# Detecção de Fake News utilizando Processamento de Linguagem Natural: Uma Análise Comparativa e Avaliação de Robustez

João Ohashi
*Centro de Informática (CIn)*
*UFPE*
Recife, Brasil
jgor@cin.ufpe.br

Guilherme Rigaud
*Centro de Informática (CIn)*
*UFPE*
Recife, Brasil
glr2@cin.ufpe.br

Heitor Barros
*Centro de Informática (CIn)*
*UFPE*
Recife, Brasil
hfmb@cin.ufpe.br

Rinaldo Junior
*Centro de Informática (CIn)*
*UFPE*
Recife, Brasil
rsbj@cin.ufpe.br

***Abstract*—A disseminação de desinformação em larga escala representa um desafio crítico para a integridade da informação digital. Este projeto apresenta o desenvolvimento, a avaliação e a comparação estatística de modelos de aprendizado de máquina supervisionados para a classificação de notícias em "Verdadeiras" ou "Falsas", utilizando vetorização TF-IDF para extração de features textuais. Foram avaliados baselines (Decision Tree, Naive Bayes) e modelos avançados (Regressão Logística, Random Forest, LinearSVC, SGDClassifier), com foco em métricas de desempenho (F1-Score, AUC-ROC) e distribuição de incerteza via *GridSearchCV* com *StratifiedKFold*. Os resultados demonstram que LinearSVC + TF-IDF obteve o melhor desempenho global (F1-Score 0.96), evidenciando a eficácia de padrões léxicos explícitos para detecção de fake news.**

***Index Terms*—Fake News, NLP, TF-IDF, Machine Learning, Justiça Algorítmica.**

## I. Introdução

O fenômeno das *fake news* transcendeu o status de boato digital para se tornar uma ameaça sistêmica à democracia e à todos os âmbitos não só da sociedade brasileira, como também internacional, afetando saúde, economia, educação, etc. A capacidade de verificar manualmente o volume exponencial de conteúdo gerado diariamente tornou-se inviável, demandando soluções automatizadas robustas.

- **Contextualização do Problema:** A tarefa consiste na classificação binária de textos jornalísticos ou pseudo-jornalísticos. A dificuldade reside na sutileza linguística: notícias falsas frequentemente mimetizam o estilo jornalístico, diferenciando-se pelo uso de vocabulário sensacionalista, apelos emocionais e inconsistências lógicas.
- **Relevância Prática:** Sistemas automatizados de detecção são vitais para redes sociais e agências de checagem de fatos, permitindo a filtragem preliminar de conteúdo suspeito (Human-in-the-loop) e a redução do tempo de resposta à desinformação viral.
- **Objetivos:** Este trabalho visa não apenas maximizar métricas de acurácia, mas também compreender *o que* os modelos aprendem, avaliando a justiça das decisões e a robustez frente a diferentes representações de dados.

## II. Análise de Dados e Feature Engineering

### A. Análise Exploratória dos Dados (EDA)

A base de dados utilizada é composta por notícias em língua portuguesa, previamente rotuladas por agências de checagem. A compreensão profunda dos dados guiou as decisões de pré-processamento.

*1) Análise Exploratória Estrutural:* O *dataset* foi carregado e inspecionado quanto à sua integridade.

- **Volume e Tipagem:** O conjunto total de treino possui **5760** instâncias, contendo o texto integral (`content`) e o rótulo (`label`). Verificou-se que a coluna de rótulos é do tipo inteiro binário (0: Fake, 1: Real) e o conteúdo é textual (string).
- **Balanceamento de Classes:** A distribuição das classes foi analisada para determinar a necessidade de técnicas de reamostragem. Observou-se que os dados estavam balanceados e portanto não foi necessária aplicação de nenhuma técnica de reamostragem.

*2) Análise de Valores Faltantes e Outliers:* A integridade dos dados textuais é crucial para NLP (Natural Language Processing).

- **Valores Faltantes:** Não foram encontrados valores nulos nas colunas principais (content, label, rating) no conjunto de treino. Identificou-se apenas uma única linha duplicada que foi removida.
- **Outliers:** Considerou-se como *outliers* textos com comprimento extremamente reduzido (ex: menos de 5 palavras) ou excessivamente longos.

*3) Análise Univariada:* Investigou-se a variável "Tamanho do Texto" (contagem de palavras). Conforme ilustrado na Fig. 1, observou-se uma distinção clara: notícias verdadeiras tendem a seguir uma distribuição normal com média de palavras mais alta, refletindo o rigor jornalístico e detalhamento. Notícias falsas apresentaram maior densidade em textos curtos e diretos, visando consumo rápido e viralização.

