Avaliação Automática de Redações Narrativas

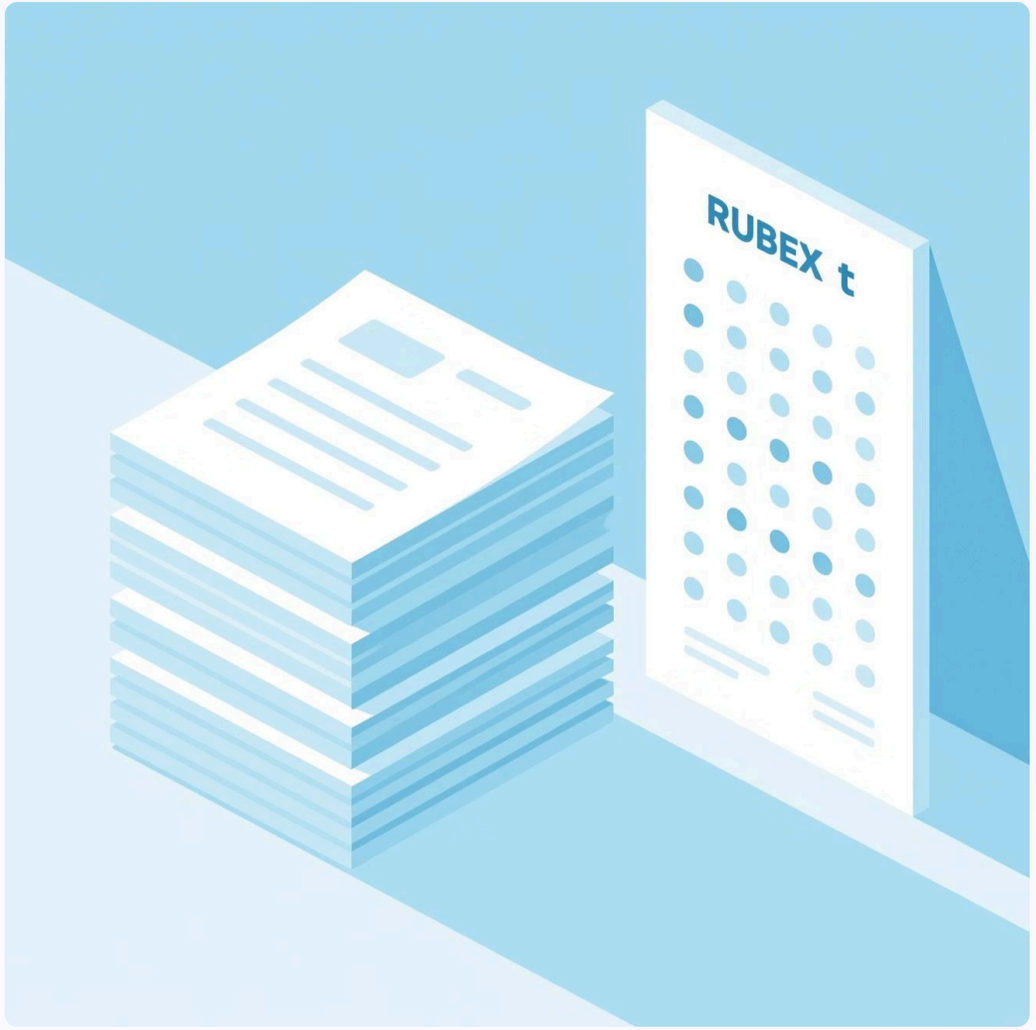
Análise da performance de diferentes formas de pré-processamento e extração de atributos pra avaliar redações narrativas.

Contexto do Projeto

Dataset de Redações Narrativas

O conjunto de dados contém **1.235 redações** escritas por estudantes do 5º ao 9º ano de escolas públicas brasileiras. Todas as redações foram digitalizadas, anonimizadas e avaliadas por dois avaliadores humanos.

As redações foram analisadas segundo uma rubrica que considera quatro competências obrigatórias, cada uma avaliada em uma escala de 1 a 5.



