

Projeto:

Automated Essay Scoring (AES) para o modelo ENEM



Processamento de Linguagem Natural
Prof. Dr. Tiago Almeida

Cinthia Costa
Laura Seto
Leonardo Oliveira
Raul Komai
Renan Almeida



Relembrando...

Sobre a proposta do projeto



Automated Essay Scoring (AES) + ENEM



- Avaliar um texto de entrada com base nas competências determinadas pelo INEP para o modelo ENEM
- Como cada competência aborda diferentes aspectos do texto e reflete problemas particulares de PLN, diversas técnicas podem ser exploradas de forma modular



Essay-BR



- 4.570 redações produzidas por alunos brasileiros do ensino médio (de 2015 a 2020) sobre 86 propostas de tema
- Anotações/avaliações feitas por especialistas humanos
- Sistema de pontos em conformidade com o ENEM
- Estrutura das amostras: *prompt*, *score*, *title*, *essay content*, C1, C2, C3, C4 e C5.



Abordagens e resultados

Estratégias aplicadas para cada competência



Competência I



- A. **Ortografia:** Análise ortográfica calculando a similaridade entre as palavras do texto com um dicionário
- B. **Gramática:** Análise gramatical utilizando modelo baseado em regras
- **Pontuação:** Mapeamento das saídas A e B + normalização das notas



Competência I: pontuação

- ◉ Nota calculada = média entre as duas análises e normalização conforme as notas do ENEM

→ Acurácia: 37,46%



Competência II: extração de *features*



A. Adequação ao gênero dissertativo-argumentativo:

1. Rótulos: pontuações de $c2 \geq 160$ determina a adequação do texto ao gênero;
2. *Features*: redação, número de parágrafos e tamanho médio de cada parágrafo;
3. Modelo para cálculo de probabilidade: Regressão Logística.

→ **Correlação probabilidade e $c2$: 0.552**



Competência II: extração de *features*



B. Cobertura do recorte temático:

1. Identificação de palavras chaves da redação e do *prompt* com TF-IDF;
2. Cálculo de *embeddings* com o modelo 'paraphrase-multilingual-MiniLM-L12-v2' do SentenceTransformer;
3. Cálculo da similaridade do cosseno entre as *embeddings*.

→ **Correlação similaridade e c2: 0.156**



Competência II: extração de *features*



C. Uso de repertório sociocultural:

1. REN com pt_core_news_md do spaCy;
2. REN com repertório personalizado;
3. Contabilização das entidades por redação.

→ **Correlação número de entidades reconhecidas e c2: 0.306**



Competência II: pontuação c2



1. *Features* extraídas:
 - a. Similaridade;
 - b. Probabilidade de adequação ao gênero;
 - c. Número de entidades reconhecidas.
2. Modelo para classificação: Regressão Logística



Competência II: desempenho



	Distribuição (%)
0	1.86
40	1.26
80	14.31
120	37.34
160	36.75
200	8.45

	Precision	Recall	F1-Score
0	0.00	0.00	0.00
40	0.00	0.00	0.00
80	0.64	0.11	0.19
120	0.49	0.70	0.58
160	0.59	0.71	0.65
200	0.00	0.00	0.00

Acurácia: 0.5401



Competência III: Features: Anatomia da Redação



- A. **Identificação dos argumentos:** heurísticas
- B. **Similaridade:** entre os diferentes componentes do texto e proposta
- C. **Árvore Sintática:** profundidade
- D. **POS Tagging, Palavras e Entidades :** quantidade e diversidade

A Redação foi traduzida em 46 features, mas correlação entre as features e C3 foi de baixa a moderada.

MAX -> 0.4413

MIN -> 0.0071



Competência III: Modelo Baseado em Regras



- A. **Lógica:** Criar um "perfil" para cada tipo de nota, usando intervalo 20%-80%
- B. **Classificação:** encontrar perfil que mais se encaixa com a amostra que deve ser classificada
- **Pontuação:** a amostra recebe a nota do perfil com quem mais teve correspondências



Competência III: Modelos de Machine Learning



- A. **Modelos:** Regressão Linear, SVM, Random Forest, Gradient Boosting, XGboost
- B. **Cenário:**
 - 1. **Grupo 1:** todas as 46 features
 - 2. **Grupo 2:** 15 features com maior correlação



Competência III: O ponto que chegamos



A. Cenário 1 (46 Features):

1. **Acurácia:** -54% (ML) vs. 27% (MR).
2. **F1-Score:** 25-32% (ML) vs. 21.5% (MR).
3. **QWK (Concordância):** 0.32-0.38 (ML) vs. 0.30 (MR) -> Performance similar nesta métrica.

