



Desafio Técnico – Cientista de Dados Sênior

Análise e Enriquecimento de Dados de Livros com IA
André Rizzo – Junho/2025



Desafio

A A3Data foi contratada para automatizar a análise de avaliações literárias com técnicas de NLP e LLMs, substituindo processos manuais por inteligência analítica que gere insights sobre autores, gêneros e leitores — com foco em agilidade e redução de custos.

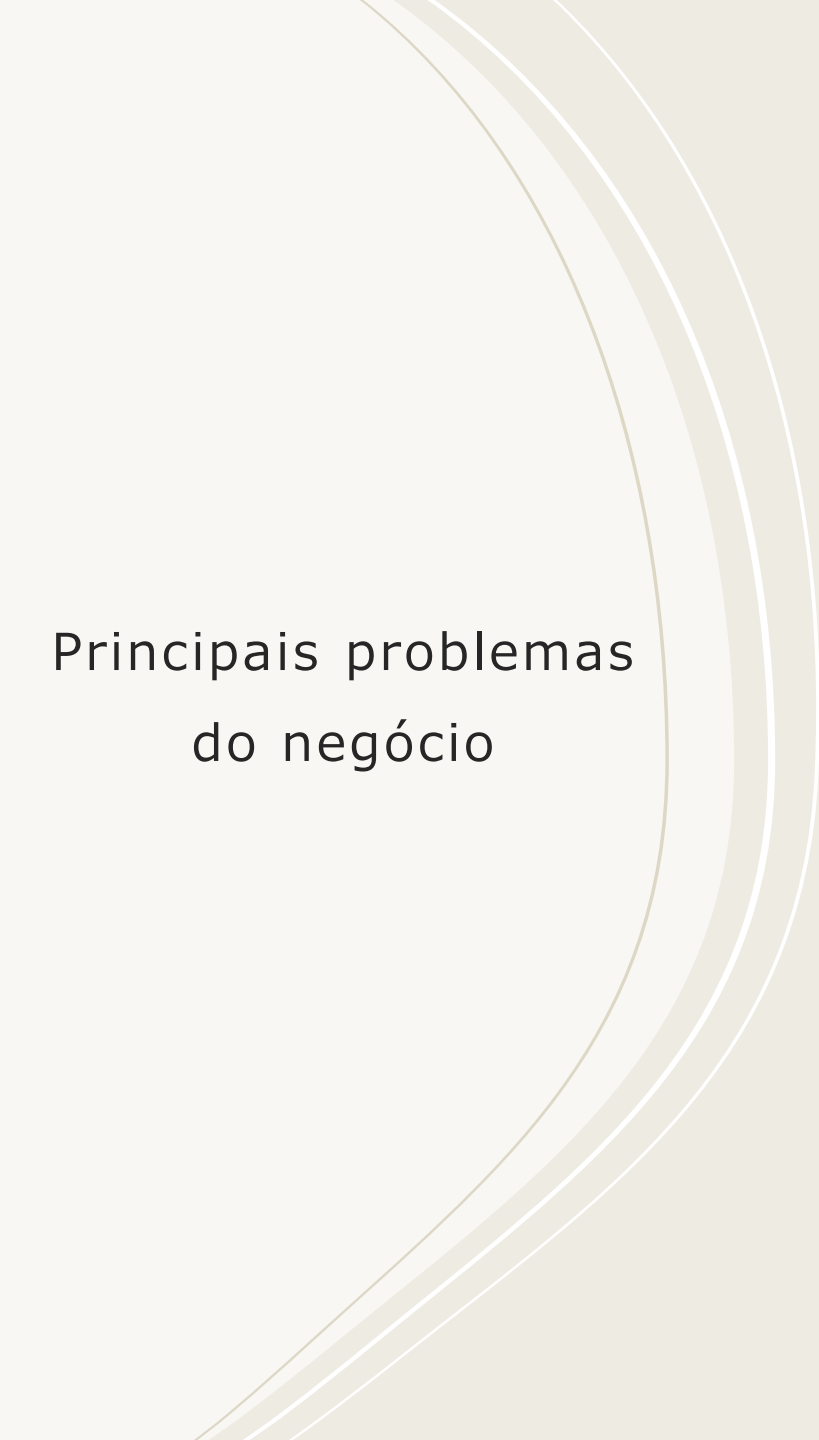
Principais problemas do negócio

1. Alto Custo e Baixa Escalabilidade

- Cada novo conjunto de avaliações exige o mesmo esforço humano.
- O modelo atual não escala com o crescimento da base de dados.

2. Subutilização de Dados Textuais

- A editora possui uma mina de ouro nas avaliações escritas, mas o foco atual está apenas nas notas (ratings).
- Faltam mecanismos para extrair valor dos **conteúdos subjetivos**.



Principais problemas do negócio

3. Tomada de Decisão Baseada em Intuição

- Sem ferramentas analíticas, os gestores tomam decisões baseadas em feeling ou análise superficial.
- Isso pode levar a erros na priorização de autores ou gêneros.

4. Falta de Visibilidade de Tendências

- Sem análise automatizada, a editora perde tempo para identificar padrões de comportamento do leitor — como tendências emergentes, mudanças no gosto do público ou feedbacks negativos recorrentes.

Estimativa de Impacto

1. Redução de Custo Operacional Direto