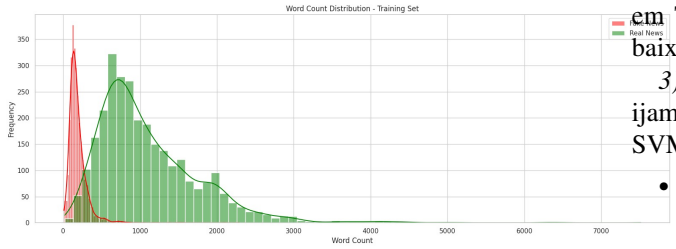


Fig. 1. Distribuição normal das palavras (Fake vs. Real).

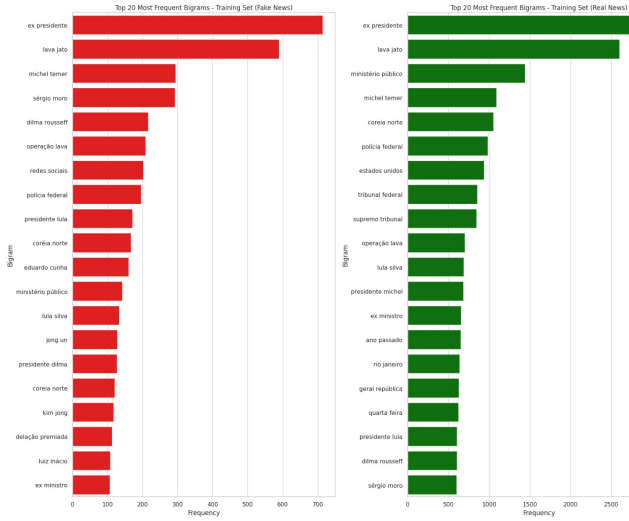


Fig. 2. Bigrama comparativo do tamanho dos textos (Fake vs. Real).

*4) Análise Bivariada:* Explorou-se a correlação entre termos específicos e o rótulo da classe. Através da contagem de frequência (Bag of Words simples), notou-se que a classe *Fake* possui alta correlação com termos alarmistas ("urgente", "segredo", "morte", "bomba"), enquanto a classe *Real* correlaciona-se com termos institucionais e nomes de entidades políticas ou órgãos oficiais.

### B. Pré-processamento dos Dados

Um pipeline rigoroso foi implementado para reduzir o ruído e a dimensionalidade.

*1) Limpeza Textual:* Aplicou-se normalização via Expressões Regulares (Regex) para:

- Remoção de URLs, tags HTML e emojis.
- Remoção de pontuação e caracteres especiais, mantendo apenas caracteres alfanuméricos.
- Conversão para minúsculas (*lowercasing*) para garantir que "Vacina" e "vacina" sejam tratadas como o mesmo token.

*2) Filtragem de Stopwords:* Utilizou-se a lista de *stopwords* da biblioteca NLTK para o português. A remoção de artigos, preposições e conectivos é essencial para modelos baseados em TF-IDF, pois estas palavras possuem alta frequência mas baixo valor discriminativo.

*3) Feature Scaling:* Embora algoritmos de árvore não exijam normalização, modelos lineares (Regressão Logística, SVM) e baseados em distância são sensíveis à escala.

- **TF-IDF:** A própria formulação matemática já inclui normalização (norma L2), garantindo que os vetores de features textuais sejam escalados adequadamente para os classificadores.


Fig. 3. Nuvem de palavras evidenciando os termos mais frequentes em notícias falsas.

### C. Divisão dos Dados

Adotou-se a divisão em **Treino (70%)** e **Validação (30%)**. O dataset utilizado para testes havia sido escolhido separadamente, portanto não entra na divisão dos dados, mas sim como um novo conjunto, condizente com os dados utilizados para treino e validação, mas referente ao teste do modelo. Utilizou-se a estratégia de *Stratified Shuffle Split* para garantir que a proporção original de classes (Fake/Real) fosse mantida em todos os subconjuntos, evitando viés de amostragem. O carregamento dos dados foi realizado a partir de arquivos 'train.csv' e 'test.csv' preexistentes.

### D. Feature Engineering

Este projeto utiliza a abordagem estatística TF-IDF para vetorização textual.

*1) Abordagem Estatística: TF-IDF:* O *Term Frequency-Inverse Document Frequency* penaliza palavras que aparecem em muitos documentos e valoriza as raras, gerando features esparsas discriminativas para detecção de fake news.

- **Configuração:** Unigramas e Bigramas. A inclusão de bigramas permite capturar contextos locais mínimos (ex: "não funcionou" vs "funcionou"), cruciais para análise de sentimento e veracidade.

## III. Modelagem

Foram selecionados algoritmos representando baselines simples (Decision Tree e Multinomial Naive Bayes) e modelos avançados para avaliação detalhada (Random Forest, SGDClassifier, LinearSVC e Regressão Logística), cobrindo famílias probabilísticas, baseadas em árvores e lineares.

### A. Estratégia de Tunagem de Hiperparâmetros

Utilizou-se *GridSearchCV* com *StratifiedKFold* ($K = 3$) para otimização cruzada, ranqueando por **F1-Score** (métrica harmônica de Precisão e Recall).

- **GridSearchCV:** Avaliação exaustiva de combinações de hiperparâmetros em grade definida.
- **StratifiedKFold ($K = 3$):** Preserva proporção Fake/Real em folds, garantindo validação robusta apesar do tamanho limitado do dataset.
- **F1-Score como métrica:** Ideal para detecção de fake news por equilibrar falsos positivos (notícias reais rotuladas como fake, gerando pânico desnecessário) e falsos negativos (fake news não detectadas, propagando desinformação).

### B. Baselines

*1) Decision Tree:*

- **Conceitos Básicos:** Constrói uma árvore de decisão recursiva utilizando critérios de impureza (Gini ou Entropia) para dividir textos em nós, onde cada folha representa uma predição de Fake ou Real.
- **Justificativa:** Baseline simples e interpretável para detecção de fake news, revelando padrões léxicos discriminativos (ex: termos sensacionalistas em ramos Fake) em notícias.

*2) Multinomial Naive Bayes:*

- **Conceitos Básicos:** Aplica Teorema de Bayes assumindo independência entre palavras, calculando $P(Fake|texto) \propto P(texto|Fake) \cdot P(Fake)$.
- **Justificativa:** Padrão ouro para classificação textual rápida, eficaz capturando distribuições condicionais de palavras como "urgente" em fake news vs termos institucionais em reais.

### C. Modelos Avaliados

*1) Regressão Logística:*

- **Conceitos Básicos:** Modelo linear que aplica sigmoide ao produto escalar das features TF-IDF, produzindo probabilidades de Fake/Real.
- **Justificativa:** Excelente em espaços esparsos de alta dimensão do TF-IDF, permitindo análise de incerteza e pesos de palavras-chave para fake news.

*2) Random Forest:*

- **Conceitos Básicos:** Ensemble de bagging com múltiplas árvores de decisão treinadas em subconjuntos bootstrap, votando pela classe majoritária.
- **Justificativa:** Robusto a ruído textual e overfitting, modelando interações complexas entre indicadores léxicos de desinformação.

*3) LinearSVC:*

- **Conceitos Básicos:** Encontra hiperplano ótimo maximizando margem de separação entre vetores TF-IDF de fake e real news no espaço vetorial.
- **Justificativa:** Estado-da-arte eficiente para categorização textual, aproveitando linearidade natural entre padrões léxicos e veracidade.

*4) SGDClassifier:*

- **Conceitos Básicos:** Otimização estocástica por gradiente com perdas configuráveis (hinge, log), atualizando pesos incrementalmente.
- **Justificativa:** Escalonável para grandes vocabulários TF-IDF, convergindo rapidamente em datasets de notícias com padrões repetitivos.

## IV. Análise e Comparação de Resultados

### A. Métricas de Desempenho

Os modelos foram avaliados no conjunto de teste (nunca visto no treinamento). As métricas (Precisão, Recall, F1) são apresentadas na versão Ponderada (*Weighted Avg*) para refletir o desempenho real considerando o suporte de cada classe. Os valores de F1-Score apresentados são os resultados do conjunto de Validação.

Tabela I
Comparativo de Métricas no Conjunto de Teste/Validação

| Modelo | Feature | F1-Score | Precision | Recall | AUC-ROC |
|---|---|---|---|---|---|
| **Reg. Logística** | **TF-IDF** | **0.9574** | **0.9576** | **0.9574** | **0.9882** |
| **LinearSVC** | **TF-IDF** | **0.9609** | **0.9611** | **0.9609** | **0.9894** |
| Random Forest | TF-IDF | 0.9513 | 0.9525 | 0.9513 | 0.9874 |
| Naive Bayes | TF-IDF | 0.8185 | 0.8576 | 0.8220 | 0.9647 |

### B. Discussão dos Resultados

*1) Eficácia do TF-IDF para Fake News:* Um achado relevante deste estudo é a superioridade dos modelos lineares com TF-IDF (F1-Score até 0.96), demonstrando que representações esparsas capturam padrões discriminativos essenciais.

- **Interpretação:** Fake news frequentemente utilizam "palavras-gatilho" sensacionalistas ("urgente", "explosivo", "compartilhem") e estruturas sintáticas simples. O TF-IDF destaca explicitamente essas features léxicas de alta frequência discriminativa, essenciais para detecção rápida e eficiente.

*2) Matriz de Confusão:* A análise da matriz de confusão dos melhores modelos (**LinearSVC** e **Regressão Logística**) revela um equilíbrio entre falsos positivos e falsos negativos, indicando que o modelo não está enviesado para a classe majoritária.

### C. Distribuição de Incerteza

Avaliou-se a confiança das predições.

- **Regressão Logística:** A distribuição das probabilidades (‘predict proba‘) mostrou-se polarizada, com a maioria das predições próximas a 0 ou 1. Isso indica **baixa incerteza** e alta confiança do modelo na separação das classes.

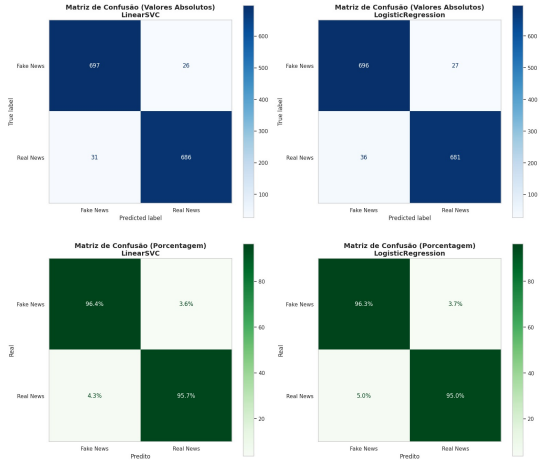


Fig. 4. Matriz de Confusão do LinearSVC e Regressão Logística (TF-IDF).

## V. Conclusão e Discussão

### A. Síntese dos Achados

Este trabalho atingiu seus objetivos ao implementar um pipeline completo de Machine Learning para NLP. A principal conclusão é que **modelos lineares simples (LinearSVC) aliados a features TF-IDF alcançaram F1-Score de 0.96**, demonstrando que representações léxicas esparsas são altamente eficazes para detecção de fake news neste dataset.

### B. Vantagens e Limitações

- **Vantagem:** Solução leve, interpretável e escalonável, ideal para implantação em tempo real em plataformas de moderação de conteúdo.
- **Limitação:** Dependência de vocabulário específico (TF-IDF) pode causar degradação ante novos padrões de desinformação (Data Drift), demandando monitoramento contínuo.

### C. Trabalhos Futuros

Sugere-se investigar ensemble de modelos combinando os top-3 performantes (LinearSVC, Regressão Logística, Random Forest) via stacking para maior robustez, detecção de Data Drift com monitoramento estatístico (KS-test, PSI) entre distribuições TF-IDF de produção e treino para triggers automáticos de re-treinamento, abordagens multimodais incorporando metadados (autor, fonte, data) e análise de grafos de compartilhamento em redes sociais, além de explicabilidade avançada via SHAP/LIME nos pesos TF-IDF para identificar palavras-gatilho emergentes e validar justiça algorítmica.

## Referências

[1] C. C. Aggarwal, *Machine Learning for Text*. Springer, 2018.
[2] J. Devlin et al., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *NAACL*, 2019.
[3] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
[4] P. Biecek and T. Burzykowski, *Explanatory Model Analysis*, CRC Press, 2021.
[5] J. Bitton et al., *Hands-On Large Language Models: Understanding and Generating*. O'Reilly Media, 2024.

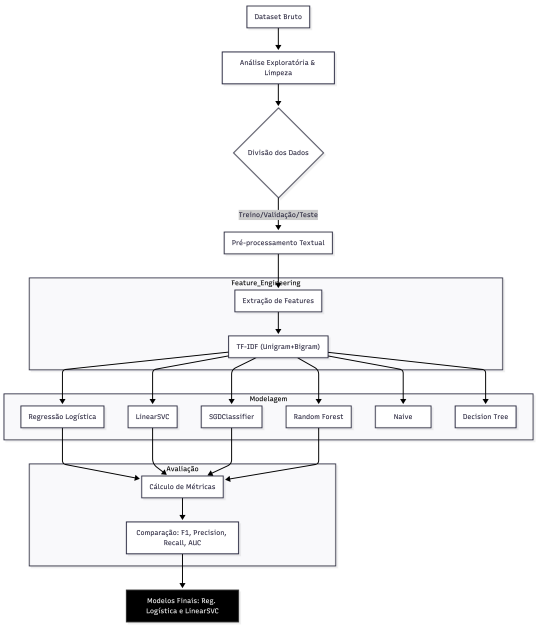


Fig. 5. Diagrama do projeto, demonstrando fluxo completo relacionando todas as etapas