

# Categorização de Textos para Análise de Sentimentos: Uma Abordagem Utilizando Aprendizado de Máquina

Vincent V. S. V. B. Tomio<sup>1</sup>, Yan O. da Costa<sup>2</sup>

<sup>1</sup>Departamento de Informática – Universidade Federal do Paraná (UFPR)  
Curitiba – PR – Brazil

{vvsbt20,yoc20}@inf.ufpr.br

**Abstract.** *Sentiment analysis is a widely used technique for extracting subjective information from texts, such as opinions and emotions. This work presents a machine learning application for text categorization with a focus on sentiment analysis, using a dataset of drug reviews. The methodology includes text pre-processing with TF-IDF, training of classification models (Naive Bayes, Random Forest, Logistic Regression, and SVM), and evaluation based on metrics such as accuracy, F1-Score, and ROC-AUC curve. The results show that Random Forest achieved the best overall performance, with 90% accuracy, while Naive Bayes stood out for its speed and simplicity. This study reinforces the importance of machine learning in practical text analysis applications and offers insights into the performance of different algorithms.*

**Resumo.** *A análise de sentimentos é uma técnica amplamente utilizada para extrair informações subjetivas de textos, como opiniões e emoções. Este trabalho apresenta uma aplicação de aprendizado de máquina para categorização de textos com foco na análise de sentimentos, utilizando um dataset de avaliações de medicamentos. A metodologia inclui pré-processamento textual com TF-IDF, treinamento de modelos de classificação (Naive Bayes, Random Forest, Logistic Regression e SVM) e avaliação baseada em métricas como acurácia, F1-Score e curva ROC-AUC. Os resultados mostram que o Random Forest alcançou a melhor performance geral, com acurácia de 90%, enquanto o Naive Bayes destacou-se pela rapidez e simplicidade. Este estudo reforça a importância do aprendizado de máquina em aplicações práticas de análise textual e oferece insights sobre o desempenho de diferentes algoritmos.*

## 1. Introdução

A análise de sentimentos é uma subárea da mineração de textos que busca identificar e extrair opiniões subjetivas, emoções ou avaliações a partir de dados textuais. Este tipo de análise tem se tornado crucial em diversos setores, como saúde, marketing e tecnologia, permitindo entender percepções e melhorar tomadas de decisão.

Uma abordagem popular para a análise de sentimentos é a categorização de textos, que utiliza algoritmos de aprendizado de máquina para classificar textos em categorias predefinidas, como "positivo" ou "negativo". No entanto, a escolha adequada de técnicas de pré-processamento e algoritmos pode impactar significativamente o desempenho da análise.

Neste trabalho, exploramos a aplicação de aprendizado de máquina para análise de sentimentos em um conjunto de dados composto por avaliações de medicamentos. Utilizamos o TF-IDF para transformar os textos em representações numéricas e treinamos diferentes algoritmos de classificação, incluindo Naive Bayes, Random Forest, Logistic Regression e SVM. O desempenho dos modelos foi avaliado com base em métricas como acurácia, F1-Score e curva ROC-AUC.

Os resultados obtidos destacam a eficiência do Random Forest em termos de acurácia, enquanto o Naive Bayes demonstrou vantagens em termos de simplicidade e tempo de execução. Este estudo reforça a aplicabilidade do aprendizado de máquina em tarefas de categorização de textos e oferece uma comparação prática entre modelos.

## 2. Metodologia

Neste trabalho, foi desenvolvido um pipeline de aprendizado de máquina para categorização de textos com aplicação em análise de sentimentos. A metodologia pode ser dividida nas seguintes etapas:

### 2.1. Dataset

O conjunto de dados utilizado neste trabalho é o *Drug Reviews Dataset*, disponibilizado pela Universidade da Califórnia, Irvine (UCI). Este dataset contém avaliações de pacientes sobre medicamentos específicos, suas respectivas condições de saúde, e uma classificação de satisfação em uma escala de 10 estrelas.

#### Características principais do dataset:

- **Tarefas associadas:** Classificação, Regressão, Agrupamento.
- **Tipo de dados:** Multivariado, com componentes textuais e numéricos.
- **Áreas:** Saúde e Medicina.
- **Número de instâncias:** 215.063 avaliações.
- **Número de atributos:** 6 atributos.

