

Quality and Emotion Visualization in Text

Principais Características:

- Base de Dados
- Algoritmos
- Heurísticas
- Modelo de Treino
- Aplicação Web

Base de Dados

- A base de dados contém 406 textos;
- Todos os textos têm mais de 5000 palavras e 60 frases.

Estrutura	Muito Simples	Pouco Complexo	Complexo	Muito Complexo
Nº Textos	8	238	138	20

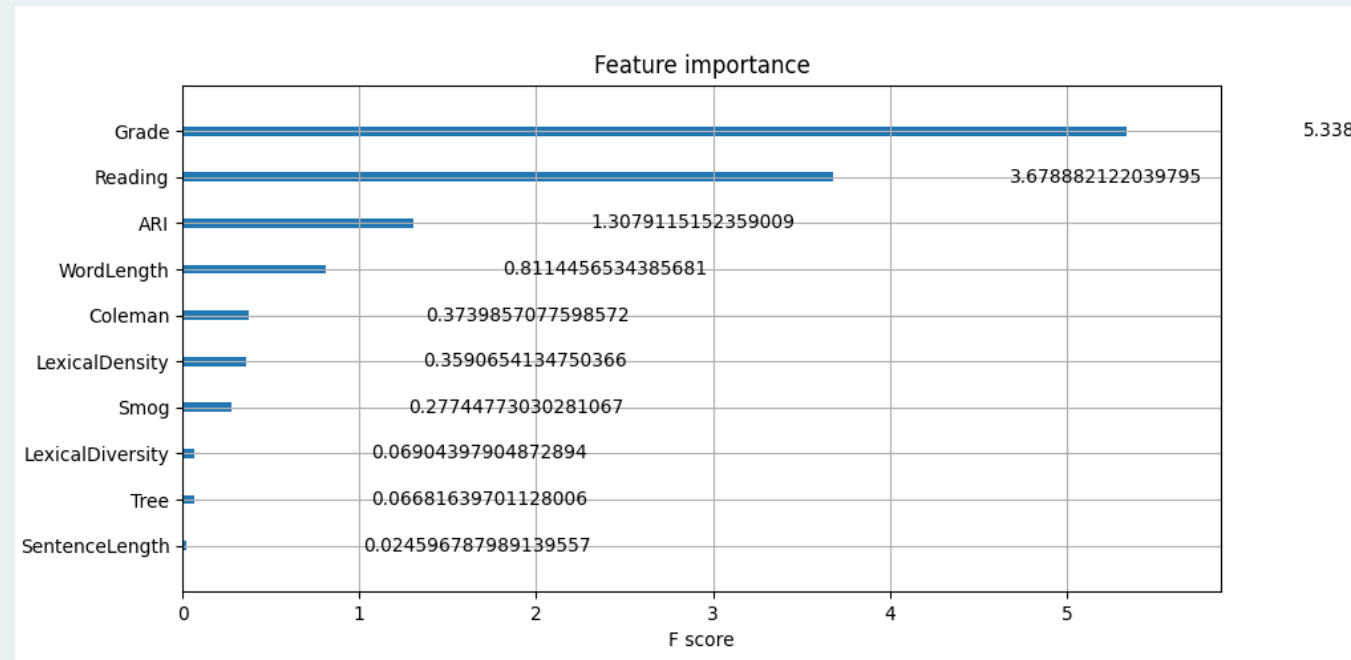
Sentimentos	Negativo	Neutro	Positivo
Nº Textos	51	141	212

Amostras

- **Tipo de Amostras:** São retiradas 60 frases aleatórias do texto;
- **Tipo de Frases:** São apenas retiradas frases que contem mais de 4 palavras;
- **Razão:** As amostras são retiradas desta forma para um funcionamento normal dos algoritmos desenvolvidos;
- **Erro:** Margem de erro comparado com a avaliação total do texto apenas de 1.2% a 10%.
- **Eficiência:** O uso de amostras demonstrou um aumento de performance de 90%.