- A análise manual atual consome **5 analistas × 3 dias**, com custo estimado de **R\$15.000 por tarefa**.
- Com a automação proposta, esse custo pode ser reduzido em mais de **90%**, com retorno quase imediato do investimento.

2. Aumento de Eficiência e Velocidade

- Hoje: processo manual e demorado.
- Com a solução: geração de insights em **minutos**, permitindo decisões em tempo real sobre lançamentos, campanhas ou curadoria editorial.

Estimativa de Impacto

3. Melhoria na Tomada de Decisão

- Análises subjetivas dão lugar a evidências baseadas em dados.
- A editora poderá identificar:
 - Autores com maior aceitação
 - Gêneros em alta
 - Críticas recorrentes que impactam as vendas

4. Valorização do Capital Intelectual

- Liberação dos analistas para atividades estratégicas: entrevistas com leitores-chave, desenvolvimento de novos produtos, planejamento de marketing de conteúdo.



Estimativa de Impacto

5. Inovação e Diferenciação no Mercado

- Uso de **LLMs e NLP avançado** posiciona a livraria como inovadora no uso de IA.
- Possibilidade de criar ferramentas internas ou B2B para curadoria, recomendação ou diagnóstico de portfólio editorial.

Análise Exploratória Resumida

```
===== BOOKS_DATA =====  
Número de registros: 212404      Número de variáveis: 10  
  
===== BOOKS_RATING =====  
Número de registros: 3000000      Número de variáveis: 9
```