#### Descrição dos atributos:

1. **drugName (categorical):** Nome do medicamento avaliado.
2. **condition (categorical):** Condição associada ao uso do medicamento.
3. **review (text):** Texto da avaliação fornecida pelo paciente.
4. **rating (numerical):** Nota de satisfação do paciente em uma escala de 1 a 10.
5. **date (date):** Data de submissão da avaliação.
6. **usefulCount (numerical):** Número de usuários que consideraram a avaliação útil.

O objetivo principal do dataset é possibilitar a análise de sentimentos a partir das experiências dos pacientes. Entre os estudos previstos com os dados, destacam-se:

1. A análise de sentimentos sobre múltiplos aspectos, como eficácia e efeitos colaterais.
2. A transferência de modelos entre domínios diferentes, como condições de saúde variadas.
3. A transferência de modelos entre diferentes fontes de dados, como outros repositórios farmacêuticos.

Os dados foram obtidos por meio de *web scraping* de sites de revisões farmacêuticas e são particionados em 75% para treinamento e 25% para teste, conforme descrito na publicação original.

**Notas importantes:** O uso deste dataset está condicionado às seguintes restrições:

- Utilizar os dados exclusivamente para fins de pesquisa.
- Não utilizar os dados para fins comerciais.
- Não distribuir os dados a terceiros.
- Citar a fonte original ao utilizar o dataset.

O código completo da aplicação está disponível no Google Colab por meio do seguinte link: [Google Colab - Aplicação Completa](#).

## 2.2. Pré-processamento dos Dados

O pré-processamento dos dados foi uma etapa essencial para transformar os textos brutos em uma forma adequada para os modelos de aprendizado de máquina. As seguintes etapas foram realizadas:

- **Remoção de valores ausentes:** Linhas com valores ausentes na coluna de condição (*condition*) foram removidas para garantir a consistência dos dados.
- **Limpeza de texto:**
  - Conversão de todo o texto para letras minúsculas.
  - Remoção de caracteres especiais, números e múltiplos espaços utilizando expressões regulares.
  - Utilização do *BeautifulSoup* para eliminar resíduos HTML presentes nos textos.
- **Remoção de palavras irrelevantes (*stopwords*):** As palavras irrelevantes foram removidas com a ajuda da biblioteca NLTK.
- **Stemming:** As palavras foram reduzidas às suas raízes com o *Snowball Stemmer*, otimizando a representação dos dados.
- **Representação numérica:** Os textos pré-processados foram convertidos para uma representação numérica utilizando o *TF-IDF Vectorizer*. Este processo:
  - Removeu palavras extremamente frequentes (*stopwords do TF-IDF*).
  - Limitou o vocabulário a 5000 características mais relevantes, para reduzir a dimensionalidade.
- **Divisão dos dados:** Os dados foram divididos em 70% para o conjunto de treinamento e 30% para o conjunto de teste.

## 2.3. Modelos de Classificação

Foram utilizados quatro algoritmos de aprendizado de máquina para realizar a categorização dos textos. Abaixo estão descritos os modelos e suas características principais:

1. **Naive Bayes Multinomial:** Modelo probabilístico eficiente para textos, ideal para tarefas de classificação simples.
2. **Random Forest:** Um *ensemble* que combina múltiplas árvores de decisão, reduzindo *overfitting*.
3. **Logistic Regression:** Modelo linear usado para classificação binária, interpretável e eficaz.
4. **SVM (Support Vector Machine):** Algoritmo que busca margens ótimas para separar classes, utilizando kernel linear.

## 2.4. Ferramentas e Bibliotecas

A implementação foi realizada em Python com as seguintes bibliotecas:

- **Scikit-Learn:** Para modelos e métricas de aprendizado de máquina.
- **NLTK:** Para pré-processamento textual, como *stopwords* e *stemming*.
- **Pandas:** Para manipulação de dados.
- **Seaborn/Matplotlib:** Para visualizações gráficas.
- **BeautifulSoup:** Para limpeza de HTML.

## 3. Análise Exploratória de Dados (EDA)

Nesta seção, apresentamos a análise exploratória do conjunto de dados, que permitiu compreender as características principais e identificar padrões relevantes para a tarefa de categorização de textos.