B. Cenário 2 (15 Features):

1. Acurácia do ML **cai** (para -35%), mas...
2. **F1-Score** do ML **melhora**, subindo para 28-35%.
3. **QWK** do ML **aumenta**, subindo para 0.48-0.58.
4. **Modelo de Regras:** Permanece estável (QWK de 0.39, F1 de 21.97%).



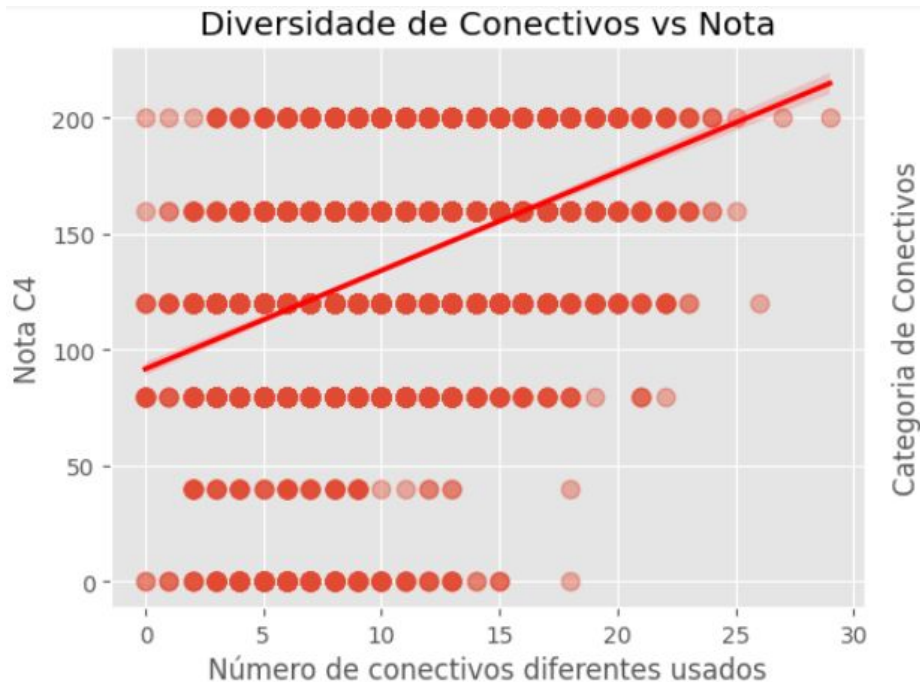
Competência IV



- A. **Análise** nos dados (redações do *dataset*) e **correlações** entre cada estratégia e sua pontuação:
1. Declaração: **lista de conectivos** e suas **categorias**.
 2. Investigação: **correlação** entre a **quantidade** e **diversidade** de conectivo e a nota: Segmentação em (tokens) e identificação dos conectivos.
 3. Correlação entre o uso de conectivos e as notas.
 - Variedade: **0,39** – Quantidade: **0,32**
 - Forte aumento em conectivos de **conclusão** (1,8 – 3,8).



Competência IV



- A. Gráfico diversidade conectivos.
- B. **Outra análise:** Adequação conectivos com sugestões (BERT).
1. Ao realizar esse processamento, a correlação encontrada é de 0.06, sendo **irrelevante**.



Competência IV



- A. Análise: **repetição** de palavras, elementos **referenciais** e **tamanho** dos parágrafos.
 - 1. Correlação de **-0,21** (repetição local) e **0,1** (mecanismos de retomada de informações).
 - 2. Tamanho dos parágrafos: **0,2**.
- B. Com base nessas métricas, cria-se um **avaliador** que quantifica as características que a análise revelou serem mais correlacionadas com as notas (sistema de pontuação).