Algoritmos

- Foram desenvolvidos 11 algoritmos no total;
- Medidas de Complexidade Textual: Diversidade e densidade lexical, tamanho médio das palavras e das frases e profundidade sintática da frase;
- Métricas de Leitura: *SMOG*, *ARI*, *Coleman-Liau Index*, *Flesch Kincaid Grade Level* e *Flesch Reading Ease*;
- Análise de Sentimentos é usado o módulo *Spacy* para essa avaliação.



Heurísticas

Estrutura Textual:

- **Característica comum:** Quanto mais elevado o valor, melhor estruturado se encontra o texto;
- Aplicada normalização nos algoritmos para um resultado mais preciso e melhor manipulação dos dados;
- Aplicado a média de todos os algoritmos utilizados.

Visualização Sentimental:

- É utilizado o *compound* como valor principal para avaliação sentimental do texto.

Heurísticas

Estrutura	Muito Simples	Pouco Complexo	Complexo	Muito Complexo
Intervalos:	[0, 0.25[[0.25, 0.50[[0.50, 0.75[[0.75, 1]

Sentimentos	Negativo	Neutro	Positivo
Intervalos:	[-1, -0.05[[-0.05, 0.05]]0.05, 1]

Modelo de Treino XGBoost

- Funções relevantes para o trabalho desenvolvido como: *plot importance*;
- Modelo mais preciso e eficiente comparado com modelos testados, árvore de decisão e regressão linear.

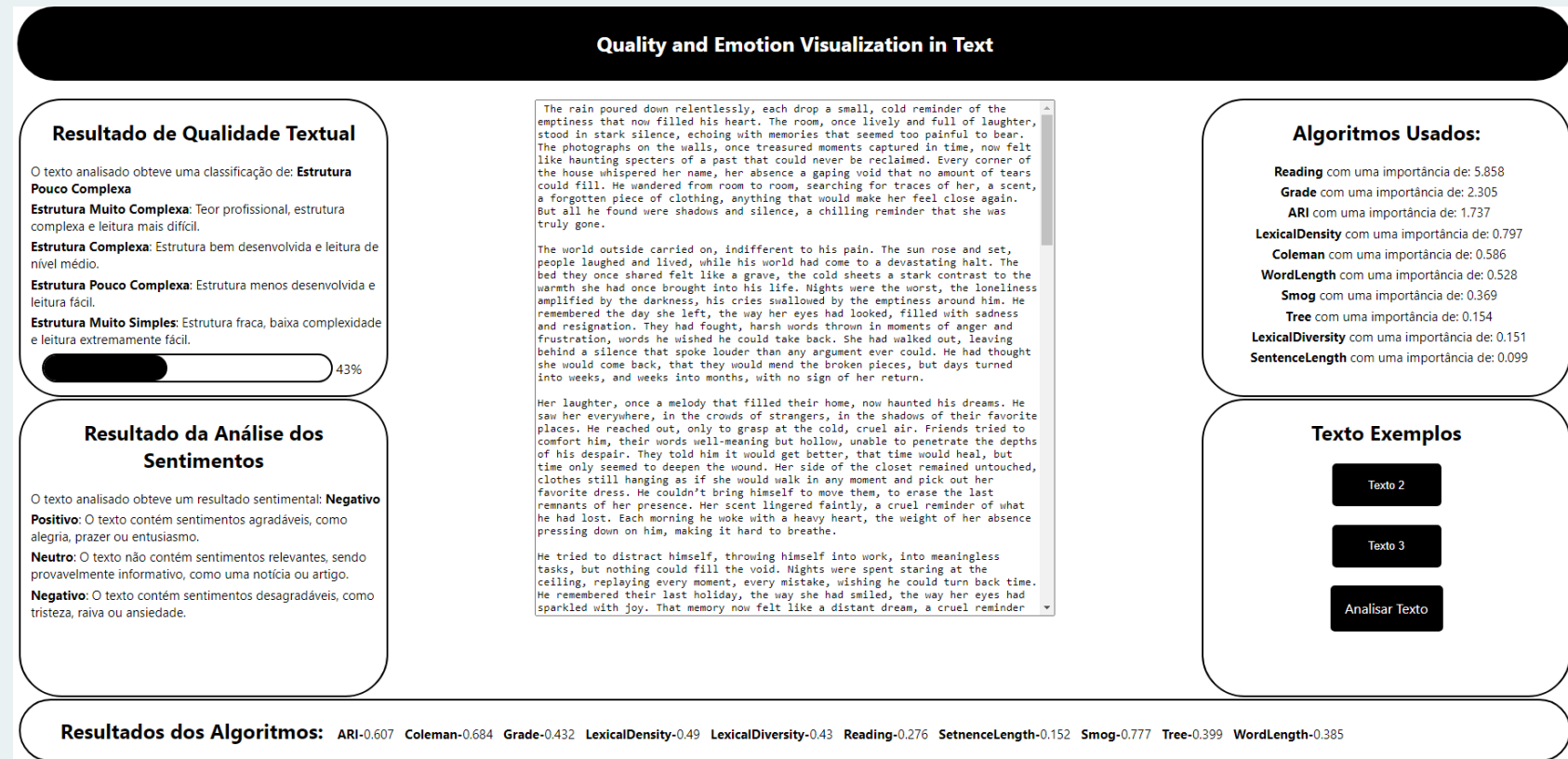
Modelos	Regressão Linear	Árvore de Decisão	XGBoost
Taxa de Acerto(%) Estrutura	73	93	96
Taxa de Acerto(%) Sentimento	43	98	99