### 3.1. Dimensões e Tipos de Dados

O dataset possui um total de 215.063 instâncias e 6 atributos:

- **drugName (categorical):** Nome do medicamento avaliado.
- **condition (categorical):** Condição associada ao uso do medicamento.
- **review (text):** Texto da avaliação fornecida pelo paciente.
- **rating (numerical):** Nota de satisfação do paciente (1 a 10).
- **date (date):** Data de submissão da avaliação.
- **usefulCount (numerical):** Número de usuários que consideraram a avaliação útil.

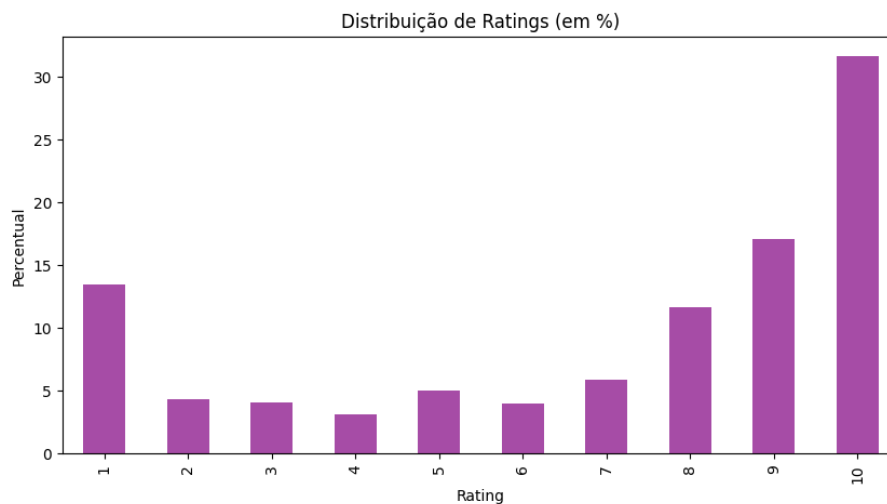
Os dados possuem alguns valores ausentes na coluna *condition* (aproximadamente 0,55% das entradas).

### 3.2. Distribuição de Classes e Ratings

A distribuição dos *ratings* apresenta os seguintes percentuais:

- Nota 10: 31,6%.
- Nota 9: 17,1%.
- Nota 1: 13,4%.

A maioria das avaliações concentra-se em notas extremas (1 e 10), sugerindo uma polarização das opiniões.



**Figure 1. Distribuição dos ratings no dataset.**

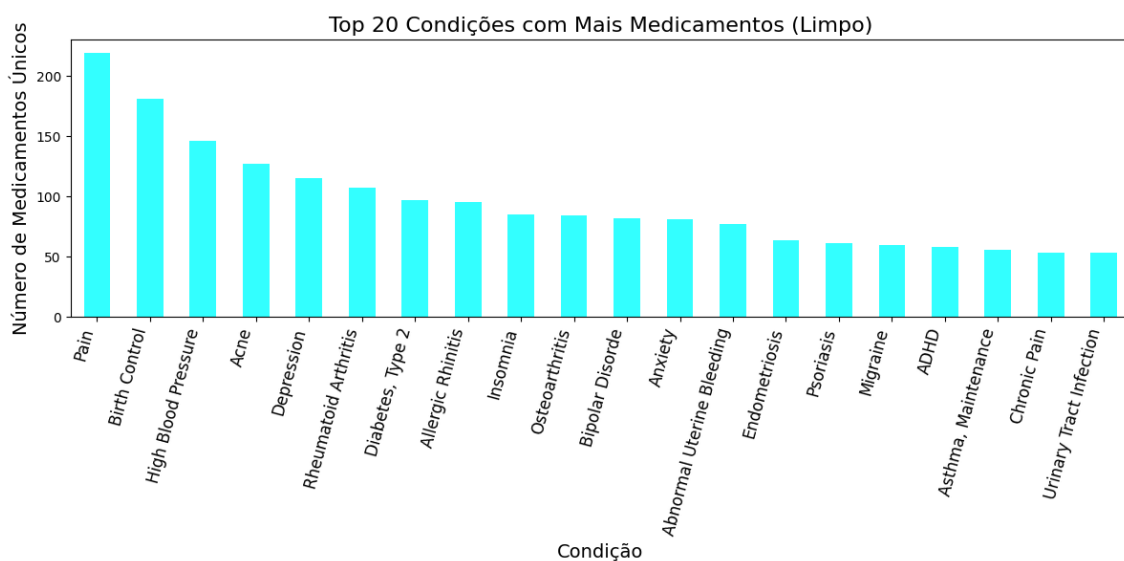
### 3.3. Análise de Medicamentos e Condições