Competências Avaliadas

Registro Formal

Adequação à norma culta e uso apropriado da linguagem formal

Coerência Temática

Desenvolvimento lógico e consistente do tema proposto

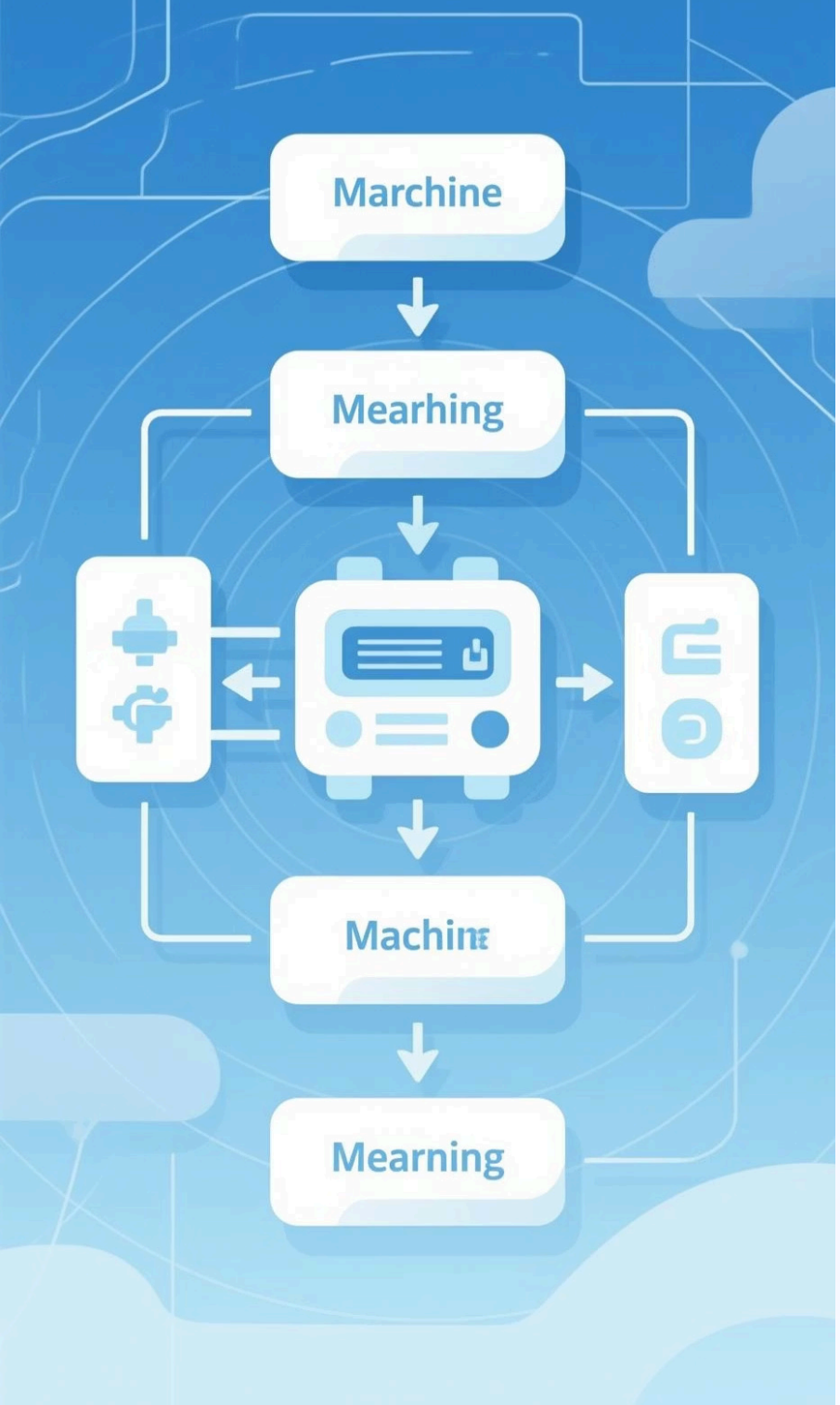
Estrutura Retórica Narrativa

Organização adequada dos elementos narrativos

Coesão

Conexão entre ideias e parágrafos do texto

Para este estudo, foi escolhida a competência de coesão como foco da classificação automática.