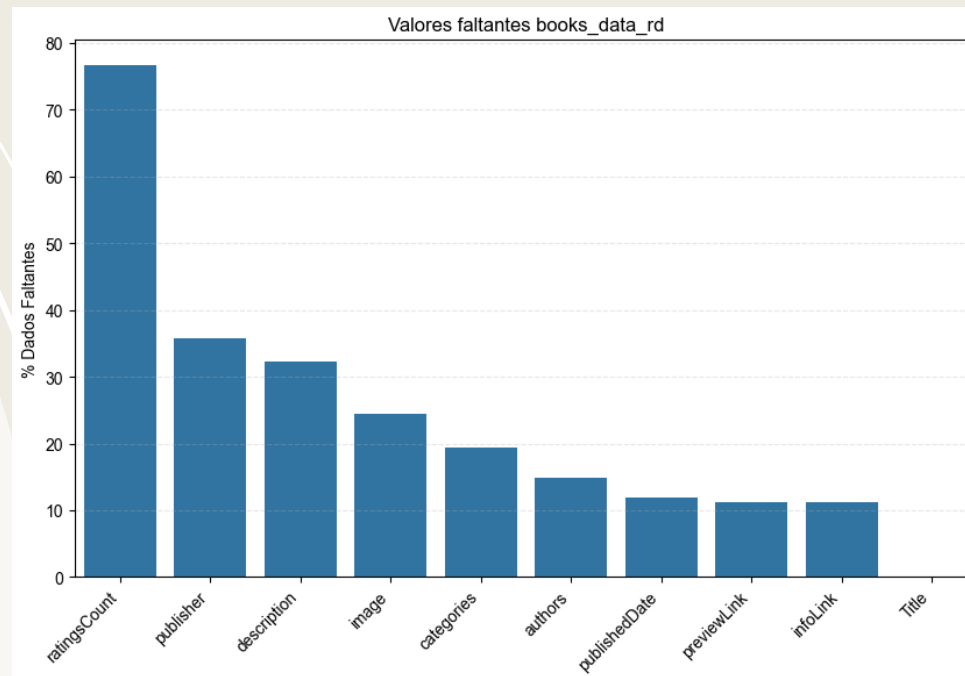
books_rating com 3 milhões de registros

books_data com mais de 200 mil registros

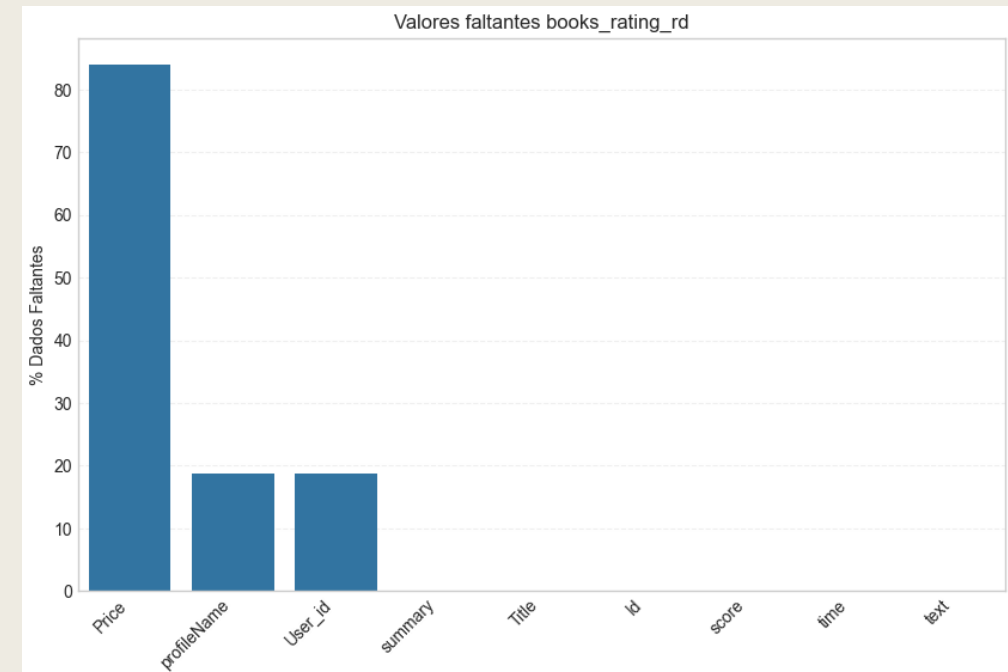
```
===== BOOKS_DATA =====  
  
<class 'pandas.core.frame.DataFrame'>  
RangeIndex: 212404 entries, 0 to 212403  
Data columns (total 10 columns):  
#   Column          Non-Null Count  Dtype  
---  ---  
0   Title            212403 non-null object  
1   description       143962 non-null object  
2   authors          180991 non-null object  
3   image            160329 non-null object  
4   previewLink      188568 non-null object  
5   publisher        136518 non-null object  
6   publishedDate    187099 non-null object  
7   infoLink         188568 non-null object  
8   categories       171205 non-null object  
9   ratingsCount     49752 non-null  float64  
dtypes: float64(1), object(9)  
memory usage: 16.2+ MB
```

```
===== BOOKS_RATING =====  
  
<class 'pandas.core.frame.DataFrame'>  
RangeIndex: 3000000 entries, 0 to 2999999  
Data columns (total 9 columns):  
#   Column          Dtype  
---  ---  
0   Id              object  
1   Title           object  
2   Price           float64  
3   User_id        object  
4   profileName     object  
5   score           float64  
6   time            int64  
7   summary         object  
8   text            object  
dtypes: float64(2), int64(1), object(6)  
memory usage: 206.0+ MB
```