O dataset inclui 3.671 medicamentos únicos e 916 condições de saúde distintas. As condições mais avaliadas são:

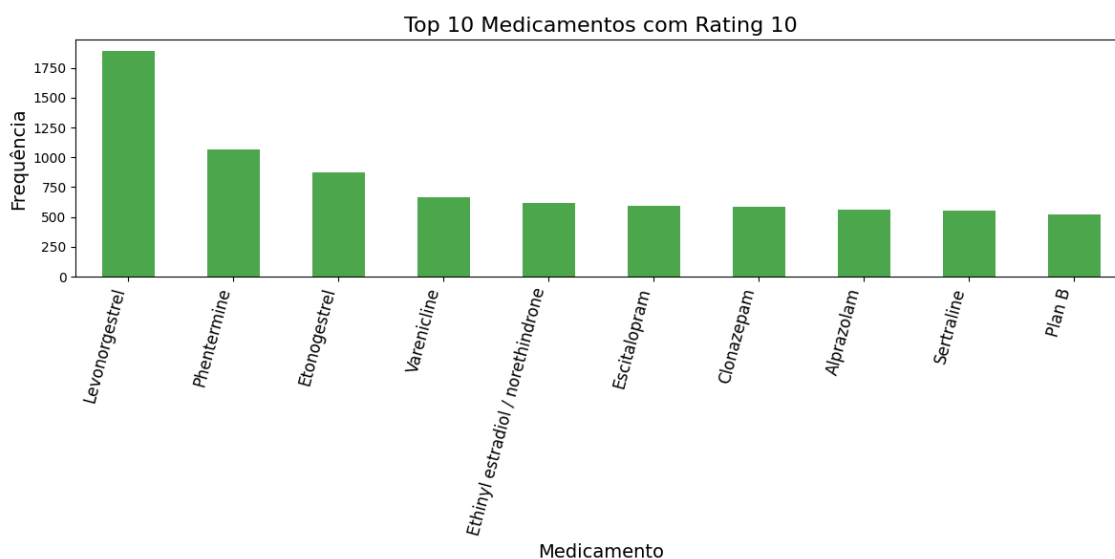
- *Birth Control*: 38.436 avaliações (aproximadamente 18% do total).
- *Depression*: 12.164 avaliações.
- *Pain*: 8.245 avaliações.

Os medicamentos com maior número de avaliações incluem:

- *Levonorgestrel*: 4.930 avaliações.
- *Etonogestrel*: 4.421 avaliações.
- *Ethinyl estradiol / norethindrone*: 3.753 avaliações.



**Figure 2. Top 20 condições mais frequentes.**



**Figure 3. Top 10 medicamentos com rating 10.**

### 3.4. Análise Textual

A análise dos textos (*reviews*) revelou:

- Comprimentos variáveis: A maioria das avaliações possui entre 50 e 150 palavras.
- Presença de ruído: Alguns textos contêm caracteres indesejados, como "</span>" e "Not Listed / Other".

Após a remoção de ruído, as dimensões do dataset foram reduzidas para 213.300 instâncias.

### 3.5. Insights Adicionais

A análise também revelou:

- **Condições associadas ao Levonorgestrel:**
  - *Birth Control*: 2.884 avaliações.
  - *Emergency Contraception*: 1.651 avaliações.
- **Distribuição temporal:** Cerca de 60% das avaliações foram registradas entre 2015 e 2017.
- **Distribuição de *usefulCount*:** A maioria das contagens de utilidade está entre 0 e 200, mas há valores extremos que podem ser tratados como outliers.

## 4. Pré-Processamento de Dados

O pré-processamento de dados é uma etapa crucial no desenvolvimento de modelos de aprendizado de máquina. Um dataset mal estruturado, com valores ausentes ou informações irrelevantes, pode comprometer a qualidade dos modelos gerados. Esta etapa transforma os dados brutos em um formato limpo e padronizado, adequado para análises e modelagem preditiva. Como diz o ditado: "*Garbage In, Garbage Out*".

### 4.1. Etapas Realizadas

Para preparar os dados, realizamos as seguintes etapas:

#### 4.1.1. Remoção de Valores Ausentes

A coluna *condition* apresentava valores ausentes em cerca de 0,55% das entradas (1.194 linhas). Essas instâncias foram removidas para garantir a consistência dos dados. Após a limpeza, não havia mais valores ausentes no dataset.