Modelo de Treino XGBoost

- Usado para treino a percentagens comuns de 30% para testes e 70% para treino;
- A média dos testes foi obtida através da media feita a partir de 1000 treinos diferentes.

Classificadores	F1 Score	Recall	Taxa de Acerto	Precisão
Percentagem(%) Estrutura	96.02	96.11	96.11	96.27
Percentagem(%) Sentimento	99.1	99.11	99.1	99.13

Aplicação Web



Aplicação Web

Resultado de Qualidade Textual

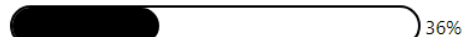
O texto analisado obteve uma classificação de: **Estrutura Pouco Complexa**

Estrutura Muito Complexa: Teor profissional, estrutura complexa e leitura mais difícil.

Estrutura Complexa: Estrutura bem desenvolvida e leitura de nível médio.

Estrutura Pouco Complexa: Estrutura menos desenvolvida e leitura fácil.

Estrutura Muito Simples: Estrutura fraca, baixa complexidade e leitura extremamente fácil.



Resultado da Análise dos Sentimentos

O texto analisado obteve um resultado sentimental: **Negativo**

Positivo: O texto contém sentimentos agradáveis, como alegria, prazer ou entusiasmo.

Neutro: O texto não contém sentimentos relevantes, sendo provavelmente informativo, como uma notícia ou artigo.

Negativo: O texto contém sentimentos desagradáveis, como tristeza, raiva ou ansiedade.

Zona responsável por mostrar de forma listada os algoritmos utilizados

Zona responsável por mostrar os resultados obtidos e explicá-los

Algoritmos Usados:

Reading com uma importância de: 5.858

Grade com uma importância de: 2.305

ARI com uma importância de: 1.737

LexicalDensity com uma importância de: 0.797

Coleman com uma importância de: 0.586

WordLength com uma importância de: 0.528

Smog com uma importância de: 0.369

Tree com uma importância de: 0.154

LexicalDiversity com uma importância de: 0.151

SentenceLength com uma importância de: 0.099

Zona responsável por mostrar os resultados normalizados dos algoritmos

Resultados dos Algoritmos: ARI-0.636 Coleman-0.679 Grade-0.461 LexicalDensity-0.497 LexicalDiversity-0.421 Reading-0.296 SetnenceLength-0.162 Smog-0.778 Tree-0.411 WordLength-0.38

Conclusões Finais

- Importância de bons algoritmos de avaliação estrutural e sentimental;
- Importância de boas heurísticas;
- Importância de um bom modelo de treino;
- Existência de uma relação entre todos os algoritmos usados.

Obrigado