Classificação Automática de Sentimentos em Resenhas de Livros em Português Brasileiro Utilizando Processamento de Linguagem Natural (PLN)

Bruno Gustavo Rocha[1], Felipe Carvalho de Alencar[2], Maria Gabriela de Barros do Amaral[1], Raphaela Polonis Marques Maria[3]

[1]Faculdade de Computação e Informática - Universidade Presbiteriana Mackenzie (UPM)
Caixa Postal 01302907 – São Paulo – SP – Brazil
{10400926@mackenzista.com.br, 10409804@mackenzista.com.br, 10409037@mackenzista.com.br, 104088423@mackenzista.com.br

Abstract. *This work presents the development of a Natural Language Processing (NLP) model for the automatic sentiment classification of book reviews written in Brazilian Portuguese. This domain was selected due to the large volume of literary evaluations available on digital platforms and the inherent difficulty of manually analyzing such content at scale. A publicly available and balanced dataset of positive and negative reviews was employed, combined with a preprocessing pipeline, TF-IDF feature extraction, and supervised machine learning algorithms. This report includes the theoretical background, exploratory data analysis, methodological description, and ethical considerations related to the use of AI models. The experimental results and final discussions are presented in the Results section, which concludes the final stage of the project.*

Resumo. *Este trabalho apresenta o desenvolvimento de um modelo de Processamento de Linguagem Natural (PLN) para a classificação automática de sentimentos em resenhas de livros escritas em português brasileiro. A escolha desse domínio justifica-se pelo grande volume de avaliações literárias disponíveis em plataformas digitais e pela dificuldade de analisar manualmente essas opiniões em larga escala. Utilizou-se um conjunto público e balanceado de resenhas positivas e negativas, aplicado a um pipeline de pré-processamento textual, vetorização por TF-IDF e algoritmos de aprendizado supervisionado. O relatório abrange a contextualização teórica, a análise exploratória dos dados, as etapas metodológicas empregadas e as considerações éticas relacionadas ao uso de modelos de IA. Os resultados experimentais e discussões complementares são apresentados na seção de Resultados, que compõem a etapa final do projeto.*

1. Introdução

1.1 Contextualização

Com o crescimento de plataformas digitais e da produção de conteúdo textual na internet, compreender automaticamente a opinião de usuários tornou-se uma necessidade para diversas organizações. Dentro desse contexto, a análise de sentimentos destaca-se como uma das aplicações mais relevantes do

Processamento de Linguagem Natural (PLN), permitindo identificar automaticamente percepções positivas, negativas ou neutras em textos.

Resenhas de livros representam uma fonte rica de opiniões espontâneas, mas sua análise manual é inviável em larga escala. Assim, modelos automatizados contribuem para resumir percepções, apoiar recomendações e facilitar decisões editoriais e comerciais.

1.2 Justificativa

A análise humana de milhares de resenhas é um processo demorado, subjetivo e suscetível a vieses individuais. A partir de dados textuais não estruturados, técnicas de PLN e aprendizado de máquina permitem construir modelos capazes de interpretar sentimentos com rapidez e precisão. Além disso, há uma carência de projetos acadêmicos focados em português brasileiro, especialmente no domínio de avaliações literárias, reforçando a relevância prática e científica deste estudo.

1.3 Objetivo

O objetivo deste projeto é desenvolver e avaliar um modelo capaz de identificar automaticamente se uma resenha de livro é positiva ou negativa. O modelo utiliza um dataset público composto exclusivamente por textos classificados nessas duas categorias, justificando a adoção de uma abordagem binária.

Objetivos específicos:

- realizar limpeza e pré-processamento textual;
- transformar textos em representações numéricas;
- aplicar e justificar algoritmos adequados;
- realizar análise exploratória do dataset;
- avaliar o desempenho do modelo

1.4 Opção do Projeto

Opção escolhida: PLN — Processamento de Linguagem Natural, com foco específico em Classificação de Sentimentos, utilizando técnicas tradicionais de aprendizado de máquina supervisionado aplicadas a textos em português brasileiro.

2. Fundamentação Teórica (Resumida)

A análise de sentimentos é uma tarefa de PLN que visa identificar a polaridade expressa em um texto, atribuindo categorias como positivo, negativo ou neutro. Bird,

Klein e Loper (2009) destacam que o PLN utiliza métodos linguísticos e estatísticos que possibilitam transformar linguagem humana em representações computáveis.