#### 4.1.2. Transformação Textual

O texto das avaliações dos pacientes (*review*) foi processado para remover ruídos e padronizar a estrutura. As seguintes operações foram realizadas:

- **Remoção de HTML:** Tags HTML foram eliminadas utilizando a biblioteca *BeautifulSoup*.
- **Remoção de caracteres especiais e números:** Apenas letras foram mantidas, para focar no conteúdo textual.
- **Conversão para letras minúsculas:** Todo o texto foi convertido para minúsculas para padronização.
- **Remoção de stopwords:** Palavras comuns irrelevantes, como "*the*", "*and*" e "*of*", foram removidas com o auxílio do *Natural Language Toolkit* (NLTK).
- **Stemming:** As palavras foram reduzidas às suas formas raiz utilizando o *Snowball Stemmer* (e.g., "*running*" → "*run*").

#### 4.1.3. Criação da Variável-Alvo (*sentiment*)

Para realizar a classificação binária de sentimentos, criamos uma nova variável chamada *sentiment*, baseada nas notas (*rating*):

- **Sentimento Positivo (1):**  $rating > 5$ .
- **Sentimento Negativo (0):**  $rating \leq 5$ .

A distribuição final da variável-alvo foi:

- **Sentimento Positivo:** 70,10%.
- **Sentimento Negativo:** 29,90%.

### 4.2. Prós e Contras da Abordagem

#### 4.2.1. Pontos Positivos

- **Limpeza e padronização:** O texto foi cuidadosamente preparado, reduzindo ruídos e inconsistências.
- **Simplicidade:** A classificação binária baseada na coluna *rating* simplifica o problema.

#### 4.2.2. Pontos Negativos

- **Perda de nuances:** Avaliações intermediárias (*ratings* de 4 a 6) podem conter informações valiosas que foram ignoradas.

- **Impacto no significado:** Técnicas como *stemming* e remoção de *stopwords* podem afetar negativamente a semântica.
- **Informação descartada:** A análise focou exclusivamente no texto (*review*), ignorando atributos como *usefulCount*.

### 4.3. Resultados do Pré-Processamento

Após aplicar as técnicas de pré-processamento, o dataset foi salvo no arquivo *drug\_reviews\_preprocessed.csv*, pronto para as próximas etapas de modelagem. Abaixo, apresentamos exemplos de textos processados:

## 5. Construção do Modelo

A construção do modelo foi realizada com o objetivo de prever o sentimento (*sentiment*) associado às avaliações dos pacientes (*review*). Para isso, utilizamos técnicas de aprendizado de máquina combinadas com um processo de vetorização textual. Esta seção descreve detalhadamente as escolhas feitas, os métodos utilizados e os resultados obtidos.

### 5.1. Raciocínio e Justificativa

O aprendizado de máquina foi escolhido como abordagem devido à sua capacidade de lidar com grandes volumes de dados textuais e identificar padrões complexos que seriam difíceis de modelar manualmente. A análise de sentimentos, por sua natureza, exige:

- Representações numéricas consistentes dos textos (*TF-IDF Vectorizer*).
- Modelos robustos e interpretáveis para classificação.

Com base na literatura e no comportamento esperado do dataset, quatro algoritmos foram escolhidos, representando abordagens probabilísticas, baseadas em árvores, lineares e de margens.

### 5.2. Técnicas Utilizadas

#### 5.2.1. Transformação Textual

Os textos foram convertidos em representações numéricas utilizando o **TF-IDF Vectorizer** (*Term Frequency-Inverse Document Frequency*). Esta técnica combina:

- **TF (Frequência do Termo):** Mede a importância de uma palavra dentro de um documento.
- **IDF (Frequência Inversa do Documento):** Reduz a importância de palavras comuns entre documentos.

*Expectativa:* O TF-IDF Vectorizer é amplamente utilizado em problemas de texto e espera-se que ele forneça uma boa base para os modelos, destacando palavras mais relevantes.

