**FACULDADE SENAC PERNAMBUCO**

**TECNÓLOGO EM ANÁLISE E DESENVOLVIMENTO DE SISTEMAS**

**ANDERSON VELOSO**

**RAMILLE CORDEIRO**

**REBEKA DIAS**

**SOPHIA ALBUQUERQUE**

# SISTEMA DE CLASSIFICAÇÃO AUTOMÁTICA DE OPINIÕES DE CLIENTES

RECIFE – PE

2025

Sumário

[SISTEMA DE CLASSIFICAÇÃO AUTOMÁTICA DE OPINIÕES DE CLIENTES 1](#_Toc199791913)

[1. Introdução 3](#_Toc199791914)

[2. Descrição do Problema 3](#_Toc199791915)

[Coleta dos dados 3](#_Toc199791916)

[Análise dos dados 4](#_Toc199791917)

[Pré-processamento do texto 5](#_Toc199791918)

[Representação dos Textos 7](#_Toc199791919)

[4. Modelos de Classificação 7](#_Toc199791920)

[Modelo supervisionado 7](#_Toc199791921)

[Modelo Pré-Treinado 7](#_Toc199791922)

[Comparação de modelo Supervisionado x Pré-treinado 7](#_Toc199791923)

[Simulação de novos reviews 8](#_Toc199791924)

[5. Conclusão 8](#_Toc199791925)

[6. Referências 10](#_Toc199791926)

## 1. Introdução

A popularização de assistentes virtuais, como o Amazon Alexa, transformou a forma como os consumidores interagem com a tecnologia no cotidiano. Com funcionalidades que vão desde o controle de dispositivos domésticos inteligentes até a execução de tarefas por comando de voz, o Alexa tem gerado uma grande quantidade de avaliações por parte dos usuários em plataformas de e-commerce. Essas avaliações, geralmente em formato textual, oferecem uma rica fonte de dados sobre a experiência do consumidor.

Este trabalho propõe um sistema automatizado para análise de reviews do Amazon Alexa, utilizando técnicas de Processamento de Linguagem Natural (PLN) para interpretar o conteúdo textual das avaliações e classificá-las automaticamente como positivas ou negativas. A metodologia envolve etapas de pré-processamento linguístico, extração de características e aplicação de algoritmos de aprendizado de máquina para a classificação de sentimentos.

## 2. Descrição do Problema

Com o crescimento exponencial do comércio eletrônico, consumidores passaram a confiar cada vez mais nas avaliações de produtos publicadas por outros usuários para tomar decisões de compra. Essas avaliações, geralmente em formato textual, contêm informações valiosas sobre a experiência do consumidor, a qualidade do produto e a confiabilidade do vendedor. No entanto, o grande volume de reviews disponíveis torna inviável a análise manual dessas informações em larga escala.

Além disso, a subjetividade e a diversidade linguística presentes nos textos tornam a tarefa de interpretação ainda mais complexa. Isso representa um desafio para empresas que desejam monitorar a satisfação dos clientes, identificar padrões de uso e tomar decisões baseadas em dados reais.

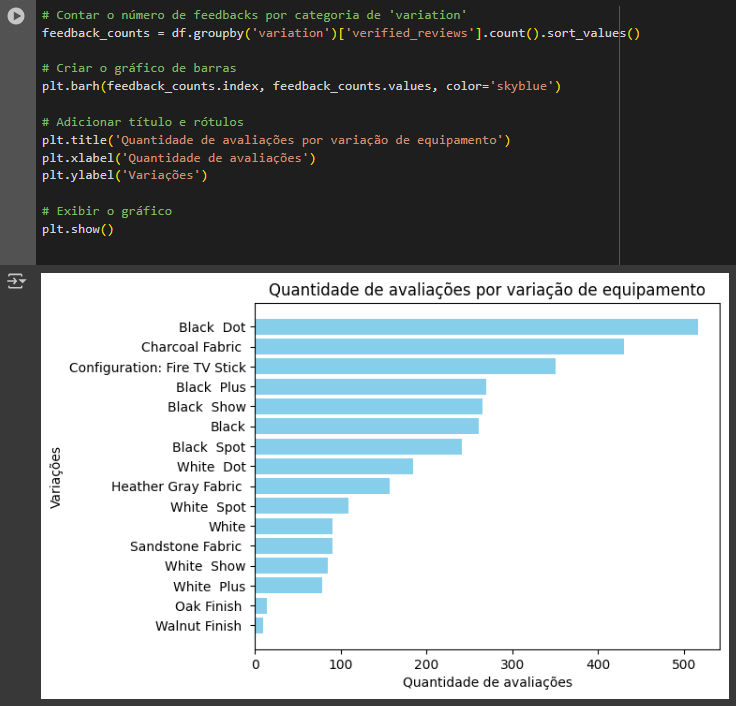
Diante desse cenário, surge a necessidade de um sistema automatizado capaz de processar e interpretar essas avaliações de forma eficiente.3. Metodologia

### Coleta dos dados

Os dados foram coletados da plataforma Kaggle, contendo informações de avaliações de usuários, como texto, data e variante do dispositivo Alexa. O arquivo utilizado estava no formato TSV (Tab Separated Values).

