Relatório: Classificação Multirrótulo

Base de Dados

Jigsaw Toxic Comment Classification Challenge

> Link: https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge/data

Objetivo

Construir um classificador multirrótulo para prever se um comentário é:

- toxic
- severe_toxic
- obscene
- threat
- insult
- identity_hate

Cada comentário pode ter **mais de uma dessas categorias** atribuídas.

Etapas realizadas

Pré-processamento dos Dados

- Remoção de pontuação, números e símbolos.
- Conversão para minúsculas.
- Tokenização e remoção de stopwords (palavras irrelevantes).
- Transformação dos textos usando vetores **TF-IDF** (Term Frequency-Inverse Document Frequency) com limite de 10.000 features.

Modelagem

- Biblioteca utilizada: **Scikit-Multilearn**, especializada em problemas multirrótulo.
- Estratégia: **Binary Relevance**, que transforma o problema multirrótulo em vários classificadores binários independentes.
- Classificador base: **Logistic Regression** (Regressão Logística) com 1000 iterações para garantir a convergência.

Modelos treinados

Modelo 1: Sem Balanceamento

Treinado diretamente no dataset original.

Modelo 2: Com Balanceamento:

Aplicado **oversampling** nas classes minoritárias:

- As instâncias das categorias menos representadas foram replicadas.
- Cada label foi balanceada individualmente e os datasets foram combinados

Avaliação:

Categoria	F1-Score (Sem Balanceamento)	F1-Score (Com Balanceamento)
toxic	0.68	**0.73**
severe_toxic	0.36	**0.42**
obscene	0.68	**0.73**
threat	0.26	**0.31**
insult	0.61	**0.66**
identity_hate	0.35	**0.40**

- Métricas Globais:

Métrica	Sem Balanceamento	Com Balanceamento
Subset Accuracy	89,5%	**91,9%**
Hamming Loss	0.025	**0.019**
F1-Score (Micro)	0.64	**0.67**
F1-Score (Macro)	0.49	**0.50**

Análise detalhada:

- O modelo **balanceado** obteve ganhos claros, especialmente nas classes minoritárias:
  - **threat (F1 subiu de 0.26 para 0.31)**
  - **identity_hate (0.35 → 0.40)**
  - **severe_toxic (0.36 → 0.42)**
- As classes mais representadas também tiveram ganhos:
  - **toxic (0.68 → 0.73)**
  - **obscene (0.68 → 0.73)**
  - **insult (0.61 → 0.66)**
- A **Hamming Loss** diminuiu (de 0.025 para 0.019), indicando menos erros médios por rótulo, uma melhoria importante.

- **Subset Accuracy** subiu (de 89,5% para 91,9%), mostrando que o balanceamento ajudou na classificação correta dos conjuntos completos de rótulos.
- **F1-Score Macro** e **Micro** também melhoraram, refletindo um modelo mais equilibrado e eficiente, especialmente na consideração das classes menos frequentes.

Conclusão:

- O modelo sem balanceamento teve dificuldades nas classes minoritárias, entregando um desempenho razoável nas categorias mais comuns, mas falhando nas menos representadas.
- Após aplicar o **oversampling**, houve uma **melhoria consistente tanto nas classes majoritárias quanto nas minoritárias**, além de uma redução dos erros médios (Hamming Loss).
- O uso de balanceamento demonstrou ser altamente benéfico neste cenário de **classificação multirrótulo com dados desbalanceados**