Metodologia Adotada

01

Seleção da Label

Escolha da competência "coesão" como variável de interesse

02

Modelo de Classificação

Regressão Logística - classificador simples, rápido e eficiente com TF-IDF/BoW

03

Pré-processamento

Teste de múltiplas técnicas de limpeza e normalização textual

04

Extração de Características

Comparação entre diferentes métodos de vetorização

Resultados - Item A

Técnicas de Pré-processamento

Foram testadas oito diferentes abordagens de pré-processamento utilizando **CountVectorizer** como extrator base. Os resultados indicam comportamentos distintos para cada técnica:

Acurácia por Técnica

- Sem processamento: baseline
- Remover stopwords: melhor resultado
- Stemming/Lemmatização: queda significativa
- Pipeline completa: sem ganho

Técnica	Acurácia
Sem pré-processamento	65.13%
Remover maiúsculas	65.13%
Remover símbolos (ex [P])	64.86%
Remover stopwords	67%
Remover pontuações	64.32%
Stemming	63.78%
Lemmatization	62.97%
Norm espaços	65.13%
Pipeline completa	65.13%

Análise do Pré-processamento

✓ Remoção de Stopwords

Único procedimento que **aumentou a acurácia**. Palavras como "de", "da" não contribuem para identificar coesão textual.

≈ Normalização Básica

Remoção de maiúsculas e normalização de espaços não geraram impacto significativo nos resultados.

Transformações Morfológicas

Stemming e lemmatização causaram **grande perda de acurácia**. A forma original das palavras é importante para avaliar coesão.

Insight Principal: A coesão textual depende mais da semântica do que da sintaxe. Transformações agressivas nas palavras prejudicam a capacidade do modelo de capturar conexões entre ideias.

Resultados - Item B

Comparação de Extratores

Performance dos Métodos

Quatro abordagens foram testadas para extração de características:

1. **ExtratorManual** - melhor resultado **(71.3%)**
2. **Embeddings** (word2vec-cbow-50d)**(70.8%)**
3. **TF-IDF (69%)**
4. **CountVectorizer (BoW) (0.67%)**

Features do ExtratorManual

- Número de caracteres
- Número de palavras
- Número de frases
- Contagem de pontuações
- Quantidade de dígitos
- Quebras de linha
- Palavras corretas

O ExtratorManual superou os demais por utilizar features específicas relacionadas à coesão, como pontuação concisa que facilita a compreensão.

Resultados - Item C

Stemming vs Lemmatização: Impacto na Coesão

1

Stemming

Corta as palavras de forma agressiva, reduzindo-as à sua raiz. Esta abordagem pode resultar em uma perda semântica significativa, prejudicando a identificação da coesão textual.

Embora simples, sua acurácia foi menor devido à simplificação excessiva.