Segundo Jurafsky e Martin (2023), modelos clássicos de aprendizado supervisionado, como Regressão Logística, Naive Bayes e SVM, continuam eficazes para tarefas de classificação textual, especialmente quando aliados a representações como Bag-of-Words e TF-IDF. Essas abordagens são adequadas em cenários com vocabulário amplo, textos curtos e datasets pequenos ou médios.

O pré-processamento textual é essencial para reduzir ruído, padronizar tokens e melhorar a qualidade das representações numéricas. Entre as etapas comuns estão: normalização, tokenização, remoção de pontuação e stopwords. A vetorização por TF-IDF calcula a relevância de termos com base em sua frequência local e inversa nos documentos, sendo amplamente utilizada devido à sua simplicidade e eficácia.

3. Descrição do Problema

O problema abordado consiste em classificar automaticamente resenhas de livros como positivas ou negativas. Essas resenhas apresentam grande diversidade de estilos, uso de expressões informais, pontuações enfáticas e variações linguísticas próprias do português brasileiro.

Embora existam modelos que tratam sentimentos em três classes (positivo, negativo e neutro), o dataset escolhido não contém avaliações neutras, e sim somente textos fortemente positivos ou negativos. Assim, o problema é tratado como uma tarefa de classificação binária.

4. Aspectos Éticos e Responsabilidade

A utilização de IA em análise de sentimentos deve considerar aspectos éticos importantes:

- Enviesamento: o modelo pode reproduzir preconceitos presentes no dataset.
- Privacidade: é essencial garantir que dados pessoais não sejam utilizados indevidamente.
- Uso responsável: decisões tomadas com base na análise automatizada devem ser complementadas por avaliação humana, principalmente em contextos críticos.

5. Dataset

O dataset utilizado é o **Amazon Brazilian Portuguese Books Reviews Dataset**, contendo:

- 2000 resenhas de livros;

- 1000 positivas;
- 1000 negativas;
- Formato: arquivos .txt com uma resenha por linha.

A análise exploratória revelou:

- textos curtos, com média de 12–20 palavras;
- presença de informalidades, pontuações repetidas e emojis;
- termos positivos frequentes como *amei*, *ótimo*, *emocionante*;
- termos negativos frequentes como *péssimo*, *confuso*, *cansativo*;
- clara distinção entre vocabulários das classes.

6. Metodologia e Resultados Esperados

A metodologia adotada segue as seguintes etapas:

6.1 Pré-processamento

O pré-processamento consistiu em etapas essenciais para preparar as resenhas para a modelagem. Inicialmente, todos os textos foram convertidos para letras minúsculas e tiveram pontuações, emojis e caracteres especiais removidos, garantindo padronização. Em seguida, os textos foram tokenizados e passaram pela remoção de stopwords da língua portuguesa, reduzindo ruído e destacando termos mais informativos. Por fim, aplicou-se a vetorização por TF-IDF, que transforma as palavras em valores numéricos ponderados conforme sua relevância, permitindo que o modelo utilize os textos de forma eficiente.

6.2 Modelagem

O modelo principal utilizado é a Regressão Logística, escolhida por sua boa performance em problemas de classificação textual binária, simplicidade, interpretabilidade e capacidade de lidar com vetores esparsos gerados via TF-IDF. A base foi dividida em 80% para treino e 20% para teste, conforme práticas recomendadas.

