

## 14º Congresso de Inovação, Ciência e Tecnologia do IFSP - 2023

### ANÁLISE PREDITIVA DE MENSAGENS FALSAS EM REDES SOCIAIS RELACIONADOS A VACINA DE COVID-19

VINICIUS MARTINS VIEIRA<sup>1</sup>, DÉBORA LUIZA MACIEL ALMEIDA<sup>2</sup>, EDUARDO LUIZ MARINHO<sup>3</sup>, THIAGO PEDRO DONADON HOMEM<sup>4</sup>, ADRIANO JOSE FERRUZZI<sup>5</sup>

<sup>1</sup> Graduando em Tecnologia em Análise e Desenvolvimento de Sistemas, IFSP - Câmpus Pirituba, v.martins@aluno.ifsp.edu.br

<sup>2</sup> Graduando em Tecnologia em Análise e Desenvolvimento de Sistemas, IFSP - Câmpus Pirituba, debora.luiza@aluno.ifsp.edu.br

<sup>3</sup> Graduando em Tecnologia em Análise e Desenvolvimento de Sistemas, IFSP - Câmpus Pirituba, e.marinho@aluno.ifsp.edu.br

<sup>4</sup> Docente em Redes de Computadores integrado ao ensino médio, IFSP - Câmpus Pirituba, thiagohomem@ifsp.edu.br

<sup>5</sup> Docente em Redes de Computadores integrado ao ensino médio, IFSP - Câmpus Pirituba, adriano.ferruzzi@ifsp.edu.br

<sup>6</sup> Grupo de informática e Tecnologia em Educação e Sociedade, GITES

Área de conhecimento (Tabela CNPq): 1.03.03.04-9 Sistemas de Informação

**RESUMO:** O problema com notícias falsas nas redes sociais não é recente e mostrou-se extremamente danoso durante o período em que ocorreu a pandemia de COVID-19. Ainda nessa linha de pesquisa, podemos identificar que quanto menor o nível de escolaridade, maior a dificuldade de reconhecer notícias falsas na internet e, conseqüentemente, maior a disseminação delas neste grupo. Neste contexto, o objetivo deste trabalho é focar precisamente no desenvolvimento de uma solução que utiliza modelos de aprendizado de máquina para identificar notícias falsas em mensagens de redes sociais e avaliar a performance desses modelos. O estudo tem como finalidade classificar *tweets* que possuem maiores indícios de *fake news* com relação às vacinas de COVID-19, dessa forma, o algoritmo pode colaborar com o combate à desinformação nas redes sociais. Por fim, uma base de dados com 585 mensagens sobre vacina de COVID-19 foi classificada e utilizada para treinamento e avaliação dos modelos, fazendo uso de técnicas de tokenização e lematização para tratamento e a técnica de vetorização da frequência do termo – inverso da frequência (TF-IDF) foi aplicada nos dados. Os algoritmos de classificação utilizados foram Naive Bayes, Árvore de Decisão e Regressão Logística com *f1-score* de: 0,85; 0,80 e 0,84; respectivamente, mostrando a eficiência dos modelos.

**PALAVRAS-CHAVE:** inteligência artificial; processamento de linguagem natural; aprendizado de máquina; análise preditiva; algoritmos de classificação.

**ABSTRACT:** The issue of fake news on social media is not recent and has proven to be highly detrimental during the period of the COVID-19 pandemic. Furthermore, within this line of research, it can be observed that the lower the level of education, the greater the difficulty in recognizing false information on the internet, consequently leading to a higher dissemination of such content within this demographic. In this context, the objective of this work is precisely focused on the development of a solution that employs machine learning models to identify fake news in social media messages and assess the performance of these models. The study aims to classify tweets that exhibit stronger indications of being fake news regarding Covid-19 vaccines. By doing so, the algorithm can contribute to the battle against misinformation on social media. Finally, a dataset containing 585 messages about COVI-19 vaccines were labeled and utilized for model training and evaluation, employing techniques like tokenization and lemmatization for data processing and Term Frequency — Inverse Data Frequency (TF-IDF) for data vectorization. The classification algorithms employed were Naive Bayes, Decision Tree, and Logistic Regression, achieving f1-scores of 0.85, 0.80, and 0.84, respectively, showcasing the effectiveness of the models.

**KEYWORDS:** artificial intelligence; natural language processing; machine learning; predictive analysis; classification algorithms.