Competência IV

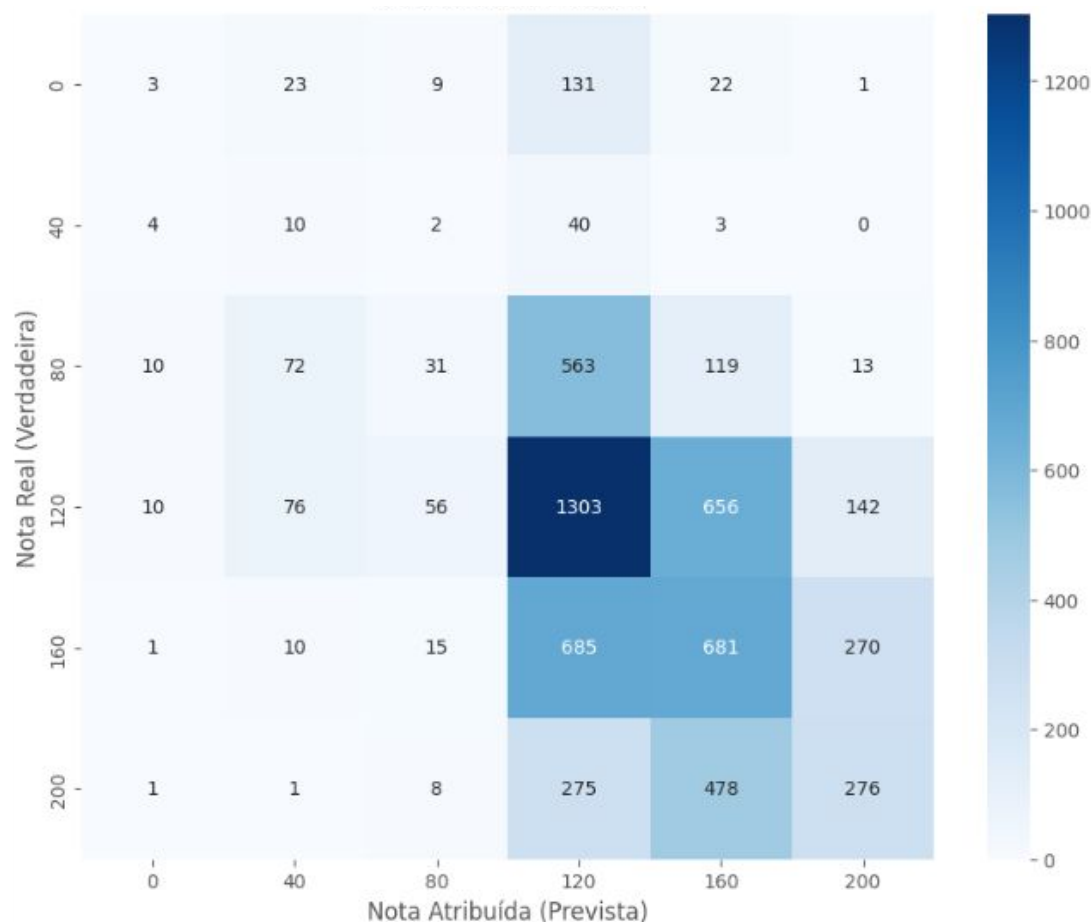
A. Métricas da pontuação:

1. Diversidade de Conectivos.
2. Tamanho dos Parágrafos.
3. Riqueza Vocabular.
4. Coesão Referencial.
5. Quantidade de Conectivos.

Acertos: 2304 de 6000

Taxa de Acerto (nota exata): 38.40%

Margem de 1 nível de nota (± 40 pontos): 85.22%





Competência IV



A. Classificador (ML):

1. Características **extraídas**: Segmentação. Para cada segmento:

- Densidade e diversidade de conectivos.
- Taxa de repetição de palavras.
- Contagem de palavras por parágrafo.
- Similaridade de significado entre os parágrafos.