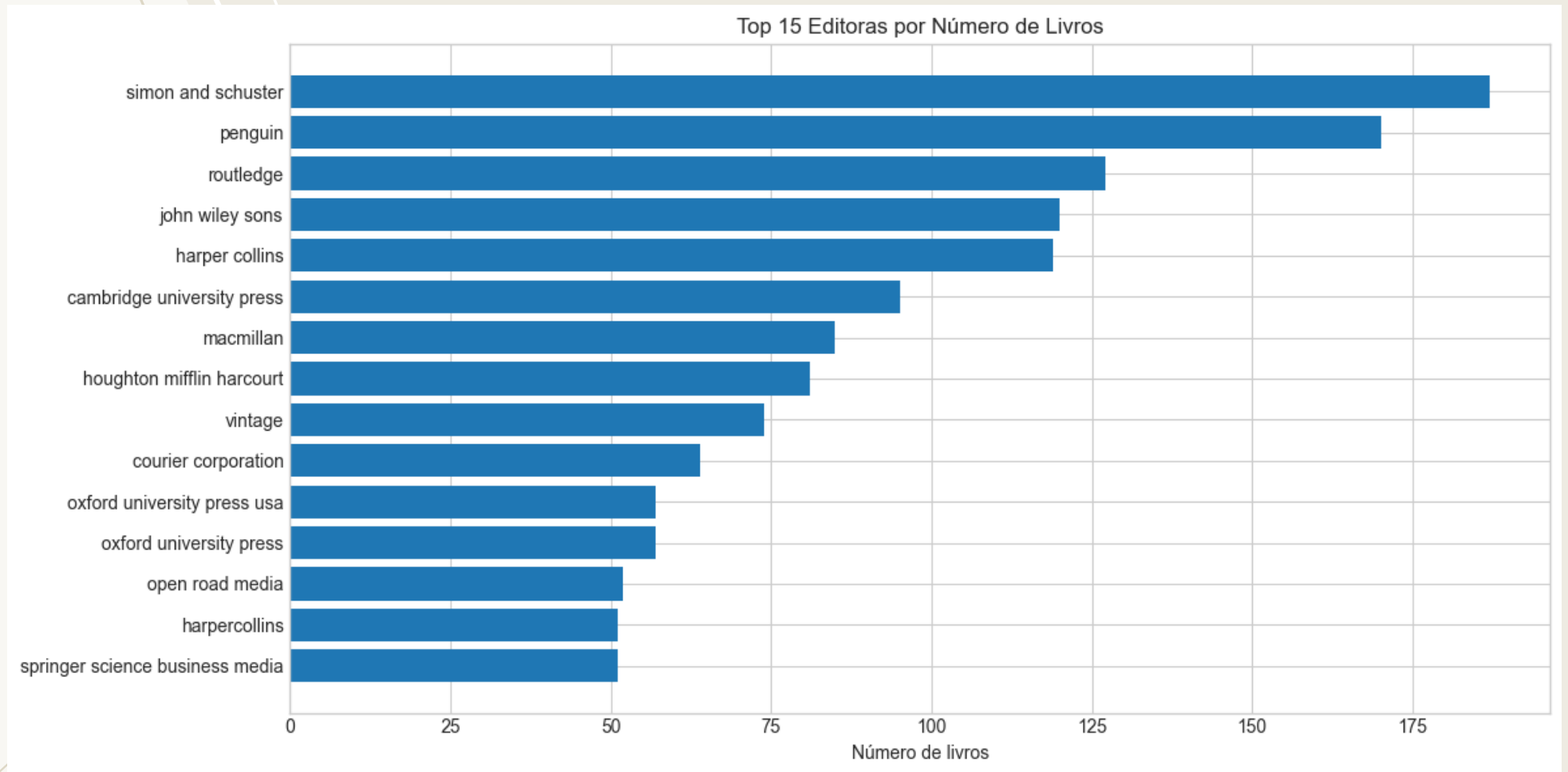

Análise Exploratória Resumida



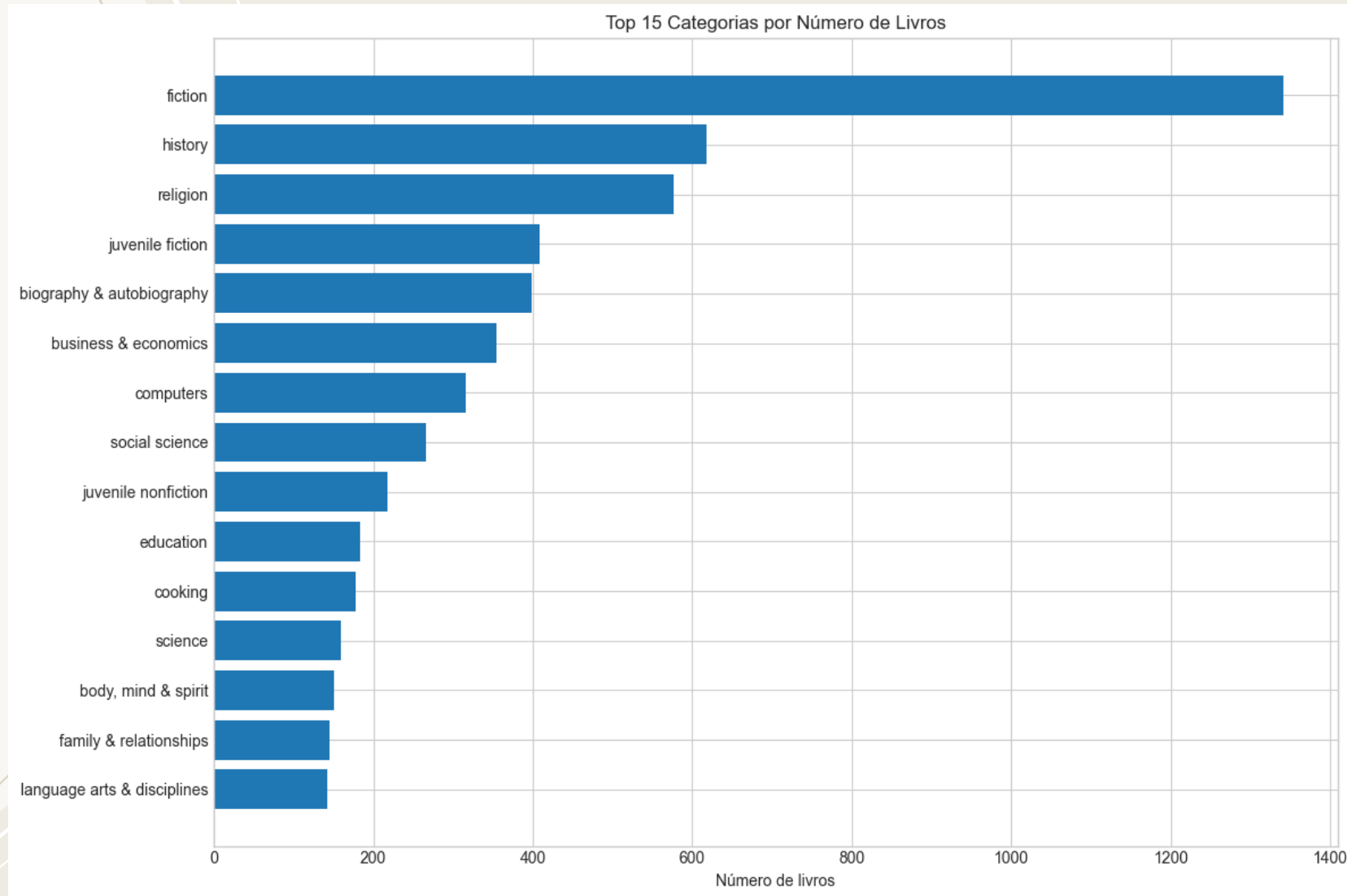
Missing Values



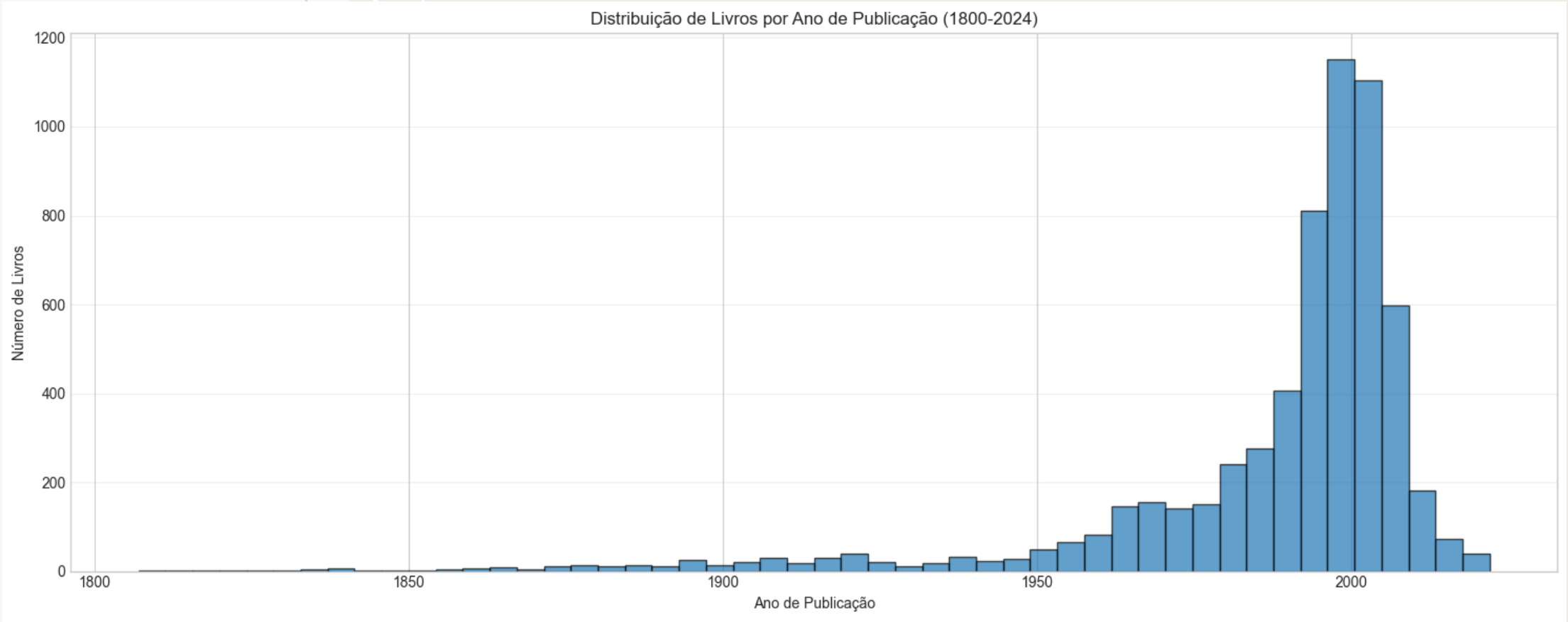
Análise Exploratória Resumida



Análise Exploratória Resumida



Análise Exploratória Resumida



Processo Utilizado

1. Inspeção e Amostragem Inteligente

Amostragem de 10% dos dados (~3 milhões de registros) para garantir agilidade com representatividade.

2. Limpeza e Padronização Textual

Normalização conservadora dos campos

3. Remoção de Duplicatas Semânticas