## INTRODUÇÃO

A desinformação disseminada nas redes sociais sobre a vacina da COVID-19 ocorreu de diversas formas notícias relacionadas a efetividade, origem, espalhamento do vírus, interesses financeiros, efeitos colaterais, de forma geral, informações incompletas e distorcidas. O rastreamento dessas informações é muito custoso e demorado, pois ocorrem de forma descentralizada e assim torna extremamente difícil para um usuário comum distinguir rapidamente se a informação é falsa ou verdadeira (Sousa, 2020).

A desinformação sobre a COVID-19, assim como a liberação do uso emergencial das vacinas, trouxe questionamentos sobre a ciência para o público geral, o resultado disso foi um aumento na desconfiança da vacina juntamente com a escalada da pandemia (Ferreira Caceres et al., 2022).

Portanto, visto que a rede social Twitter possui uma fonte valiosa de informações, este estudo tem como finalidade classificar tweets que possuem maiores indícios de *fake news* com relação às vacinas de COVID-19, dessa forma o algoritmo pode colaborar com o combate à desinformação nas redes sociais.

O trabalho está desenvolvido com preceitos na literatura, baseando-se em artigos científicos e estudos de casos reais. O modelo foi desenvolvido utilizando a linguagem de programação Python em conjunto com o uso de bibliotecas de tratamento de dados como o Pandas e de bibliotecas de aprendizado de máquina como o *scikit learn*. Para atingir tais objetivos, o trabalho está dividido em etapas em que a primeira se trata deste preâmbulo que apresenta a justificativa da temática e objetivos. Em seguida, são apresentados os materiais e métodos que apresentam trabalhos relacionados a respeito do uso de aprendizado de máquina no uso de detecção de *fake news*, seguido da explicação a respeito da coleta, tratamento e classificação dos dados que são partes fundamentais do trabalho. Por fim, em resultados, a avaliação dos modelos utiliza técnicas estatísticas para interpretação dos resultados e avaliação da performance dos modelos de forma objetiva, permitindo identificar o modelo com a melhor precisão na classificação das *fake news*.

## MATERIAL E MÉTODOS

As principais etapas do projeto podem ser classificadas como: 1. Levantamento de artigos semelhantes, 2. Coleta dos Dados, 3. Pré-processamento dos dados, 4. *Feature Engineering* e Vetorização, 5. Treinamento e Avaliação dos Modelos.

### Metodologia da Revisão Bibliográfica

Este estudo foi baseado em estudos teóricos e práticos, partindo de uma revisão de literatura e levantamento bibliográfico de natureza qualitativa, centrado-se na compreensão dos trabalhos relacionados ao tema. Nesse sentido, o estudo foi realizado com base em artigos publicados e já revisados por pares nos últimos 5 anos, ou seja, de 2018 a 2023, nas bases Periódicos Capes e Google Acadêmico.

A respeito da seleção dos artigos, foram utilizadas as seguintes palavras-chave nos portais: *fake news*, *machine learning*, *vaccine*, covid-19. Para o desenvolvimento da revisão de literatura foram selecionados os artigos que mais se assemelham com os objetivos deste estudo. Sobre os critérios de inclusão, foram utilizados artigos nos idiomas de português e inglês relacionados ao tema, compreendidos na faixa temporal estipulada e que tiveram proximidade com o contexto deste estudo. Em relação aos critérios de exclusão, foram desconsiderados os artigos que não estavam nos idiomas determinados, fora da faixa de tempo estabelecida e que não traziam a temática base para esta pesquisa.

### Coleta dos Dados

Uma coleção de mensagens (*tweets*), relacionados às vacinas COVID-19, foi coletada usando a Tweepy API. Essas mensagens foram publicadas entre 01 de janeiro de 2021 e 31 de maio de 2021. O período destacado é importante para análise, pois contém o início da vacinação e a chegada das primeiras vacinas no Brasil. Estas foram acompanhadas de muitas informações falsas e campanhas de conscientização e estímulo à vacinação.

No processo de coleta dos *datasets*, alguns conjuntos de palavras chaves foram definidos para otimizar o processo de coleta dos dados evitando mensagens cujo o foco do conteúdo não fosse vacina, então os conjuntos utilizados foram: 1. vacina sem eficácia, 2. vacina sem comprovação, 3. vacina

cobaia, 4. vacina. Desta maneira, foi possível obter um *dataset* entre informações verdadeiras e falsas nas mensagens, o que melhora a performance dos modelos de aprendizado de máquina.