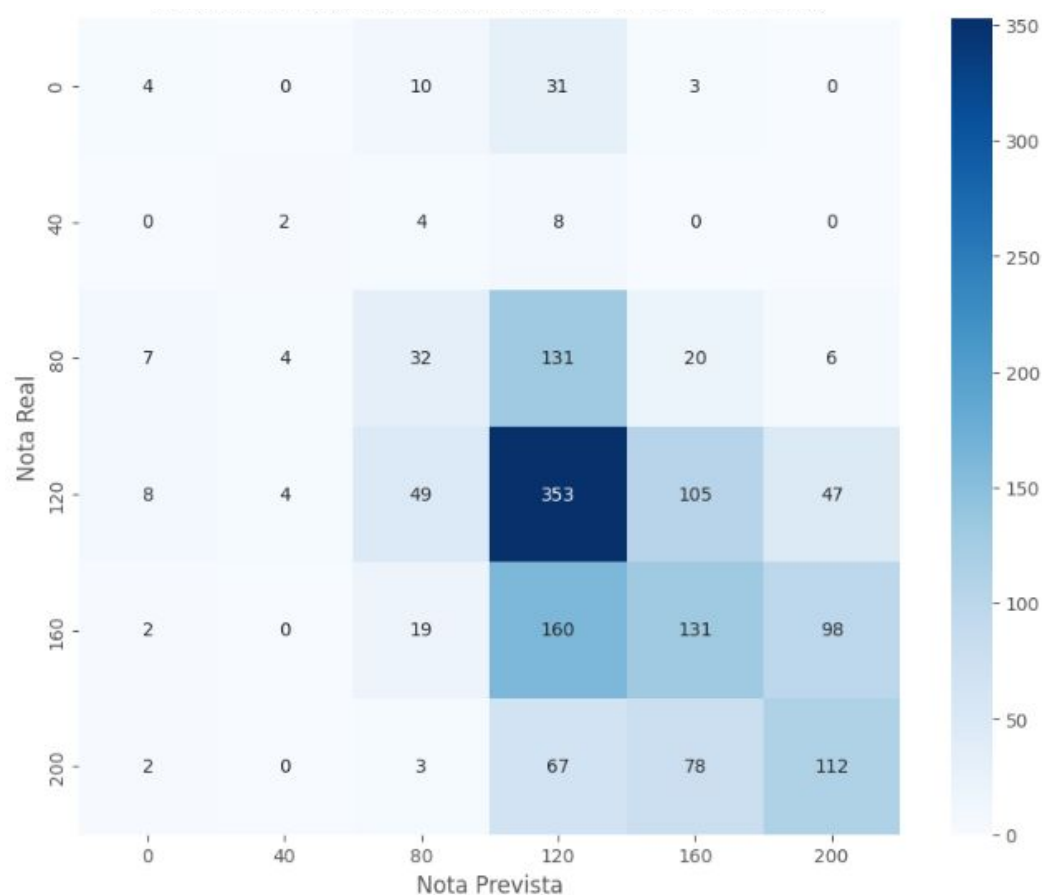
B. **LightGBM** (Light Gradient Boosting Machine): modelo baseado em árvores de decisão.