#### 5.2.2. Modelos de Aprendizado de Máquina

Quatro modelos foram treinados:



1. **Naive Bayes Multinomial:** Um modelo probabilístico baseado na aplicação da *Regra de Bayes*, assumindo independência condicional entre palavras. É eficiente para tarefas textuais devido à simplicidade de seu cálculo. *Expectativa:* Bom desempenho inicial, mas com possíveis limitações em casos de relações complexas entre palavras.
2. **Random Forest:** Um *ensemble* de árvores de decisão, onde múltiplas árvores votam para gerar a previsão final. Este método reduz o *overfitting*, comum em árvores individuais. *Expectativa:* Alta acurácia devido à robustez contra variabilidades nos dados.
3. **Logistic Regression:** Um modelo linear que utiliza a função logística para prever probabilidades de classes binárias. É simples, interpretável e eficiente. *Expectativa:* Desempenho sólido em dados bem estruturados.
4. **SVM (Support Vector Machine):** Um algoritmo que busca margens ótimas para separar as classes. O kernel linear foi escolhido devido à alta dimensionalidade do texto vetorizado. *Expectativa:* Boa performance em problemas lineares, mas sensível a desbalanceamento de dados.

### 5.3. Divisão dos Dados e Métricas de Avaliação

Os dados foram divididos em:

- **Treinamento:** 70% (148.468 instâncias).
- **Teste:** 30% (63.630 instâncias).

Para avaliar os modelos, utilizamos:

- **Acurácia:** Percentual de previsões corretas.
- **F1-Score:** Média harmônica entre precisão e recall, útil para classes desbalanceadas.
- **Curvas ROC e AUC:** Avaliam a capacidade de discriminação entre classes.

### 5.4. Resultados Esperados e Obtidos

#### 5.4.1. Naive Bayes Multinomial

*Resultados:* Acurácia de 77,8% e AUC de 0,84. Embora rápido e eficiente, apresentou dificuldades com relações mais complexas entre palavras.

#### 5.4.2. Random Forest

*Resultados:* Acurácia de 90,6% e AUC de 0,96. Este modelo superou as expectativas, oferecendo alta performance, mas com maior custo computacional.

#### 5.4.3. Logistic Regression

*Resultados:* Acurácia de 83,1% e AUC de 0,88. Um equilíbrio sólido entre simplicidade e desempenho.

#### 5.4.4. SVM

*Resultados:* Acurácia de 53,9% e AUC de 0,54. A performance ficou abaixo do esperado, indicando que o kernel linear não foi suficiente para capturar as relações nos dados.