Além disso, foram utilizados outros filtros nas coletas, o filtro de idioma foi definido para português, as duplicatas foram removidas, por fim os dados estão armazenados em um *database* MongoDB na nuvem.

### Pré-processamento dos dados

Esta primeira etapa de tratamento visa filtrar dados que não são úteis nas análises, assim como padronizar o formato dos dados. Dessa forma, o primeiro passo é realizar a remoção de elementos não textuais e padronização dos dados, descapitalização de todas as palavras, remoção de caracteres especiais e pontuação, remoção de espaços múltiplos. Em seguida é realizado o tratamento de abreviações, substituindo por palavras correspondentes, melhorando assim a qualidade dos dados. Após essa limpeza inicial dos dados, ocorre a remoção de palavras de parada (*stopwords*), remoção de ruídos, essas são palavras comumente usadas em um idioma e possuem pouca ou nenhuma importância na classificação das mensagens, para isso foi utilizado o dicionário de *stopwords* do pacote *Natural Language Toolkit* (nltk). Por fim, foi realizada a etapa de tokenização e lematização dos dados, em que os tokens foram gerados por palavras após o tratamento dos dados, as palavras geradas passam por um processo de lematização, uma técnica avançada para diminuir o ruído dos dados, com essa técnica as palavras são convertidas a sua forma equivalente ou original sem perda de significado, reduzindo número de tokens e facilitando o treinamento do modelo. O pacote *spacy* com a base de dados “pt\_core\_new\_sm” foi utilizado para realizar esse tratamento.

### Feature Engineering e Vetorização dos Dados

Para transformação dos elementos textuais em uma matriz numérica com o peso de cada termo (*feature*) para processamento dos algoritmos de *machine learning*, foi utilizada a técnica TF-IDF (*Term Frequency - Inverse Data Frequency*). Aplicado o algoritmo da *sklearn* com *n-grams* igual a 1, unigramas, assim foi obtido um mapeamento de um para um entre tokens e termos resultantes da vetorização. A equação básica utilizada no modelo para o cálculo do TF-IDF é mostrada a seguir:

Equação para o cálculo do TF-IDF:

$$tf - idf(t, d) = tf(t, d) \times idf(t) \quad (1)$$

em que:

t - termo;

d - documento;

tf(t,d) - frequência do termo no documento;

idf(t) - inverso da frequência do termo.

Equação para o cálculo do inverso da frequência do termo:

$$idf(t) = \log \frac{(1+n)}{1+df(t)} + 1 \quad (2)$$

em que,

t - termo;

idf(t) - inverso da frequência nos dados do termo avaliado;

n - número total de documentos na coleção de dados.

df(t) - frequência do termo avaliado;

Ao final do processo o resultado do TF-IDF é normalizado pela distância euclidiana.

Equação para normalização dos dados pela distância Euclidiana:

$$w_{norm} = \frac{w}{\sqrt{w_1^2 + w_2^2 + \dots + w_n^2}} \quad (3)$$

### Treinamento do Modelo e Avaliação

Os algoritmos de classificação implementados para avaliação da performance foram: Árvore de Decisão, Naive Bayes e Regressão Logística. O *dataset* foi previamente classificado e rotulado pelos integrantes do grupo para treinamento e avaliação do modelo. Assim, 75% do *dataset* foi utilizado para treinamento do modelo e o restante, 25%, foi utilizado para o teste dos modelos. Os modelos foram avaliados pelas métricas de precisão, revocação e *f1-score*.

Para o cálculo dessas métricas precisamos comparar os resultados do *dataset* de teste, onde os resultados obtidos pelo modelos foram comparados com os resultados esperados. As seguintes equações foram utilizadas para calcular as métricas:

Equação para o cálculo da acurácia.

$$Acurácia = \frac{VP + VN}{VP + FP + VN + FN} \quad (4)$$

Equação para o cálculo da precisão.

$$Precisão = \frac{VP}{VP + FP} \quad (5)$$

Equação para o cálculo da revocação.

$$Revocação = \frac{VP}{VP + FN} \quad (6)$$

Equação para o cálculo do *f1-score*.

$$f1 - score = \frac{2 \times Precisão \times Revocação}{Precisão + Revocação} \quad (7)$$

em que:

- VP - número de elementos classificados como verdadeiros positivos;
- VN - número de elementos classificados como verdadeiros negativos;
- FP - número de elementos classificados como falso positivos;
- FN - número de elementos classificados como falso negativos;

### RESULTADOS E DISCUSSÃO

Os modelos foram avaliados com uma base de dados de 585 mensagens, sendo 364 mensagens classificadas como desinformação e 221 como mensagens verdadeiras, assim, 62% das mensagens dos *dataset* estão rotuladas como positivas com relação a desinformação, ou seja possuem desinformação sobre as vacinas de COVID-19. O volume das amostras podem ser observados na ilustração da Figura 1.

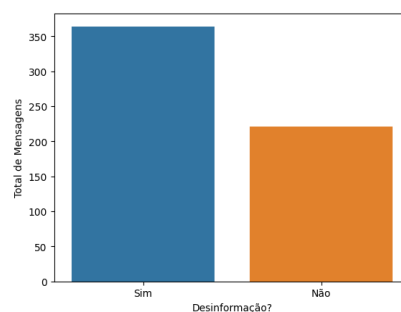


FIGURA 1 - Volume das classes do dataset.

O resultado dos modelos está descrito na tabela abaixo:

TABELA 1. Resultados das análises de performance dos modelos de predição.

Teste	Naive Bayes	Árvore de Decisão	Regressão Logística
Acurácia	0,79	0,74	0,76
Precisão	0,78	0,79	0,74
Revocação	0,95	0,82	0,98
f1-score	0,86	0,81	0,84

Analisando os resultados, pode-se observar que o modelo Naive Bayes apresentou a melhor performance, com a maior acurácia (0,79) e o maior f1-score (0,86). A Árvore de Decisão teve a pior performance entre os modelos testados, com a menor acurácia (0,74) e o menor f1-score (0,81). A Regressão Logística teve uma performance intermediária tanto na acurácia (0,76) quanto no f1-score (0,84). Vale ressaltar que o f1-score é de extrema importante para avaliar o desempenho dos modelos, pois considera tanto a precisão quanto a revocação, sendo especialmente útil quando as classes estão desbalanceadas.

Os resultados obtidos pelos diversos modelos encontram-se em linha com os resultados observados na literatura, apesar destes estudos analisarem a eficácia de modelos em outros idiomas.

Hayawi et al. (2021) observou resultados com *f1-score* de 0,58 a 0,98 classificando mensagens de diversas fontes de dados diferentes com relação a desinformação de COVID-19. Sendo que o *f1-score* mais baixo ocorreu com fonte de dados de mídias sociais e sites de desinformação, enquanto o melhor resultado ocorreu em um fonte de dados proprietária do estudo com mensagens relevantes e aplicação de um modelo BERT na análise.

Khanam et al. (2020) realizou um estudo semelhante com aplicação de diversos algoritmos para identificar notícias falsas, algoritmos como *XGboost*, *Random Forest*, *k-Nearest Neighbors*, Árvore de Decisão e *Support Vector Machine* (SVM). Todos apresentaram acurácia na faixa de 0,65 e 0,75. Com destaque para *XGboost* com 0,75 de acurácia, SVM e *Random Forest* apresentaram 0,73 de acurácia, em seguida Árvore de Decisão com 0,71 e *Naive Bayes* obteve 0,69 de acurácia.

Assim é possível perceber que os resultados obtidos encontram-se dentro do esperado, apesar do *dataset* limitado à poucas amostras, de acordo com o apresentado na Tabela 1. Como a análise e aplicação dos modelos depende de uma série de fatores como pré-processamento dos dados, tratamento e eliminação de ruídos, podendo conter ainda tratamento mais complexos como *stemming*, ou seja, extração da raiz das palavras, ou um processo de lematização como aplicado neste estudo, em que as palavras são transformadas em outras para demonstrar a origem do seu significado, assim ao tratar os *tokens*, é natural que pequenas divergências possam ocorrer nos resultados com relação a outros estudos semelhantes.