TF-IDF + Similaridade Cosseno para eliminar registros redundantes e manter os mais completos.

4. Enriquecimento de dados com OpenLibrary API

Pipeline paralelo (40 threads) com matching semântico para imputação de metadados ausentes.

Processo Utilizado

5. Análise de Sentimentos com VADER

Classificação dos reviews em positivo, neutro e negativo + análise cruzada com as notas numéricas.

6. Geração de Resumos com LLM (GPT-4o)

Uso de LLM para sintetizar os principais elogios, críticas e recomendações executivas por sentimento.

7. Armazenamento em Banco SQLite

Todos os dados processados foram organizados em uma base local SQLite para facilitar consultas e integração com o frontend.

8. Criação de Frontend Interativo

Protótipo funcional em Streamlit que permite buscar livros, visualizar resumos de IA e insights de negócio em tempo real.

Aplicação – Livros com Pior Avaliação

POC - Análise de Livros e Reviews

Dashboard para tomada de decisões baseada em dados

✓ Banco de dados conectado com sucesso

Atualizar Status

Versão: 1.1.0

⚠ Livros Mais Problemáticos

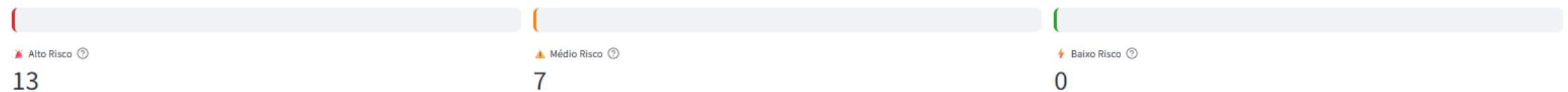
Identificação de livros com alta taxa de reviews negativos e baixo sentimento.

