CTir v0.1

Classificação Textual

em documentos com relevante

Incidência de Ruído

Projeto didático apresentado no Curso:

Inteligência Artificial na prática: Machine Learning

ESMPU - Maio/2023

Professores:
Erick Muzart
Fernando Melo

Tutoria:
Thiago Vieira

Alunos:
Christiano Maia
Denard Soares

Roteiro

1. Desafio
2. Solução Proposta
3. Dados: pré-processamento e Ruído
4. Modelo: Treinamento e Métricas de performance
5. Abordagem: pontos de corte
6. Abordagem: margem máxima
7. Resultados
8. Publicação e próximos passos

1. Desafio

- Classificação de pronunciamentos judiciais a partir de seu texto.
  - Documentos padronizados (mesma origem e formato);
  - Documentos não padronizados:
    - Elevada variabilidade estrutural;
    - **Relevante incidência de ruído**
      - **Erros no reconhecimento óptico de caracteres (OCR);**
      - **Erros de digitação (como substituição ou exclusão aleatória de caracteres em palavras).**

Problema a ser solucionado

Recuperação de acurácia em classificação de textos com alta incidência de ruído

2. Solução Proposta

- Treinamento supervisionado de modelo de classificação a partir de texto com relevante incidência de ruído.
    - Variável alvo (target): categoria/classe do texto.
        - Dados já classificados/rotulados;
        - **Texto COM ruído.**

3. Dados: pré-processamento e Ruído

- Fonte de dados: *<Vide README.MD> do projeto no GitHub*

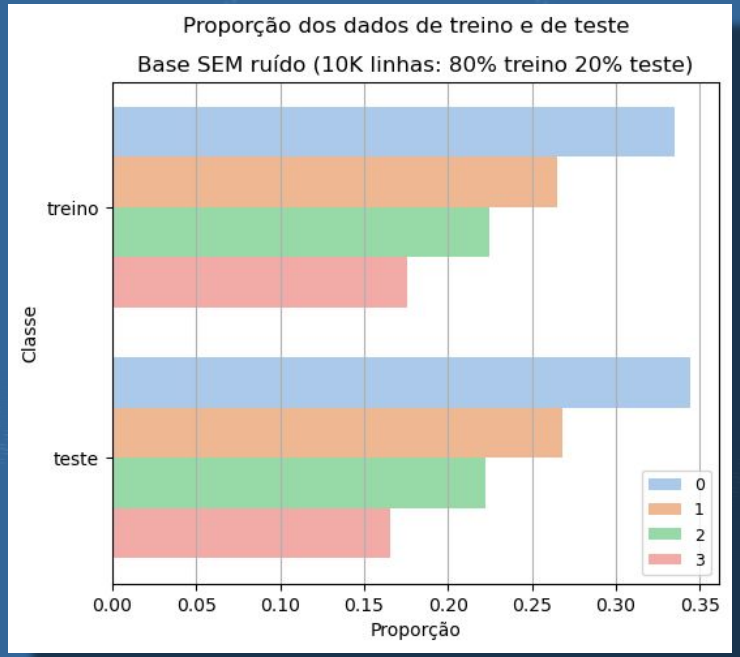
Informações técnicas:

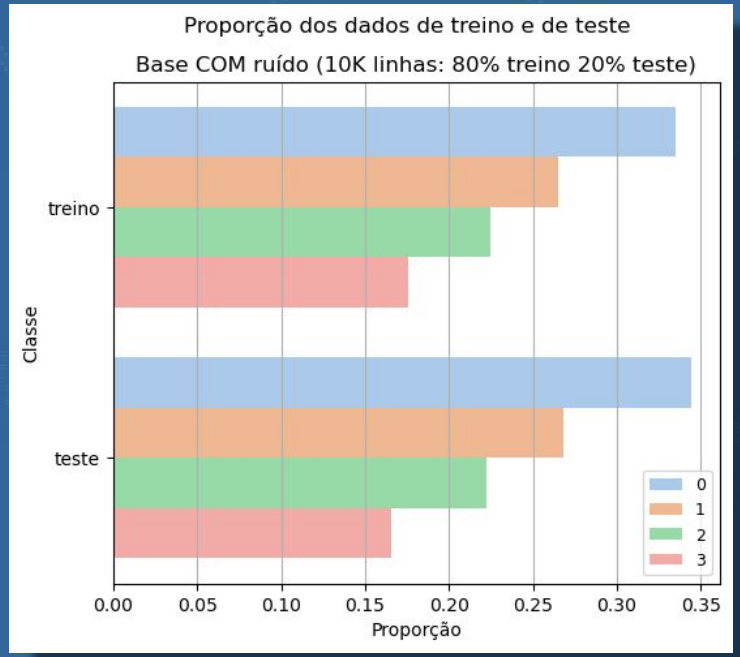
- *Remoção de Stopwords após a inserção do ruído:*
  - *algumas palavras com ruído deixam de ser stopwords e, portanto, não são removidas no pré-processamento.*
- *Bibliotecas python utilizadas para:*
  - *remoção de stopwords:* ***nltk***
  - *inserção de ruído aleatório:* ***nlpaug***

4. Modelo: Treinamento e Métricas de performance

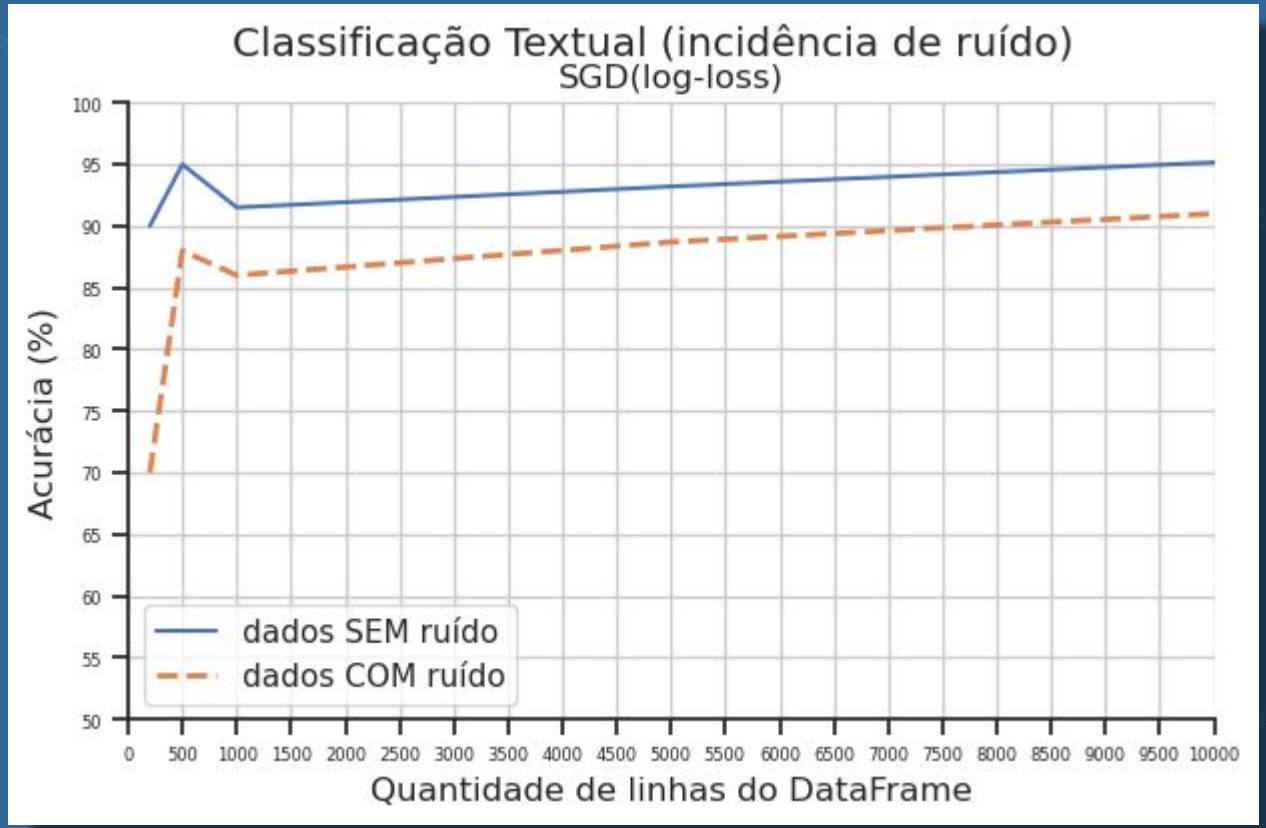
- Trade-off SGD classifier X Random Forest
- Modelo de referência:
  - SGDClassifier(loss='log_loss')
    - Classificação linear com treinamento SGD.
      - Regressão logística
- Pontos de corte do Dataset = 200, 500, 1000, 5.000 e 10.000 linhas.
- Acurácias comparadas com a do modelo de referência.
- Matriz de confusão.