6.3 Avaliação

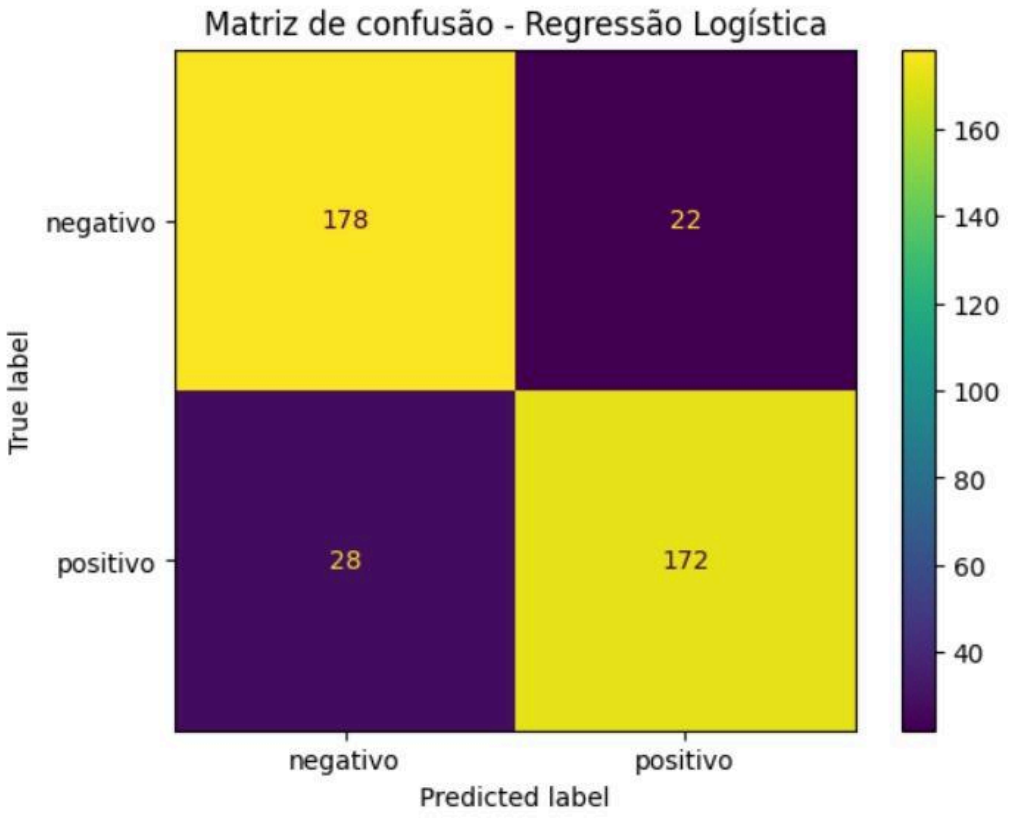
As métricas utilizadas serão:

- acurácia;
- precisão;
- recall;
- F1-score;
- matriz de confusão.

Predições que o modelo realizou:

- [positivo] Amei a história, personagens cativantes e final emocionante!
- [negativo] Péssimo, confuso e cansativo. Não recomendo.
- [negativo] Achei ok, mas esperava mais do enredo.
- [negativo] Não gostei do livro. A história é arrastada e pouco interessante.
- [negativo] Personagens sem graça e enredo previsível. Esperava bem mais.
- [negativo] Foi uma decepção. Escrita confusa e capítulos cansativos.
- [negativo] O livro não prende a atenção. Muito repetitivo e mal desenvolvido.
- [negativo] Final extremamente fraco. Não recomendo a leitura.
- [positivo] Adorei este livro! A escrita é envolvente e os personagens são muito cativantes.
- [positivo] Uma leitura maravilhosa. História emocionante e bem construída do início ao fim.
- [positivo] Simplesmente fantástico. Recomendo para todos que gostam de uma boa narrativa.
- [positivo] Um dos melhores livros que já li. Profundo, inteligente e muito inspirador.
- [positivo] Leitura leve e prazerosa, impossível parar até terminar.

Imagens dos Resultados:



Nuvem de Palavras - Negativas (limpas)

Nuvem de Palavras - Positivas (limpas)


```
Acurácia: 0.875
              precision    recall  f1-score   support

    negativo       0.86      0.89      0.88       200
    positivo       0.89      0.86      0.87       200

    accuracy                           0.88       400
   macro avg       0.88      0.88      0.87       400
weighted avg       0.88      0.88      0.87       400
```

7. Conclusão e Próximos Passos

O modelo de Regressão Logística, combinado com vetorização TF-IDF, apresentou desempenho satisfatório na tarefa de classificação de resenhas em *positivas* e *negativas*.

Como próximos passos, seria interessante:

- comparar com outros algoritmos (por exemplo, Naive Bayes ou SVM);

- ajustar hiperparâmetros e testar diferentes configurações de TF-IDF;

- explorar representações mais avançadas, como embeddings (Word2Vec, FastText, BERT);

- ampliar o conjunto de dados para aumentar a robustez do modelo.

Este notebook funciona como um pipeline completo e didático de análise de sentimentos em português.

8. Endereço GitHub e Endereço do vídeo no Youtube

Link GitHub:

https://github.com/raphaelapolonis/Projeto-IA

Link do vídeo no Youtube:

https://youtu.be/lPA1Q5Jggxc

9. Referências

- BIRD, Steven; KLEIN, Ewan; LOPER, Edward. Natural Language Processing with Python. O'Reilly Media, 2009.
- FELICIANA, Larissa. Amazon Brazilian Portuguese Books Reviews Dataset. GitHub, 2020.
- JURAFSKY, Daniel; MARTIN, James H. Speech and Language Processing. 3rd ed., Draft, 2023.