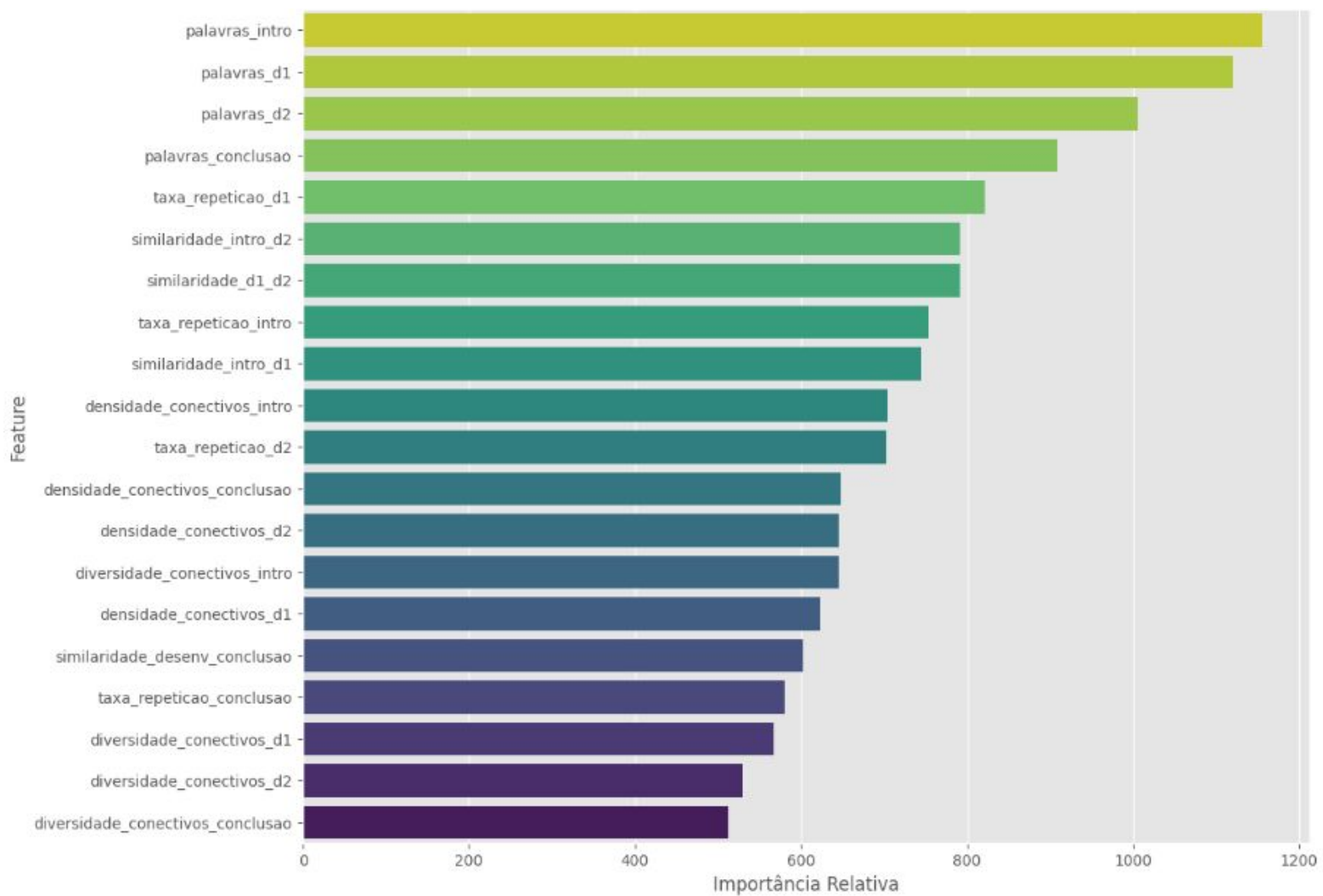
C. Analisa milhares de redações e suas notas, aprendendo quais combinações de features são os melhores indicadores para cada nível de nota.



Competência IV

- A. Acurácia: **42.27%**
- B. Extrema dificuldade com as notas mais raras, especialmente 0 e 40.
- C. Dados extremadamente desbalanceados.







Competência V



A. Proposta de Intervenção

Ação

o que será feito?

Agente

quem fará?

Meio

como será feito?

Finalidade

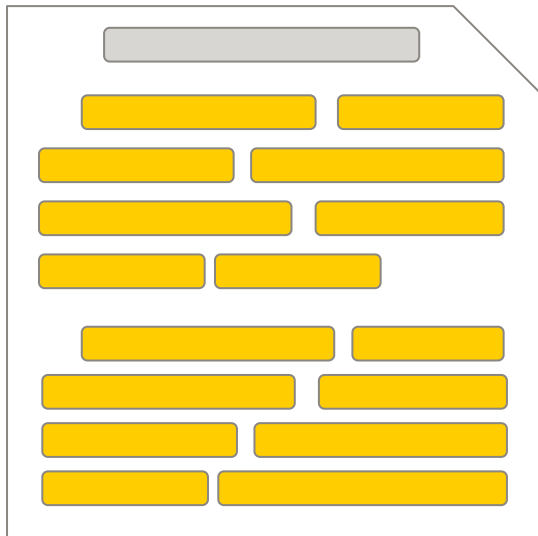
qual o objetivo a ser alcançado?