Proporção entre as bases de Treino e de Teste





CTir v0.1 - Classificação Textual em documentos com relevante Incidência de Ruído



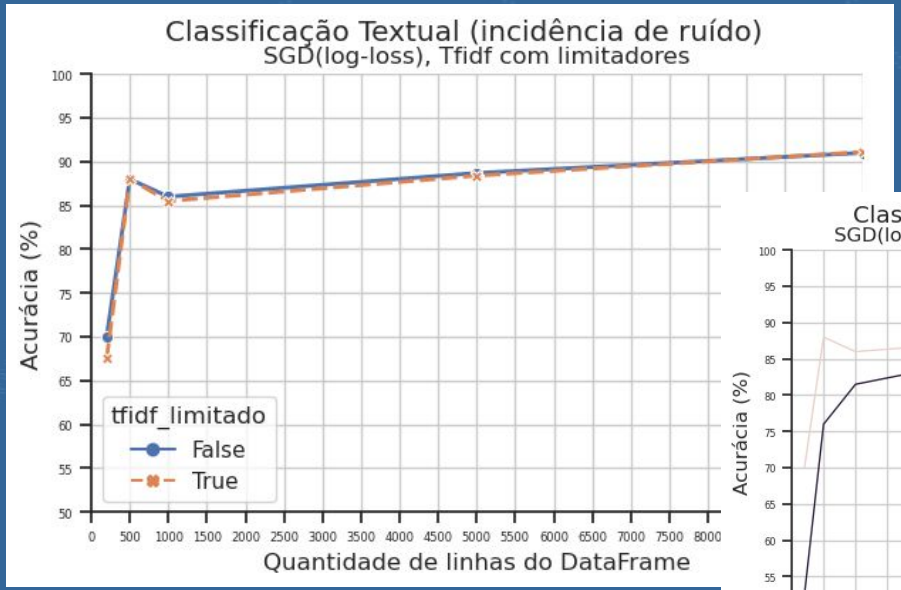
10 K linhas:

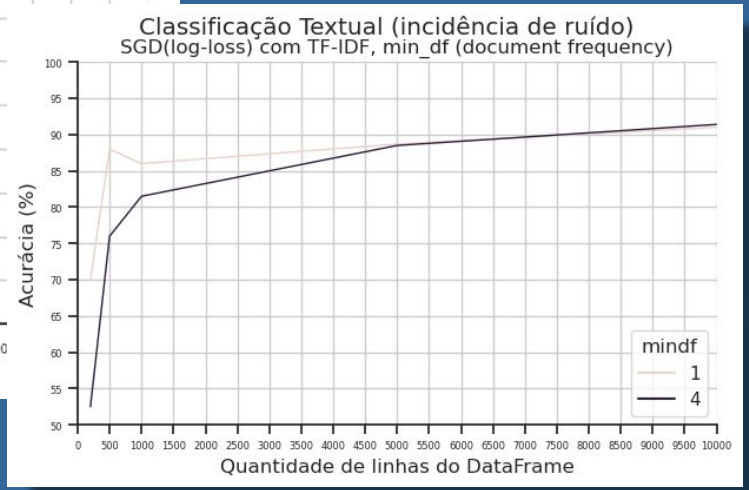
SEM ruído:
95,15 %

COM ruído:
91,00%

5. Abordagem: pontos de corte

- Mínimo DF (document frequency)
- TF-idf limitado





10 K linhas:

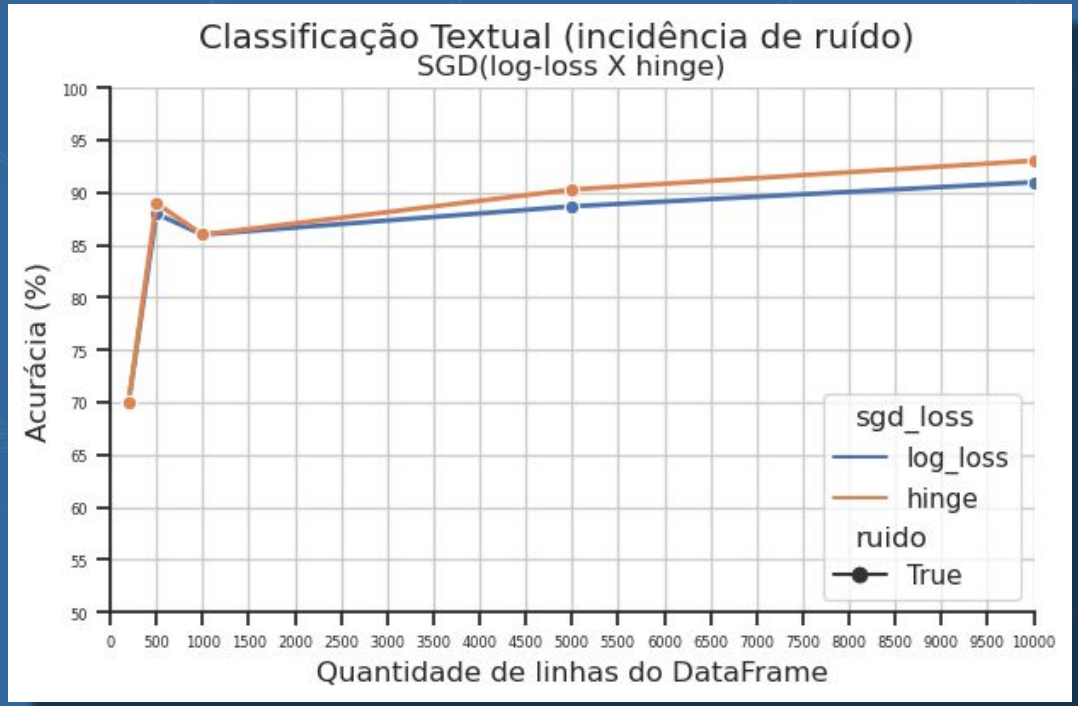
COM ruído: 91,00%

C/ tf-idf limitado: 91,10 %

C/ min_df = 4: 91,40 %

6. Abordagem: margem máxima

- **Hinge**



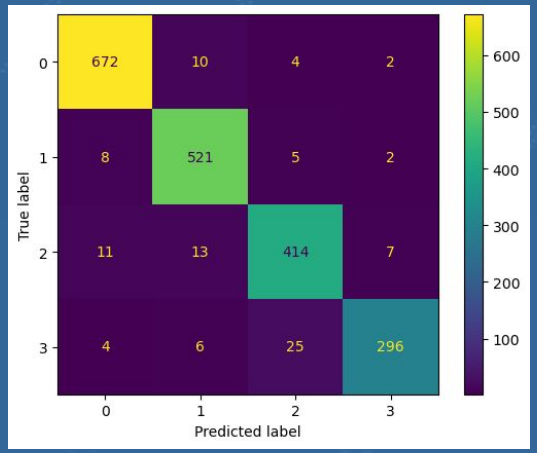
10 K linhas:

LOG-LOSS COM ruído: 91,00 %

HINGE COM ruído: 93,05 %

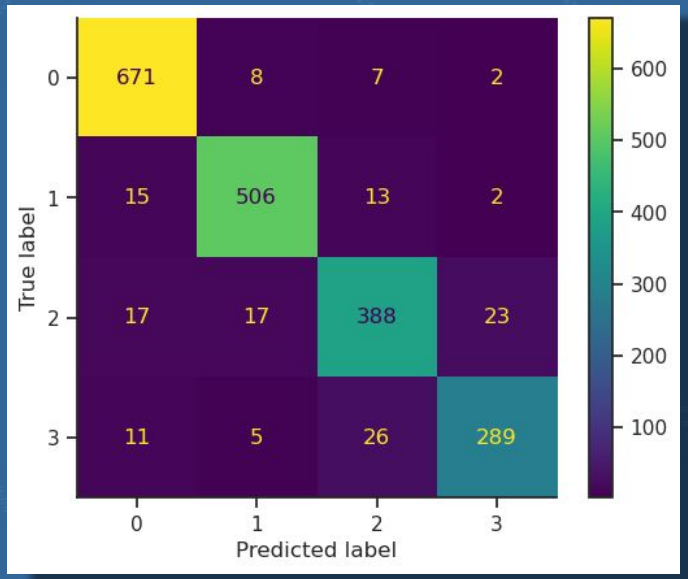
CTir v0.1 - Classificação Textual em documentos com relevante Incidência de Ruído

Modelo de referência
SEM ruído

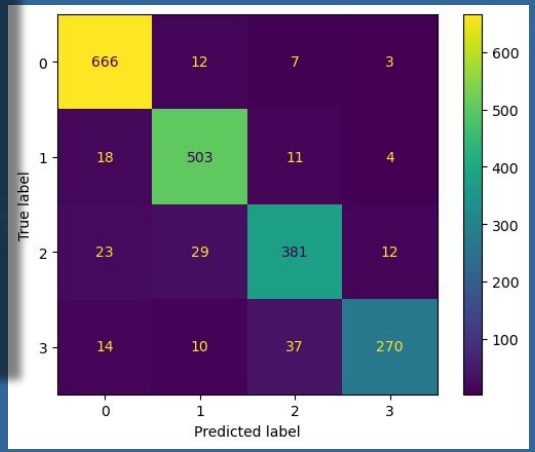


Modelo de referência = SGDClassifier(loss='log_loss')