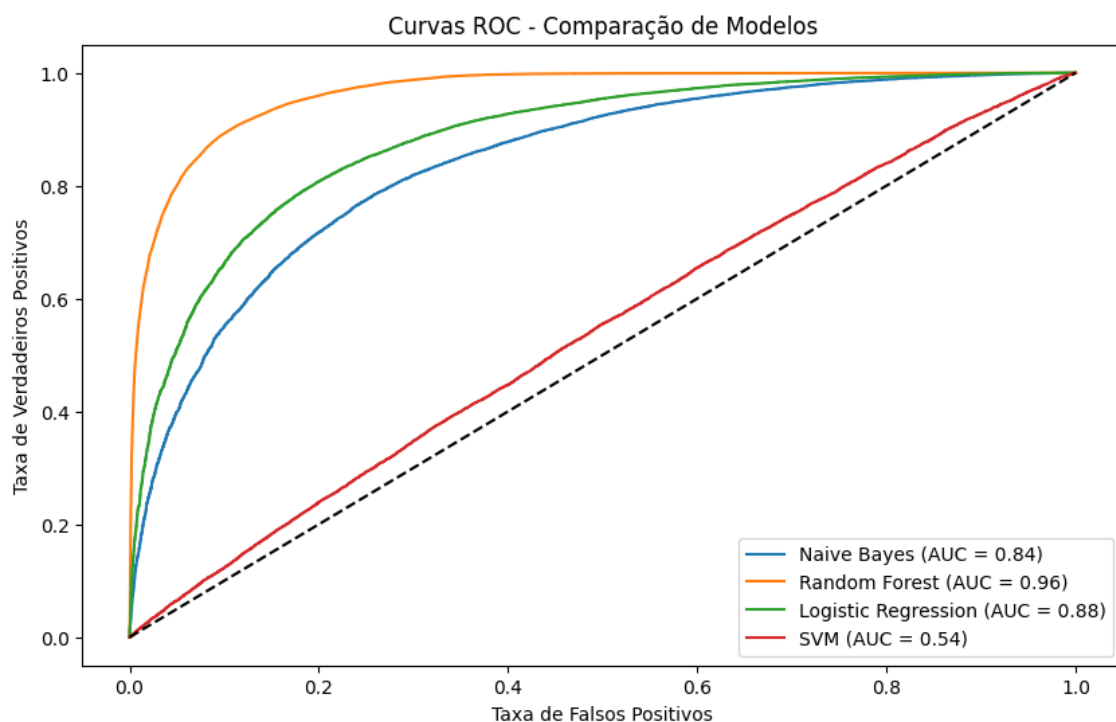
#### 5.5. Validação Cruzada

A validação cruzada foi aplicada para verificar a robustez dos modelos. As acurácias médias foram:

- **Naive Bayes:** 77,4%.
- **Random Forest:** 86,5%.
- **Logistic Regression:** 82,5%.
- **SVM:** 59,6%.

#### 5.6. Curvas ROC e AUC

As curvas ROC (Figura 4) demonstram que o **Random Forest** foi o modelo com maior capacidade de discriminação entre classes, seguido por **Logistic Regression** e **Naive Bayes**.



**Figure 4. Curvas ROC para os modelos treinados.**

### 6. Resultados e Discussão

Nesta seção, discutimos os resultados obtidos na etapa de avaliação dos modelos, com base nas métricas calculadas e nas curvas ROC. Além disso, comparamos os desempenhos dos diferentes algoritmos e refletimos sobre os desafios e limitações encontrados.

#### 6.1. Desempenho dos Modelos

Os modelos apresentaram os seguintes desempenhos em termos de acurácia, F1-Score e AUC:

Modelo	Acurácia	F1-Score	AUC
Naive Bayes	77,8%	0,74	0,84
Random Forest	90,6%	0,90	0,96
Logistic Regression	83,1%	0,83	0,88
SVM	53,9%	0,56	0,54

**Table 1. Resumo do desempenho dos modelos.**

## 6.2. Análise Comparativa

### 6.2.1. Naive Bayes Multinomial

O modelo Naive Bayes apresentou um bom desempenho inicial, com acurácia de 77,8% e AUC de 0,84. Apesar de sua simplicidade, mostrou-se eficiente em lidar com texto. No entanto, sua suposição de independência condicional entre palavras limitou sua capacidade de capturar relações mais complexas.

### 6.2.2. Random Forest

O modelo Random Forest superou as expectativas, com acurácia de 90,6% e AUC de 0,96. Sua capacidade de reduzir *overfitting* e lidar com relações não lineares destacou-se, tornando-o a melhor escolha para este problema. No entanto, o custo computacional foi mais alto em comparação aos demais modelos.

### 6.2.3. Logistic Regression

A Logistic Regression teve um desempenho consistente, com acurácia de 83,1% e AUC de 0,88. Este modelo é uma escolha sólida para problemas binários devido à sua simplicidade e interpretabilidade, mas ficou atrás do Random Forest em termos de capacidade de discriminação.

### 6.2.4. SVM

O SVM teve o pior desempenho, com acurácia de 53,9% e AUC de 0,54. Embora esperado um desempenho razoável com o kernel linear, o modelo mostrou dificuldades em lidar com o desbalanceamento e a dimensionalidade dos dados.

## 6.3. Discussão dos Resultados

- **Desbalanceamento de classes:** A predominância de sentimentos positivos (70%) pode ter favorecido modelos como o Random Forest e Logistic Regression, que têm boa capacidade de generalização.
- **Limitações do TF-IDF:** Embora eficiente, o TF-IDF pode não capturar totalmente o contexto semântico dos textos, o que impacta especialmente modelos como SVM.
- **Complexidade computacional:** Random Forest obteve o melhor desempenho, mas exigiu mais recursos computacionais em comparação aos demais.

- **Simplicidade x Desempenho:** Enquanto Logistic Regression e Naive Bayes oferecem soluções rápidas e interpretáveis, o Random Forest se mostrou mais robusto e eficiente para este caso.

#### 6.4. Curvas ROC e AUC

A Figura 4 apresenta as curvas ROC dos modelos. O Random Forest destacou-se com uma curva mais acentuada ( $AUC = 0,96$ ), seguido pela Logistic Regression ( $AUC = 0,88$ ) e pelo Naive Bayes ( $AUC = 0,84$ ). O SVM teve um desempenho insatisfatório ( $AUC = 0,54$ ), indicando dificuldades em discriminar entre classes positivas e negativas.