Competência V



A. Análise sobre o texto completo



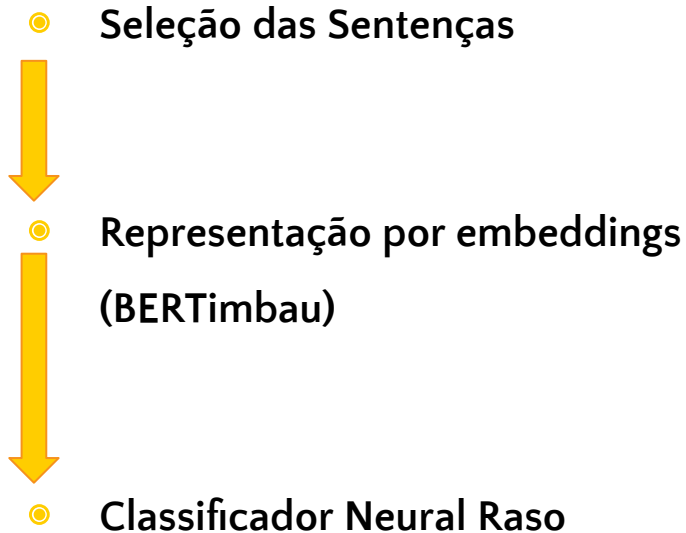
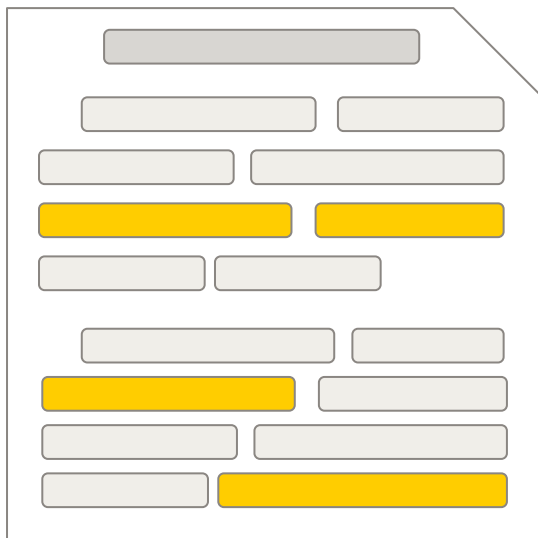
- Representação por embeddings (BERTimbau)
- 
- Classificador Neural Raso



Competência V



A. Análise sobre sentenças relevantes

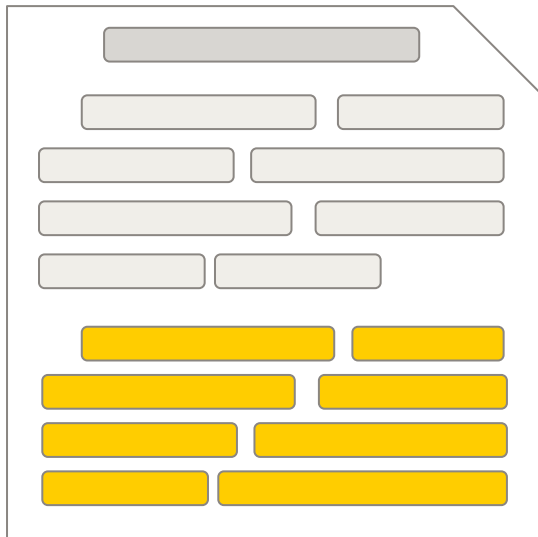




Competência V



A. Análise sobre paragrafos relevantes



Seleção dos Parágrafos



Representação por embeddings
(BERTimbau)



Classificador Neural Raso



Competência V



A. Desempenho Obtido

Abordagem	Acuracia	Quadratic Weighted Kappa
Texto Completo	48%	30%
Sentenças Relevantes	45%	22%
Parágrafos Relevantes	45%	25%



Considerações Finais



Desafios e Limitações do Modelo

- A. **Baixo desempenho:** Modelos não superam significativamente a predição da classe majoritária.
- B. **Features fracas:** Correlações fracas com o atributo alvo dificultam a inferência.
- C. **Desbalanceamento:** Prejudica a performance em classes minoritárias (precisão e recall = 0 em muitos casos).
- D. **Ruído nos dados:** Presença de amostras inválidas e anotações incoerentes (ex: outros gêneros textuais com notas altas).