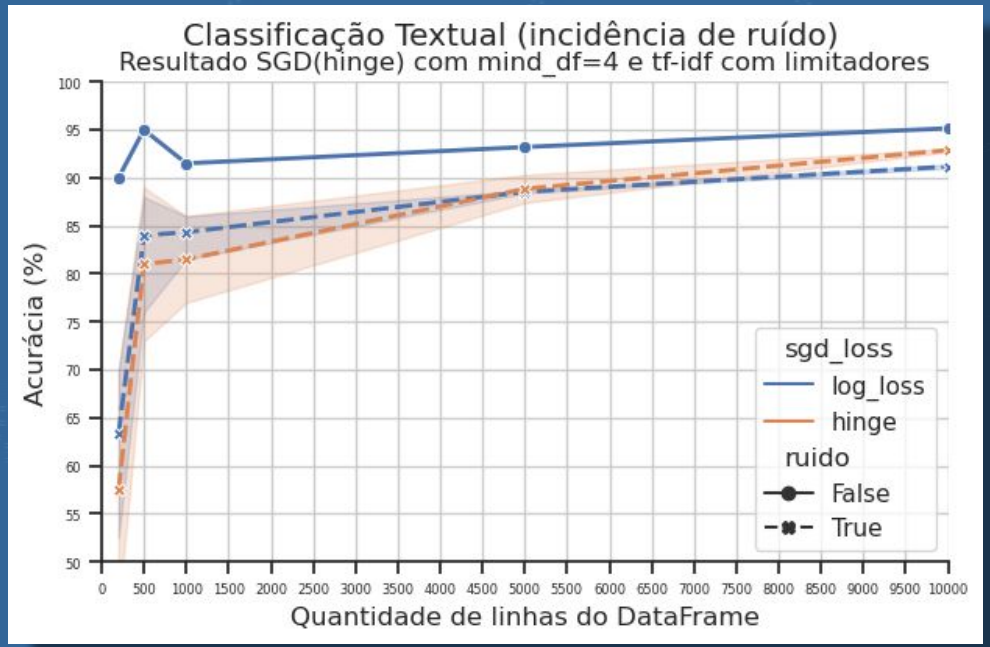
SGDClassifier(loss='hinge') [default]
COM ruído



Modelo de referência
COM ruído



7. Resultados



Interpretação/percepções sobre a acurácia:

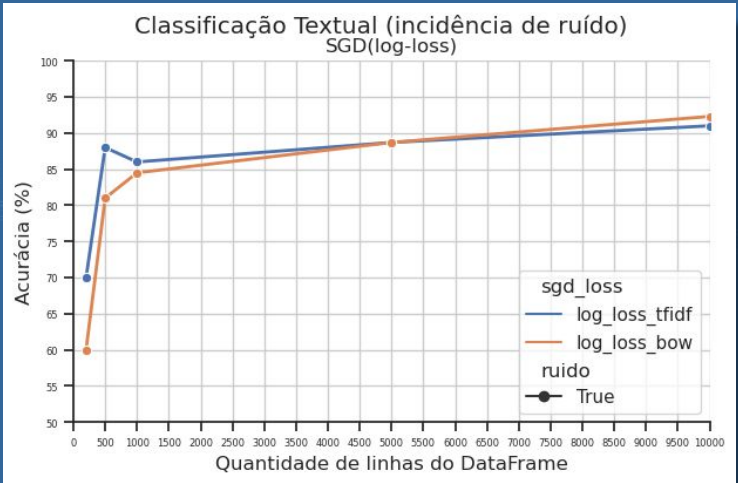
- Queda sensível para dados COM ruído.
- Recuperação progressiva em função do aumento da quantidade de dados.
- Melhoria pouco relevante para métodos de pontos de corte ("cut-off").
- Considerável melhoria para método que maximize a margem de separação entre as classes
  - Hinge
    (margem/fronteira máxima em SVM)

8. Publicação e próximos passos

https://github.com/pro-chsmaia/projeto_cursoml_ctir

- SVMs com otimização de hiperparâmetros
- Pré-processamento com lematização

- Bag of words X Tf-idf (impacto do idf)



CTir - Classificação Textual em documentos com relevante incidência de ruído

STATUS EM DESENVOLVIMENTO FINALIDADE DIDÁTICA

Projeto didático apresentado no Curso:

Inteligência Artificial na prática: Machine Learning / ESMPU - *Maio/2023*

Equipe:

> Christiano Maia
>
> Denard Soares

Dentre as 20 menores médias da quantidade de ocorrências de palavras em cada documento (considerando que há muitos empates), foram identificadas as seguintes quantidades de palavras/expressões que existem no vocabulário para os seguinte tamanhos do conjunto de dados:

- *Para 200 linhas: 9 / 20*
- *Para 1000 linhas: 8 / 20*
- *Para 10.000 linhas: 1 / 20*

CTir v0.1

Classificação Textual

em documentos com relevante

Incidência de Ruído

Final da apresentação