Número de livros

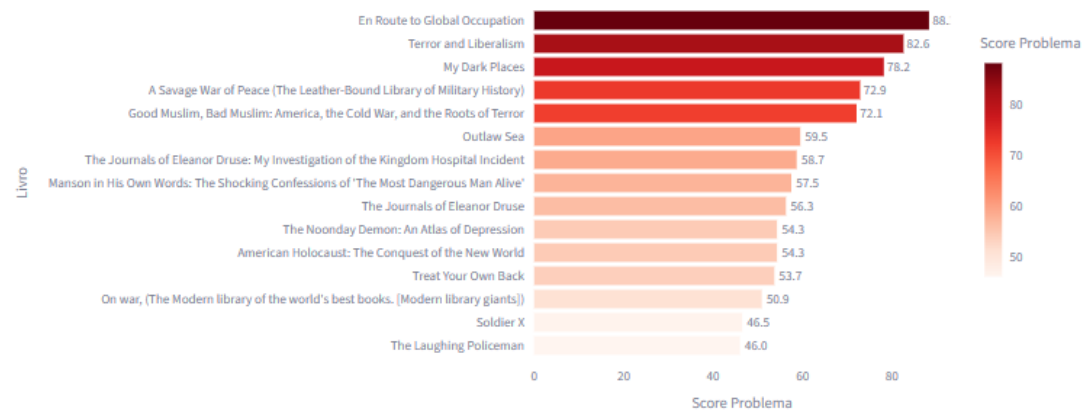


☐ Mostrar detalhes

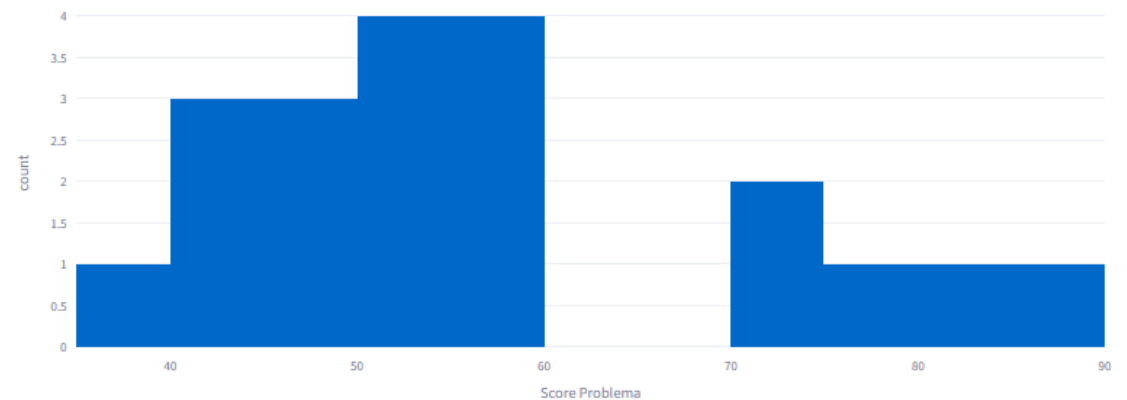
💣 Níveis de Risco



Score de Problema por Livro



Distribuição dos Scores de Problema



Aplicação – Resumo por LLM

☹️ Análise IA: The Hobbit

📊 Resumo Executivo

Total de Reviews	Reviews Positivos	Reviews Negativos	Sentimento Médio
2151	1977	133	0.723

👜 Recomendação de Negócio

✅ **PROMOVER** - Livro com excelente recepção (Alto volume - dados confiáveis)

Prioridade de Negócio: Alta

☹️ Resumos Gerados por IA

👍 Reviews Positivos 🙄 Reviews Negativos 😐 Reviews Neutros





😊 10 reviews analisados


😊 Resumo da IA:

Resumo Executivo

- Os leitores destacam a interpretação inspiradora de Martin Freeman como Bilbo Baggins e a riqueza do estilo de escrita de Tolkien, que, apesar de complexo, é apreciado por sua profundidade. Muitos expressam entusiasmo pela adaptação cinematográfica e a conexão com a obra original, reforçando a relevância de "The Hobbit" no universo literário.
- Os três aspectos mais elogiados pelos leitores são:
 - A atuação de Martin Freeman como Bilbo Baggins.
 - A complexidade e riqueza do estilo de escrita de Tolkien.
 - A conexão emocional e a nostalgia geradas pela obra.
- O público-alvo ideal inclui fãs de literatura fantástica, admiradores de Tolkien e aqueles que apreciam narrativas ricas e complexas.

GitHub


 README  MIT license  



Análise Automatizada de Reviews de Livros




Desafio Técnico — A3Data



Solução completa de NLP e IA para automatizar a análise de avaliações literárias, reduzindo o tempo de análise de 3 dias para apenas 5 minutos, com economia de até R\$15 mil por análise.



O Desafio de Negócio

Editoras enfrentam um processo **lento e custoso** para extrair insights de milhares de reviews. Entre os objetivos:

-  Identificar livros com recepção negativa
-  Selecionar usuários com opiniões relevantes
-  Avaliar performance de autores, gêneros e editoras

 Custo estimado da análise manual: R\$ 15.000,00  Tempo médio atual: 3 dias

<https://github.com/andrerizzo/A3Data>

[illegible]