Além disso, outra fonte de atenção aos resultados é o idioma, em que ambos estudos citados, não focaram em mensagens em português, mas sim em inglês. Assim, o resultado quanto a avaliação de performance dos modelos se mostra em níveis satisfatórios, ao ser comparado com o resultado obtido pelos estudos semelhantes citados, contendo ainda uma margem significativa de aprimoramento, principalmente com relação a calibração de modelos com *datasets* maiores, onde seria possível utilizar configurações mais avançadas de *n-grams* com dados suficiente para calibrar mais elementos (*features*) nos modelos, além de implementar novos modelos de aprendizado de máquina.

A seguir estão relacionados os links do repositório da aplicação utilizada na elaboração deste trabalho, o repositório contém todos os módulos citados no artigo e ainda contém uma versão do *dataset* em arquivo CSV que pode ser usado para replicar o ambiente utilizado na elaboração deste trabalho.

<<https://github.com/vinicius-mv/FakeNewsVacinas>>.

<[https://github.com/vinicius-mv/FakeNewsVacinas/blob/main/src/datamanagement/bkps/tcc\\_tweets\\_10-22-2023\\_bkp.csv](https://github.com/vinicius-mv/FakeNewsVacinas/blob/main/src/datamanagement/bkps/tcc_tweets_10-22-2023_bkp.csv)>.

## CONCLUSÕES

O objetivo deste estudo foi a aplicação de modelos de aprendizado de máquina supervisionado para a classificação de mensagens de redes sociais com foco em mensagens sobre vacinas de COVID-19. A solução computacional desenvolvida aborda diversos modelos e técnicas de coleta, análise e tratamento dos dados, eliminação de ruídos, aplicação de modelos de *machine learning* e cálculo de métricas para avaliar os modelos que atingiram o propósito do estudo.

É possível utilizar solução com algoritmos e aprendizagem para auxiliar o combate às notícias falsas e desinformação, apesar de os modelos atuarem com um filtro inicial, eles ainda não são capazes de eliminar a intervenção humana.

A respeito dos modelos de predição analisados, o modelo de *Naive Bayes* de maneira geral performou melhor que seus pares neste estudo. Com destaque para o seu índice de revocação, indicando assim além de um bom percentual de acerto dos verdadeiros positivos (possuem desinformação e o algoritmo classificou corretamente), um baixo índice de falsos negativos (possuem desinformação, mas o algoritmo classificou como verdadeira a mensagem). O que é especialmente importante em uma possível aplicação de algoritmos de *machine learning* na análise de mensagens nas redes sociais para atuarem no combate e prevenção na disseminação de notícias falsas. Futuramente, o trabalho utilizará de técnicas de explicabilidade para determinar como os algoritmos tomaram as decisões e quais as variáveis mais relevantes para a classificação.

## AGRADECIMENTOS

Os autores agradecem ao Instituto Federal de Educação, Ciência e Tecnologia de São Paulo - Câmpus Pirituba e ao corpo docente pela dedicação, conhecimento e pela oportunidade de estudo. E um agradecimento especial aos professores pela orientação e apoio contínuo.

## REFERÊNCIAS

- FERREIRA CACERES, M. M.; SOSA, J. P.; LAWRENCE, J. A.; SESTACOVSKI, C.; TIDD-JOHNSON, A.; RASOOL, M. H. U.; GADAMIDI, V. K.; OZAI, S.; PANDAV, K.; CUEVAS-LOU, C.; PARRISH, M.; RODRIGUEZ, I.; FERNANDEZ, J. P. **The impact of misinformation on the COVID-19 pandemic**. AIMS Public Health, [s. l.], v. 9, n. 2, p. 262–277, 2022.
- HAYAWI, K.; SHAHRIAR, S.; SERHANI, M. A.; TALEB, I.; MATHEW, S. S. **ANTI-Vax: a novel Twitter dataset for COVID-19 vaccine misinformation detection**. Public Health, [s. l.], v. 203, p. 23–30, 2022.
- KHANAM, Z.; ALWASEL, B. N.; SIRAFI, H.; RASHID, M. **Fake News Detection Using Machine Learning Approaches**. IOP Conference Series: Materials Science and Engineering, [s. l.], v. 1099, n. 1, p. 012040, 2021.
- SOUSA, D. **62% dos brasileiros não sabem reconhecer fake news, diz pesquisa**. 2020. Disponível em: <<https://canaltech.com.br/seguranca/brasileiros-nao-sabem-reconhecer-fake-news-diz-pesquisa-160415/>>. Acesso em: 18 nov. 2022.