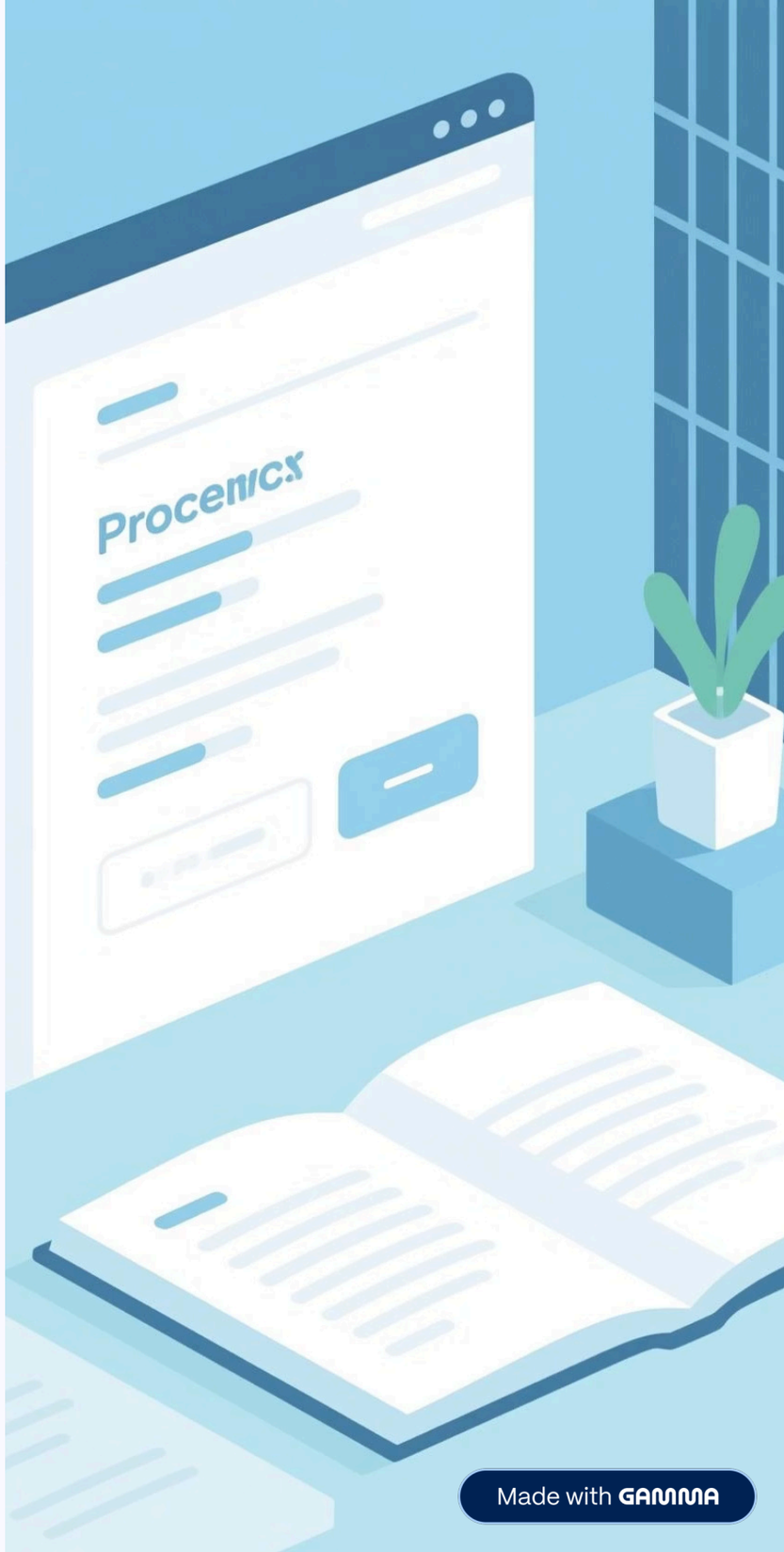
2

Lemmatização

Preserva melhor o significado das palavras, convertendo-as para sua forma canônica (lema) no infinitivo. Isso garante uma compreensão mais precisa do contexto.

Apresentou um desempenho levemente superior na classificação em comparação ao Stemming.

Ambas as técnicas foram testadas utilizando o **ExtratorManual**, o extrator que demonstrou o melhor desempenho no item anterior, para avaliar seu impacto na acurácia da classificação.



Conclusões e Limitações

Acurácia Atual: 71%

Resultado **abaixo do esperado** considerando a importância da coesão textual. Pode prejudicar a avaliação dos estudantes.

Limitação do Modelo

Apenas Regressão Logística foi testada. Outros algoritmos podem oferecer melhor performance.

Falta de Otimização

Hiperparâmetros dos extratores e do modelo não foram refinados sistematicamente.

Questões dos Dados

Dataset com 1.235 amostras pode ser insuficiente para capturar toda a complexidade da coesão.

Próximos Passos

Combinações de Pré-processamento

Testar combinações específicas, como stopwords + maiúsculas ou stopwords + símbolos

Expandir ExtratorManual

Adicionar contagem de parágrafos [P] e analisar importância das features via coeficientes

Combinar Extratores

Unir features de TF-IDF com ExtratorManual e usar outras competências como preditoras

Explorar LLMs

Testar modelos de linguagem modernos para classificação mais sofisticada

Sistema Web

Criar plataforma de upload de redações com correção automática em tempo real, utilizando todas as competências