### 7. Conclusão e Trabalhos Futuros

Este trabalho explorou a aplicação de técnicas de aprendizado de máquina para análise de sentimentos, utilizando um conjunto de dados de avaliações de medicamentos. A construção do pipeline incluiu etapas de pré-processamento textual, vetorização com TF-IDF e treinamento de modelos supervisionados. Os resultados obtidos permitiram avaliar a eficácia de diferentes abordagens e compreender as limitações do problema.

#### 7.1. Principais Conclusões

- **Melhor Desempenho:** O modelo *Random Forest* obteve a maior acurácia (90,6%) e a maior área sob a curva ROC ( $AUC = 0,96$ ), destacando-se como a melhor escolha para este problema.
- **Modelos Alternativos:** A *Logistic Regression* apresentou um desempenho sólido (Acurácia = 83,1%,  $AUC = 0,88$ ), sendo uma alternativa mais leve em termos computacionais.
- **Limitações do SVM:** O *Support Vector Machine* com kernel linear mostrou-se insuficiente para capturar as complexidades do dataset, obtendo uma  $AUC$  de apenas 0,54.
- **Impacto do Pré-Processamento:** As etapas de remoção de ruído, stemming e vetorização com TF-IDF foram cruciais para transformar os textos em um formato adequado para modelagem.

#### 7.2. Desafios Enfrentados

- **Desbalanceamento de Classes:** A predominância de sentimentos positivos (70%) afetou o desempenho de alguns modelos, como o SVM.
- **Dimensionalidade dos Dados:** A representação com TF-IDF (5000 características) aumentou a complexidade computacional, especialmente no treinamento do Random Forest.
- **Perda de Semântica:** Embora eficiente, o TF-IDF não captura relações contextuais profundas entre palavras, limitando o desempenho em alguns casos.

#### 7.3. Trabalhos Futuros

Com base nos resultados obtidos e nos desafios enfrentados, destacamos as seguintes oportunidades de melhoria e extensão:

- **Incorporação de Modelos Baseados em Redes Neurais:** Métodos como *Word Embeddings* (e.g., Word2Vec, GloVe) ou *Transformers* (e.g., BERT) poderiam ser

explorados para capturar a semântica contextual dos textos, potencialmente melhorando o desempenho. Trabalhos recentes, como o de Shiju e He [1], demonstraram a eficácia de modelos baseados em *Transformers* para a classificação de avaliações de medicamentos.

- **Utilização de Dados Complementares:** Outras colunas do dataset, como *usefulCount* e *condition*, poderiam ser integradas ao modelo para fornecer contexto adicional.
- **Balanceamento de Classes:** Técnicas como *oversampling* (e.g., SMOTE) ou *undersampling* poderiam ser aplicadas para lidar com o desbalanceamento do dataset.
- **Análise Multiclasse:** Estender a análise para prever mais de duas categorias de sentimentos (e.g., *negativo*, *neutro* e *positivo*).
- **Avaliação de Eficiência Computacional:** Comparar o desempenho dos modelos em termos de tempo de treinamento e inferência, especialmente em implementações para sistemas de produção.

#### 7.4. Contribuições do Trabalho

Este estudo contribuiu para:

- Demonstrar a aplicação de técnicas clássicas de aprendizado de máquina para análise de sentimentos.
- Comparar o desempenho de diferentes algoritmos em um problema textual real.
- Identificar os desafios e limitações inerentes a problemas de categorização de textos.

**Conclusão Final:** Este trabalho reforça o potencial do aprendizado de máquina para tarefas de análise textual e aponta direções para avanços futuros, especialmente no uso de técnicas mais sofisticadas, como modelos baseados em redes neurais, para explorar o contexto semântico dos dados.

#### References

- [1] Akhil Shiju e Zhe He. *Classifying Drug Ratings Using User Reviews with Transformer-Based Language Models*. Proceedings of the 2023 International Conference on Health Informatics, 2023.
- [2] K. Faceli, A. C. Lorena, e J. Gama. *Inteligência Artificial - Uma Abordagem de Aprendizado de Máquina*. Elsevier, Brasil, 2011, 1ª edição.
- [3] Stuart Russell e Peter Norvig. *Artificial Intelligence: A Modern Approach*. Pearson, EUA, 2020, 4ª edição.