. BERT/SVM + DADOS PRÉ-PREPARADOS.

Uma segunda hipótese para ser testada é juntar os parâmetros retirados manualmente com os que o BERT criou. Para isso:

1. Concatenamos os parâmetros extraídos pelo *BERT* com os parâmetros antigos.
2. Treinar o modelo SVM (que foi eficaz da última vez) com esses novos parâmetros.
3. Treinar o modelo BERT (usando as mesmas técnicas e ferramentas) com esses novos parâmetros.

Concluindo, o modelo SVM teve maior acurácia do que o modelo BERT treinado. O modelo clássico teve 76% de acurácia, enquanto o modelo BERT teve, na sua melhor época, 64%.

f. API e sua extensão

Para a aplicação prática do modelo, desenvolvemos uma extensão para o Google Chrome, que permite ao usuário selecionar um texto de interesse diretamente no navegador. Ao selecionar o texto, a extensão captura essa informação e a envia para um modelo de classificação previamente treinado por meio de uma API dedicada. O modelo então processa o texto e retorna uma previsão sobre a veracidade da notícia, classificando-a como verdadeira ou fake news. Essa previsão é apresentada diretamente na interface da extensão, permitindo uma verificação rápida e eficiente da informação pelo usuário.

V. Resultados e discussões

1. Desempenho do Modelo

O melhor modelo avaliado foi o BERT utilizando parâmetros retirados por ele. Dessa forma,

atingiu-se uma acurácia de 82%. O outro modelo que atingiu um resultado satisfatório foi o de SVM, treinado pelos parâmetros retirados manualmente. No fim, o melhor modelo e o escolhido para integrar a API foi o BERT, por conseguir se adaptar de maneira mais eficaz ao dado em si.

Esses resultados sugerem que, apesar do bom desempenho geral, o modelo ainda possui limitações na detecção de fake news, o que é crítico para aplicações onde a precisão dessa identificação é vital. A baixa taxa de recall para a classe "fake news" pode levar a uma sub detecção de notícias falsas, comprometendo a eficácia da solução em cenários reais.

2. Aplicações e impactos

Com base nos resultados, o modelo foi integrado a uma extensão do Google Chrome, permitindo que usuários verifiquem a veracidade de notícias diretamente em seus navegadores. Ao selecionar um texto, a extensão envia os dados para o modelo via API, e a classificação é retornada ao usuário em tempo real. Esta implementação prática demonstrou-se eficaz na facilitação da verificação de notícias, embora as limitações no recall da classe "fake news" sugerem que melhorias futuras são necessárias para aumentar a confiabilidade do sistema em detecções críticas.

Dado o caráter político do tema, é fundamental garantir a precisão na execução deste trabalho. A relevância desse tópico se destaca na luta contra a disseminação de fake news, que representa uma ameaça significativa para várias áreas da democracia.

VI. Conclusões e trabalhos futuros

O desenvolvimento e a implementação do modelo de classificação baseado no BERT demonstraram avanços significativos na detecção de fake news em textos. O modelo alcançou uma precisão notável para a classe "fake news", com 88%, embora o recall para esta classe tenha sido mais baixo, 57%. Por outro lado, o modelo teve um desempenho sólido na identificação de "news", com um recall de 93%, mas com uma precisão de 70%, refletindo a necessidade de refinamento para reduzir falsos positivos.

A integração do modelo em uma extensão do Google Chrome oferece uma ferramenta prática para a verificação de notícias em tempo real, demonstrando a aplicabilidade do modelo no combate à desinformação. A implementação mostrou que o modelo pode fornecer insights úteis aos usuários, embora haja espaço para aprimoramento.

A expansão do conjunto de dados aumenta a diversidade e o volume do conjunto de dados de treinamento, podendo ajudar a melhorar o desempenho do modelo. Incluir exemplos mais variados de fake news pode ajudar o modelo a generalizar melhor.

Além disso, incorporar feedback dos usuários da extensão pode fornecer insights valiosos sobre a eficácia do modelo e áreas de melhoria. Os ajustes baseados em feedback real podem ajudar a otimizar a performance do modelo em cenários do mundo real.

References

[1] BRASIL, Lucas Cordeiro.” Comparação entre modelos com diferentes abordagens para classificação de fake news”. Biblioteca Digital de Teses e Dissertações da UFCG, Campus Campina Grande, v. 1, n. 1, p. 13, out. 2021. Acesso em: 10 jul. 2024.

[2] ANG, William Yang. “Liar, Liar Pants on Fire”: A New Benchmark Dataset for Fake News Detection. ArXiv, v. 1705.00648, 2017. Acesso em: 10 jul. 2024.

[3] Fernandes, Márcia, ‘*Saiba como identificar fake news ou desinformação’*. Acesso em: 10 jul. 2024. [Online] Available: https://www.tre-sp.jus.br/comunicacao/noticias/2023/Agosto/saiba-como-identificar-fake-news-ou-desinformacao

[4] ‘*Scikit-learn: Machine Learning in Python’*. Acesso em: 10 jul. 2024. [Online] Available: https://scikit-learn.org/stable/

[5] ‘*BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding’*. Acesso em: 10 jul. 2024. [Online] Available: https://arxiv.org/abs/1810.04805

[6] ‘*Support Vector Machines’*. Acesso em: 10 jul. 2024. [Online] Available:https://scikit-learn.org/stable/modules/svm.html

[7] ‘*KFold’*. Acesso em: 10 jul. 2024. [Online] Available:https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.KFold.html

[8] 3.1. Cross-validation: evaluating estimator performance. Acesso em: 10 jul. 2024. [Online] Available:https://scikit-learn.org/stable/modules/cross_validation.html

[9] SILVA, Maria João; PEREIRA, Ana Catarina. Fake news: uma abordagem sobre o impacto da desinformação na sociedade. Revista Portuguesa de Enfermagem de Saúde Mental, n. 19, p. 12-20, 2018. Disponível em: https://scielo.pt/scielo.php?script=sci_arttext&pid=S2183-54622018000100012. Acesso em: 2 set. 2024.