### Análise dos dados

Inicialmente, foram removidos os dados nulos. Em seguida, foram analisados os valores únicos das colunas relevantes e a distribuição das respostas. Gráficos foram gerados para visualizar essas informações.

  
Uma imagem contendo Interface gráfica do usuário

O conteúdo gerado por IA pode estar incorreto.Texto

O conteúdo gerado por IA pode estar incorreto.

### Pré-processamento do texto

Foram aplicadas técnicas de PLN como:

**Tokenização**: divisão do texto em palavras.

**Remoção de stopwords**: eliminação de palavras irrelevantes.

**Normalização**: remoção de pontuações, números, caracteres especiais e conversão para minúsculas.

**Lematização**: redução das palavras à sua forma base.

Texto

O conteúdo gerado por IA pode estar incorreto.

Também foi gerada uma nuvem de palavras para visualizar a frequência dos termos mais comuns nas avaliações.

Uma imagem contendo jornal

O conteúdo gerado por IA pode estar incorreto.

## Representação dos Textos

Para representar numericamente os textos, utilizou-se o método TF-IDF (Term Frequency-Inverse Document Frequency), que permite transformar os dados textuais em vetores de características, facilitando sua aplicação em modelos de aprendizado de máquina.

## 4. Modelos de Classificação

### Modelo supervisionado

Foi desenvolvido um modelo supervisionado utilizando o algoritmo SVM (Support Vector Machine), aplicado sobre os textos representados com TF-IDF. Abaixo, apresentamos suas métricas de desempenho:

Acurácia: 0.9397

Precisão: 0.9450

Recall: 0.9913

F1-Score: 0.9676

Apesar dos bons resultados nas métricas, observamos que o modelo frequentemente classificava avaliações negativas como positivas. Ao investigar a causa, identificamos um desequilíbrio na base de dados, onde mais de 90% das avaliações eram positivas. Isso levou o modelo a se "viciar" em prever a classe positiva, comprometendo sua efetividade em identificar corretamente avaliações negativas.

### Modelo Pré-Treinado

Como alternativa, utilizamos o modelo pré-treinado BERT (Bidirectional Encoder Representations from Transformers), conhecido por seu excelente desempenho em tarefas de NLP. Em um primeiro teste com uma frase claramente positiva, o BERT classificou corretamente. Em seguida, aplicamos o modelo nas 100 primeiras avaliações da base, e os resultados foram mais coerentes e satisfatórios quando comparados ao modelo SVM.

### Comparação de modelo Supervisionado x Pré-treinado

Apesar das boas métricas do modelo SVM, seu desempenho prático foi limitado pelo viés causado pelo desbalanceamento dos dados. Por outro lado, o modelo BERT, por já possuir um entendimento profundo da linguagem natural, demonstrou maior precisão, equilíbrio e robustez nas classificações, destacando-se como a melhor escolha para este cenário.

### Simulação de novos reviews

Implementamos uma função que permite ao usuário inserir uma nova avaliação e visualizar a classificação feita por ambos os modelos. Assim como nos testes com a base original, o modelo BERT apresentou maior assertividade nas classificações em comparação ao modelo supervisionado SVM.

Interface gráfica do usuário, Texto

O conteúdo gerado por IA pode estar incorreto.

Interface gráfica do usuário, Texto

O conteúdo gerado por IA pode estar incorreto.

## 5. Conclusão

A partir da análise de avaliações sobre o Amazon Alexa, conseguimos compreender a importância do pré-processamento textual, da escolha dos algoritmos e do balanceamento dos dados para alcançar resultados mais confiáveis.

Entre os principais aprendizados, destacamos o impacto que o desbalanceamento de classes pode ter sobre o desempenho de modelos supervisionados, como o SVM, que apesar de boas métricas iniciais.

Entre as limitações do trabalho, podemos citar a simplicidade dos dados utilizados, a quantidade relativamente pequena de textos analisados e a ausência de um processo formal de balanceamento da base para o modelo supervisionado.

Como possíveis melhorias futuras:

* Aplicar técnicas de balanceamento como SMOTE ou undersampling para melhorar a performance do modelo SVM;
* Testar outros modelos de linguagem pré-treinados, como RoBERTa ou DistilBERT;
* Ampliar o conjunto de dados e incluir uma etapa de validação cruzada mais robusta;
* Desenvolver uma interface interativa para uso real do sistema em e-commerce.

**Link do Colab:** https://colab.research.google.com/drive/1TvD54KcV7fpylLh5PWJFcXv0wTAZn6fY?usp=sharing

## 6. Referências

<https://medium.com/@lainetnr/pln-pr%C3%A9-processamento-de-texto-315238b1f6cd>

<https://matplotlib.org/>

<https://scikit-learn.org/stable/supervised_learning.html>

<https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html>

**10. Apêndices (opcional)**

* Trechos de código
* Prints de telas
* Logs de execução