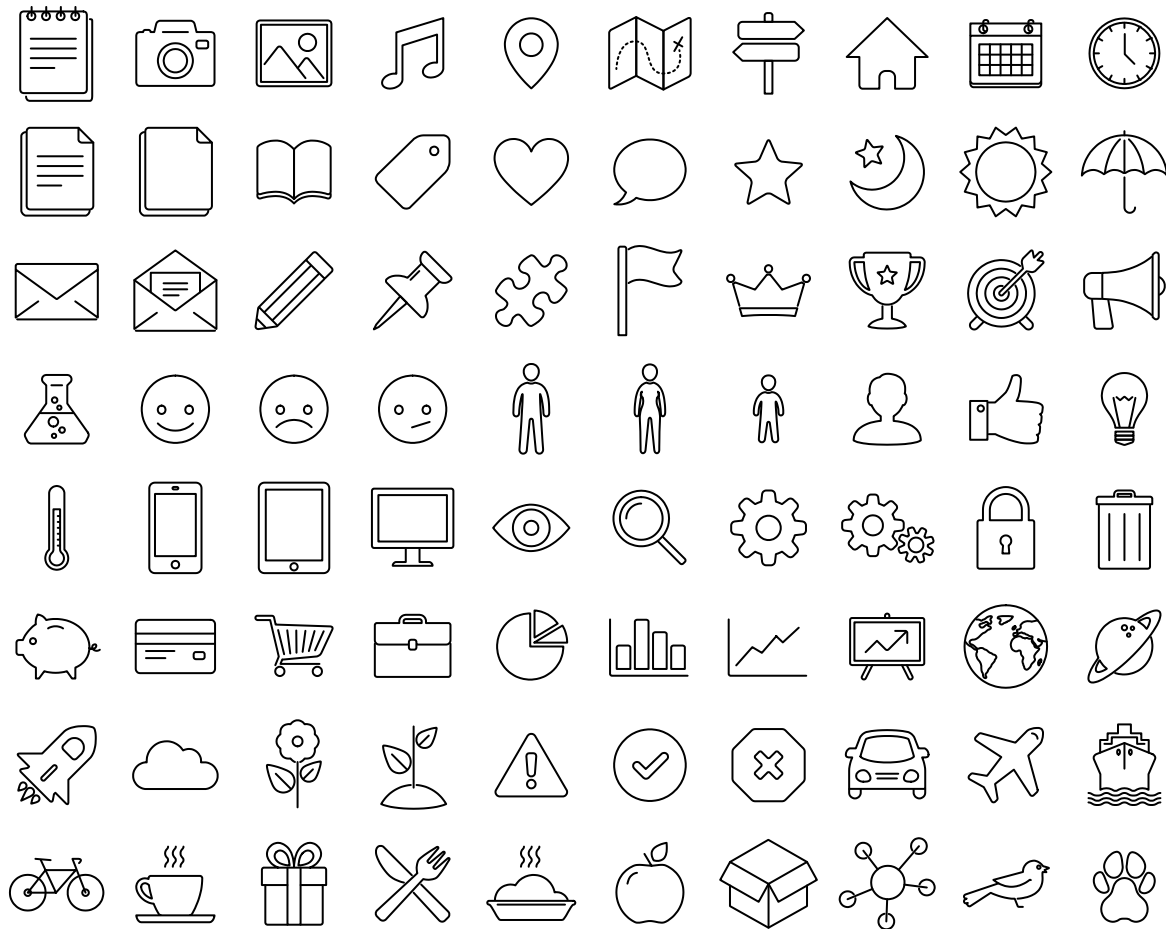
Considerações Finais



O Sonho

Recomendações

- A. **Envolver especialistas** na anotação (ex: marcação de argumentos e tese).
- B. **Melhorar a seleção de features** com base em conhecimento linguístico e pedagógico.
- C. **Aprimorar qualidade dos dados** antes da modelagem.



SlidesCarnival icons are **editable shapes**.

This means that you can:

- Resize them without losing quality.
- Change line color, width and style.

Isn't that nice? :)

Examples:

