

# O uso de Machine Learning na classificação de textos com ênfase em Fake News.

Antonio L. Basile, Marco P. Altieri, Paulo H. Apro.

Faculdade de Computação e Informática  
Universidade Presbiteriana Mackenzie (UPM) – São Paulo, SP – Brasil  
antonioluiz.basile@mackenzie.br, marcoaltieri96@hotmail.com,  
pauloapro@gmail.com

**Abstract.** *Currently, the term Fake News is increasingly entering our daily lives, and news travels great distances in a short period of time on the internet and social media. The present work sought to compare the Random forest, Decision Tree and Logistic Regression classification algorithms in the classification of news described as fake news. aiming at exploring these models and their effectiveness.*

**Keywords:** *Fake News, Random forest, Decision tree, Logistic regression, news classification, Machine Learning.*

**Resumo.** *Atualmente o termo Fake News vem adentrando cada vez mais em nosso cotidiano, e as notícias percorrem grandes distancias em um período curto de tempo na internet e nas mídias sociais. O presente trabalho buscou comparar os algoritmos de classificação Random forest, Árvore de decisão e Regressão logística na classificação de notícias descritas como fake news. visando a exploração desses modelos e sua efetividade.*

**Palavras – chave:** *Fake News, Random forest, Árvore de decisão, Regressão logística, classificação de notícias, Aprendizado de Máquina.*

## 1. Introdução

A deflagração de notícias falsa, popularmente chamadas de fake news, tornaram-se uma grande preocupação devido à sua capacidade de criar impactos devastadores no que se refere às desinformações que são disseminadas no mundo virtual. Nos últimos anos as aplicações de aprendizado de máquina apresentaram sucesso no reconhecimento de padrões, por esse razão elas estão sendo cada vez mais utilizadas para a análise de textos, principalmente em textos de notícias falsas (El Naqa et al 2015).

Devido à facilidade que as pessoas têm em acessar, compartilhar e comentar, às notícias são mais fáceis e rápidas de serem consumidas. (Shu et al 2017). Segundo (Bondielli et al 2019) o compartilhamento de conteúdo falso tem maior adesão dos usuários nas mídias sociais do que páginas que têm conteúdo jornalístico real. No Brasil, 92% dos usuários de internet estão em alguma mídia social (C. Rock, 2018).

Neste contexto, a utilização de ferramentas de machine learning podem contribuir para identificar a disseminação de notícias falsas. Desta forma, o presente trabalho objetivou comparar algoritmos de aprendizado de máquina utilizados para classificação de fake news.

## **2. Objetivo Geral**

O presente trabalho busca aplicar e comparar ferramentas de machine learning utilizadas nos processos de identificação e classificação de Fake News.

## **3. Objetivo Específico**

Comparar os algoritmos de classificação Random forest, Árvore de decisão e Regressão logística na classificação de notícias. Visando a exploração desses modelos e sua efetividade, levado a um possível direcionamento para futuras pesquisas sobre Fake News.

## **4. Referencial Teórico**

### **4.1. Fake News.**

A expressão fake news é relativamente recente e sua caracterização se faz necessária, pois existem diferentes interpretações para fake news. Mesmo com as mídias tradicionais, já existiam pessoas que divulgavam notícias falsas propositalmente. Independentemente do surgimento, fake news apresenta muitas definições que podem ser divididas em dois grupos. (Golbeck et al., 2018).

No primeiro grupo é considerado o aspecto intencional, pois define as publicações intencionais e verificadas como falsas. Assim, não basta a notícia ser falsa para ser caracterizada uma fake news, é preciso ser intencional, podendo também ser divulgada em uma mídia digital (Zhou & Zafarani, 2020).

O segundo grupo, classifica as fake news como notícias falsas independentemente da sua natureza intencional, como, por exemplo, falta de interpretação sobre o conteúdo abordado (Sharma et al., 2019). programadas para determinadas ações, utilizam algoritmos complexos para a melhor tomada de decisão.

### **4.2. Detecção de notícias falsas usando aprendizado de máquina.**

Muitos estudos já estão focados em detectar fake news, mas ainda dependem da detecção manual dessas notícias, milhares de artigos são publicados e removidos a cada segundo na internet, o que torna inviável manualmente.

Uma solução poderia ser o desenvolvimento de um sistema para fornecer uma pontuação de índice automatizada confiável, ou classificação de credibilidade de diferentes editores e contexto de notícias (Z Khanam et al., 2021).

Segundo Zeba Khanam, pesquisador em Harbin Institute of Technology, Shenzhen, China, em seu artigo “ Fake News Detection Using Machine Learning Approaches, 2021” propôs uma análise relacionada à detecção de notícias falsas explorando os modelos tradicionais de aprendizado de máquina para escolher o melhor.

### **4.3. Linguagem natural.**

O processamento da linguagem natural é uma área de inteligência artificial que tem por objetivo a interpretação e manipulação das línguas humanas, envolve traduzir linguagem natural em dados numéricos que um computador pode usar para aprender sobre o mundo, é um processo chamado de vetorização. (Lane et al 2019). O bag of

words é uma técnica que transformará um texto original em um conjunto de palavras e a frequência que uma palavra aparece no texto é calculada.

A saída do método é uma matriz em que cada coluna representa uma palavra no vocabulário e a linha corresponde a um texto. E ao final sairá o número de vezes que a palavra aparece. (Ghosh et al 2019).

De acordo com Freire e Goldschmidt (2019, apud Laura D et al., 2020) a identificação automática de notícias falsas pode ser entendida como um problema de classificação binária, em que uma notícia é dada como uma entrada ( $\epsilon$ ), a tarefa de identificar notícias falsas é prever se essa notícia é falsa ou não, ou seja,  $f$ :  $\epsilon$  de modo que  $f$  é a função de previsão.

#### **4.4. Árvore de decisão**

Árvores de decisão são modelos estatísticos que compreende a uma série de decisões lógicas, utilizados em problemas de predição supervisionada, onde um conjunto de atributos é utilizado para prever o valor de um atributo de saída (resultado), sendo o mapeamento destas entradas para a saída denominado modelo preditivo.

O algoritmo utilizando a representação de uma árvore, dividida em raiz, nós internos e folhas. A árvore é construída da raiz para as folhas, ou seja, num método top-down. (Oliveira, Érick Ritir 2019).

#### **4.5. Random forest.**

Random Forests é um tipo específico de ensemble learning no qual treinamos várias árvores de decisão com um único conjunto de dados, ou seja, Random forest é uma coleção de árvores de decisão. Cada árvore é treinada com um subconjunto do conjunto de dados e as diversas árvores obtidas constituirão um único classificador (Uma Sharma et al., 2021).

#### **4.6. Regressão logística.**

Também chamado de Regressão Logit, é utilizada para estimar valores discretos (valores binários como 0/1, sim/não, verdadeiro/falso) com base em determinado conjunto de variáveis independentes.

Segundo Aurélien Géron (2019, p.139) estima a probabilidade de uma instância pertencer a uma determinada classe (por exemplo, qual é a probabilidade desse e-mail ser spam?). Se a probabilidade estimada for maior que 50%, então o modelo prevê que a instância pertence a essa classe (chamada de classe positiva, rotulada como “1”), ou então ela prevê que não (isto é, pertence à classe negativa, rotulada “0”). Isso o transforma em um classificador binário.

### **5. Metodologia**

Para a realização desta trabalho optou-se por dividi-lo em duas etapas. A primeira etapa se inicia com um levantamento bibliográfico sobre o tema da pesquisa. A comparação e análise da fundamentação teórica foi fundamental para a escolha das ferramentas e algoritmos utilizados para a realização deste trabalho. A partir do levantamento de artigos

e livros, seguimos para segunda etapa. Identificamos os modelos tradicionais de aprendizado de máquina mais utilizados para classificação de fake news. Os algoritmos de classificação aplicados para comparação e análise foram Árvore de Decisão, Random Forests e Regressão logística.

### 5.1. Desenvolvimento.

Neste estudo, utilizou como linguagem de programação Python e suas bibliotecas Scikit-Learn e como ambiente de desenvolvimento foi usado o Google Colab, um framework online onde se pode escrever, executar códigos de Deep Learning e Machine Learning. O Python possui um enorme conjunto de bibliotecas e extensões, que podem ser facilmente usadas em Machine Learning. Segundo Uma Sharma et al., 2021 a biblioteca Scikit-Learn é a melhor fonte para algoritmos de aprendizado de máquina, onde quase todos os tipos de algoritmos de aprendizado de máquina estão prontamente disponíveis para Python, portanto, é possível uma avaliação fácil e rápida de algoritmos de Machine Learning.

A base de dados utilizada para treinamento e testes dos algoritmos neste estudo é da Kaggle.com, uma plataforma de comunidade online da Google para cientistas de dados que, disponibiliza conjuntos de dados de notícias falsas e reais para trabalhos e pesquisas, todos os artigos e notícias retirados da base de dados para este trabalho são de origem norte americana de língua inglesa. Segundo Huang, Jeffrey (2020), esses conjuntos de dados têm sido amplamente utilizados em diferentes trabalhos de pesquisa para determinar a veracidade de notícias.

O conjunto de dados contém 3 arquivos no formato .CSV denominados Train.csv para treinamento, Fake.csv e True.csv para teste. Todo o conjunto de dados de notícias contém 23 mil linhas de notícias descritas como sabidamente falsas e 21 mil linhas de notícias sabidamente verdadeiras. Os dados referem-se às categorias de notícias variadas de política, governo e mundo, demonstradas em porcentagem na Figura 1.

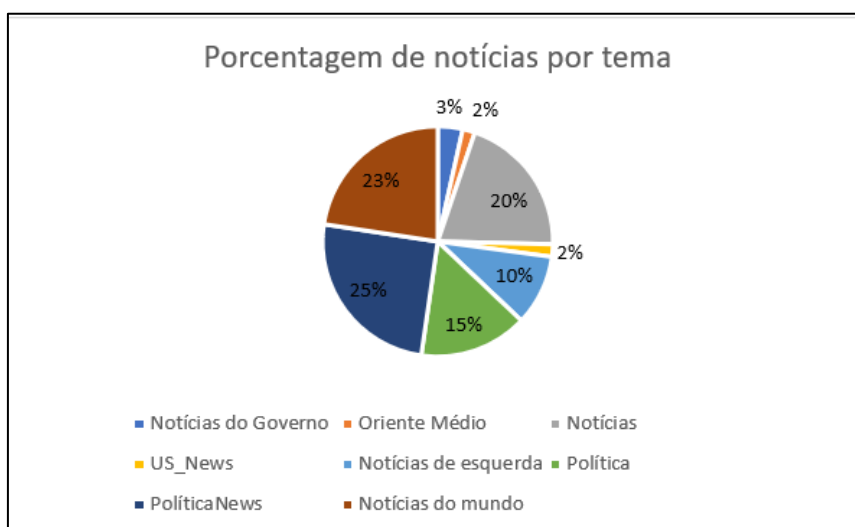


Figura 1. Notícias. Fonte: Autoria Própria.

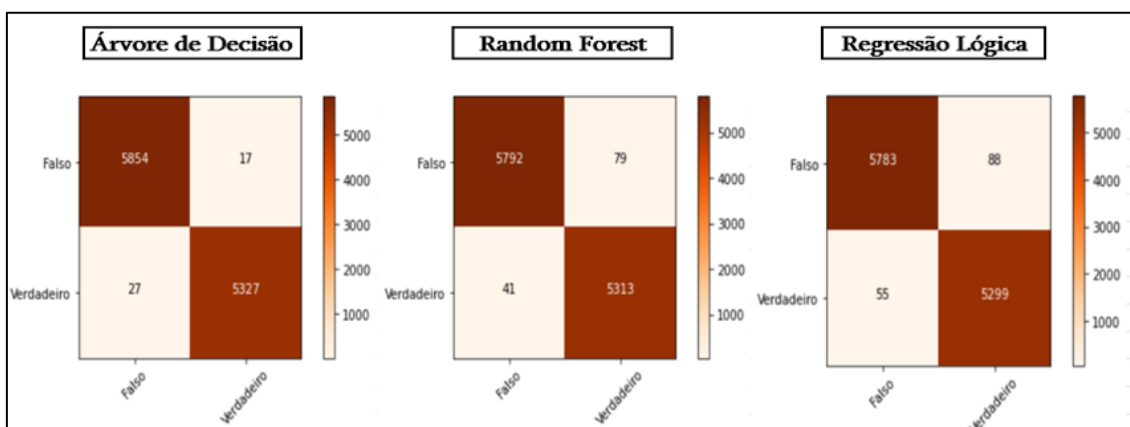
Após a fase de coleta de dados, é realizado o pré-processamento dos dados utilizando o CountVectorizer, usado para transformar um determinado texto em um vetor com base na frequência (contagem) de cada palavra que ocorre em todo o texto. A limpeza dos dados também inclui Tokenização para separar o texto em unidades, como frases ou palavras. Utilização de Stopwords para remover palavras irrelevantes, o TfidfTransformer é aplicado no corpo do texto, de modo que a contagem relativa de cada palavra nas frases é armazenada na matriz do documento, ou seja, calcula a frequência com que um termo aparece em um documento. Ao final do pré-processamento, a quantidade de dados é reduzida para análise de testes, chegando em torno de 12 mil dados analisados.

Para medir a performance dos três modelos classificadores utilizou-se da acuracidade para tratar o total de acertos obtidos em relação ao total de experimentos e, avaliando a qualidade dos resultados obtidos, a matriz de confusão apresenta os dados reais e a previsibilidade dos algoritmos. A precisão, Erro, Recall e F1-Score, também foram tratados para medir a performance dos modelos.

## **5.2. Resultados.**

Neste estudo, foram analisados os desempenhos de modelos de aprendizagem de máquina, Árvore de decisão, Random Forest e Regressão Lógica. Aproximadamente 11.225 dados foram testados para comparação entre os modelos. A Matriz de confusão para todos os algoritmos está representada abaixo na figura 2, mostrando duas linhas e duas colunas que descrevem a quantidade de falsos positivos, falsos negativos, verdadeiros positivos e verdadeiros negativos.

Nos testes realizados a Árvore de decisão obteve um total de acertos de 11.181, dos quais 5854 identificados como fake news e 5327 identificados como notícias verdadeiras. Random Forest registrou um total de 11.105 acertos, dos quais 5792 identificados como fake news e 5313 identificados como notícias verdadeiras. A Regressão Lógica indicou um total de 11.082 acertos, dos quais 5783 identificados como fake news e 5299 identificados como notícias verdadeiras. Quanto ao total de erros registrados nos testes, o modelo de Árvore de decisão obteve os menores erros dentre os modelos, 44 erros em relação a Random Forest com 120 erros e Regressão Lógica com 143 erros.



**Figura 2. Resultados da Matriz de Confusão. Fonte: Autoria Própria.**

Ao medir a performance dos modelos com as técnicas de Acuracidade, Erro, Recall, Precisão e F1-Score, vistos na Tabela 1. Identificou-se uma precisão de 0,99 em todos os modelos analisados quanto a precisão, além do F1-Score que obteve 0,99 para Árvore de decisão e 0,98 para Random Forest e Regressão Lógica. Estes valores de precisão poderão ser entendidos como modelos de precisão alta, poucas taxas de falso positivo e falso negativo, vistos na matriz de confusão, figura 2. Destaca-se a acuracidade do modelo Árvore de decisão com 99,61%, seguido dos modelos de Random Forest 98,93% e Regressão Lógica 98,73%.

Modelos	Acuracidade	Total Acertos	Total Erro	Recall	Precisão	F1-Score
Árvore de decisão	99,61%	11181	44	0,99	0,99	0,99
Random Forest	98,93%	11105	120	0,98	0,99	0,98
Regressão Lógica	98,73%	11082	143	0,98	0,99	0,98

**Tabela 1. Comparação da performance para todos os três modelos.**

## 6. Conclusão

O presente trabalho buscou comparar os algoritmos de classificação Random forest, Árvore de decisão e Regressão logística na classificação de notícias descritas como fake news. Visando a exploração desses modelos e sua efetividade.

Ao analisar os resultados obtidos pode-se concluir que, dentre os algoritmos selecionados, a Árvore de decisão foi o que demonstrou o melhor desempenho na classificação de fake news da base supervisionada, registrando uma acuracidade de 99,61% e obtendo um bom desempenho na matriz de confusão com os menores erros. Não muito distantes a acuracidade registradas nos modelos Random Forest e Regressão Lógica foram de 98,93% e 98,73% respectivamente.

Estes valores poderão ser entendidos como modelos de precisão alta, em comparação com os resultados obtidos por Uma Sharma et al., (2021) utilizando o Twitter

como fonte de dados, a Regressão Logística e Random Forest tiveram precisão de 65% e 59% respectivamente. De acordo com outras obras utilizadas como base para o projeto e analisando seus resultados, a variedade de algoritmos de aprendizagem de máquinas demonstram uma grande capacidade de detectar fake news. É possível que no futuro as técnicas de aprendizado de máquinas se tornarão um método padrão na identificação de notícias intencionalmente falsas.

## 7. Referências

Bondielli, A., & Marcelloni, F. (2019). A survey on fake news and rumour detection techniques. *Information Sciences*, 497, 38-55. Disponível em: <<https://www.sciencedirect.com/science/article/abs/pii/S0020025519304372?via%3Dihub>>. Acesso em: 2 abril 2022.

C. Rock. Social media trends 2018. Technical report, 2018. Disponível em: <<https://cdn2.hubspot.net/hubfs/355484/Ebooks%20MKTC/Social%20Media.pdf>>. Acesso em: 2 abril 2022.

El Naqa, I., & Murphy, M. J. (2015). What is machine learning?. In *machine learning in radiation oncology* (pp. 3-11). Springer, Cham. Disponível em: <<https://books.google.com.br/books?hl=ptBR&lr=&id=1N7yCQAAQBAJ&oi=fnd&pg=PR5&dq=What+is+machine+learning%3FIn+machine+learning+in+radiation+oncology&ots=n4qabanL4&sig=Ml5b9XRke5aNZJGgDBhReMq0KTK#v=onepage&q=What%20is%20machine%20learning%3FIn%20machine%20learning%20in%20radiation%20oncology&f=false>>. Acesso em: 2 abril 2022.

Freire, P. M. S., & Goldschmidt, R. R. (1). Combate automático às Fake News nas mídias sociais virtuais: uma revisão do estado da arte. *Revista Militar De Ciência E Tecnologia*, 38(2), 3-12. Disponível em: <<http://ebrevistas.eb.mil.br/CT/article/view/8639>> Acesso em: 2 abril 2022.

GÉRON, Aurélien. *Hands-On Machine Learning with Scikit-Learn and TensorFlow*. Sebastopol: O'Reilly, 2017.

Ghosh, S., & Gunning, D. (2019). *Natural Language Processing Fundamentals: Build intelligent applications that can interpret the human language to deliver impactful results*. Packt Publishing Ltd. Disponível em: <<https://www.perlego.com/book/955577/natural-language-processing-fundamentals-build-intelligent-applications-that-can-interpret-the-human-language-to-deliver-impactful-results-pdf>>. Acesso em: 2 abril 2022.

Golbeck, J., Mauriello, M., Auxier, B., Bhanushali, K. H., Bonk, C., Bouzaghrane, M. A., Visnansky, G. (2018). Fake news vs satire: A dataset and analysis. In *Proceedings of the 10th ACM Con on Web Science* (pp. 17–21). New York, NY, USA: ACM. Disponível em: <<https://www.eecis.udel.edu/~mlm/docs/2018-Golbeck-WebSci-FakeNewsVsSatire.pdf>>. Acesso em: 2 abril 2022.

Jeffrey Huang (2020). *Detecting Fake News With Machine Learning*. Disponível em: <<https://iopscience.iop.org/article/10.1088/1742-6596/1693/1/012158/meta>> Acesso em: 2 Set 2022.

Kaggle.com (2020). Fake and real news dataset .Disponível em:<<https://www.kaggle.com/datasets/clmentbisailon/fake-and-real-news-dataset?select=True.csv> > Acesso em: 10 Set 2021.

Lane, Hobson. Howard, Cole. Hapke, Hannes. (2019). Natural Language Processing in Action: Understanding, analyzing, and generatint text with Python. Manning. Disponível em:<<https://pt.scribd.com/book/511817149/Natural-Language-Processing-in-Action-Understanding-analyzing-and-generating-text-with-Python> > Acesso em: 4 abril 2022.

Laura D. de Almeida, Victor Fuzaro, Falmer V. Nieto, André L. M. Santana.(2020) Identificação de “Fake News” no contexto político brasileiro: uma abordagem computacional. Disponível em:<<https://sol.sbc.org.br/index.php/wics/article/view/15966/15807>>. Acesso em: 2 abril 2022.

Oliveira, Érick Ritir (2019).Reconhecimento de Bots no Twitter: uma abordagem utilizando aprendizagem de máquina. Disponível em:<<https://ud10.arapiraca.ufal.br/repositorio/publicacoes/3082>> Acesso em: 4 abril 2022.

Sharma, K., Qian, F., Jiang, H., Ruchansky, N., Zhang, M., & Liu, Y. (2019). Combating fake news: A survey on identification and mitigation techniques. ACM Trans. Intell. Syst. Technol., 10(3), 21:1–21:42. Disponível em: < <https://arxiv.org/pdf/1901.06437.pdf> >.Acesso em: 2 abril 2022.

Sharma, K., Seo, S., Meng, C., Rambhatla, S., Dua, A., & Liu, Y. (2020). Coronavirus on social media: Analyzing misinformation in Twitter conversations. arXiv preprint arXiv:2003.12309. Disponível em: < <https://arxiv.org/pdf/2003.12309.pdf> >.Acesso em: 2 abril 2022.

Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media: A data mining perspective. ACM SIGKDD explorations newsletter, 19(1), 22-36 Disponível em: < [https://www.researchgate.net/publication/318981549\\_Fake\\_News\\_Detection\\_on\\_Social\\_Media\\_A\\_Data\\_Mining\\_Perspective](https://www.researchgate.net/publication/318981549_Fake_News_Detection_on_Social_Media_A_Data_Mining_Perspective) >.Acesso em: 2 abril 2022.

Zhou, X., & Zafarani, R. (2020). Fake news: A survey of research, detection methods, and opportunities. ACM Computing Surveys, 53(5), 1–40. Disponível em: <<https://arxiv.org/pdf/1812.00315.pdf>>. Acesso em: 2 nov 2022.

Z Khanam, B N Alwasel, H Sirafi and M Rashid (2021). Fake News Detection Using Machine Learning Approaches. Disponível em: <<https://iopscience.iop.org/article/10.1088/1757-899X/1099/1/012040/meta>>. Acesso em: 2 nov 2022.