FACULDADE DE ENGENHARIA DE SOROCABA

Trabalho de Conclusão de Curso - Especialização em Ciência de Dados

HATE SPEECHES

Adriano Valério Santos da Silva
Allan Flores de Jesus
Vinícius Targa Gonçalves

Prof. Orientador Fernando Vieira da Silva Msc.

Sorocaba (SP), Outubro de 2020.

“Eu não sei o que quero ser, mas sei muito bem o que não quero me tornar.” -

Friedrich Nietzsche

Agradecimentos

Agradecemos por todos os nossos professores e colegas de equipe participantes deste trabalho, onde se colocaram à disposição de auxiliar no desenvolvimento do mesmo, bem como disponibilidade de participar de todas as possíveis discussões envolvidas. Em especial, aos nossos orientadores e mestres Fernando Vieira da Silva e Johannes Von Lochter - no qual se dedicaram imensuravelmente à proposta do projeto.

Pelos esclarecimentos relacionados ao dataset escolhido (bem como técnicas utilizadas) - Rogers Prates de Pelle, criador e mestre na Universidade Estadual do Rio Grande do Sul, que nos ajudou ao entendimento detalhado do contexto.

Introdução

As redes sociais se popularizaram por possibilitar a interligação de pessoas de todos os lugares do mundo, buscando de forma facilitada, a interação social entre elas. Por ser um espaço amplo e diversificado, a dessemelhança entre pessoas se torna mais acentuada gerando conflitos virtuais, estes que são oriundos do mundo real. Como no mundo virtual o contato físico é inexistente, seus usuários se sentem encorajados em expressar suas opiniões sem qualquer limitação ou ponderação, sendo estas opiniões tão diversas quanto seu público, devido a isso, opiniões preconceituosas, discriminatórias e intolerantes se tornaram extremamente comuns.

Problemas identificados

- Identificação de sentimentos em discursos ofensivos / não ofensivos.
- Dificuldades quando exposto à um grande volume de dados.
- Separação de palavras obsoletas.

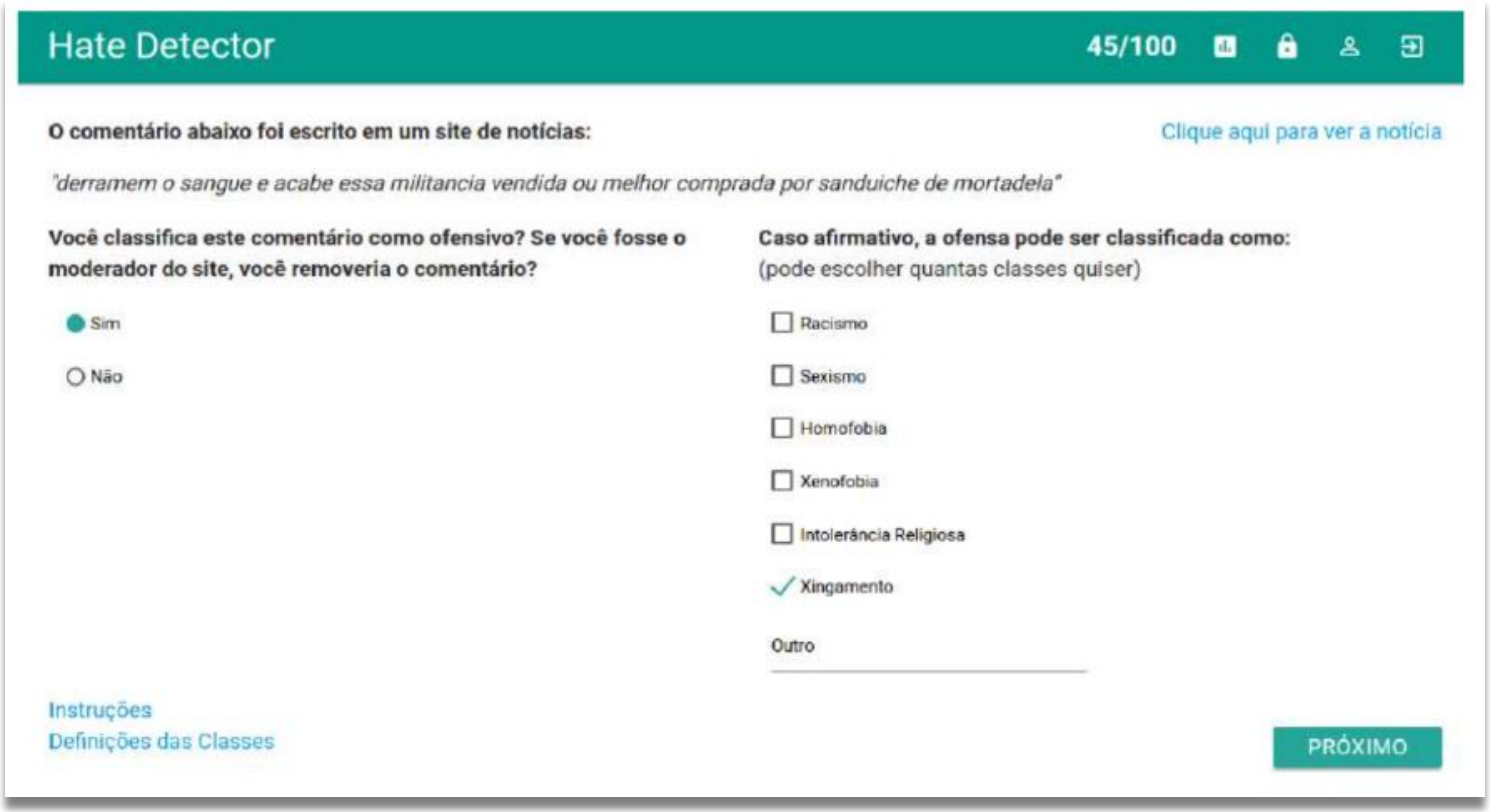
Objetivos

- Aplicação de técnicas de aprendizado de máquina e processamento de linguagem natural.
- Uso de Redes Neurais / Classificadores que auxiliem na identificação de textos ofensivos.
- Deploy de modelos (observando as predições) para serem utilizados de modo externo.
- Comparativos entre as diferentes técnicas aplicadas.

Dataset escolhido

- O dataset utilizado chama-se **OFFCOMBR-2**.
- Utilizado na dissertação de mestrado em Ciência da Computação (UFRGS).
- Dados coletados do Portal Web G1 Notícias (política e esportes).
- Técnicas de web scraper para coleta.
- 10 mil registros, sendo 1.250 utilizados (escolhidos aleatoriamente).
- Classificação por juízes humanos.
- Colunas: *ID, Class, Document.*

Dataset escolhido



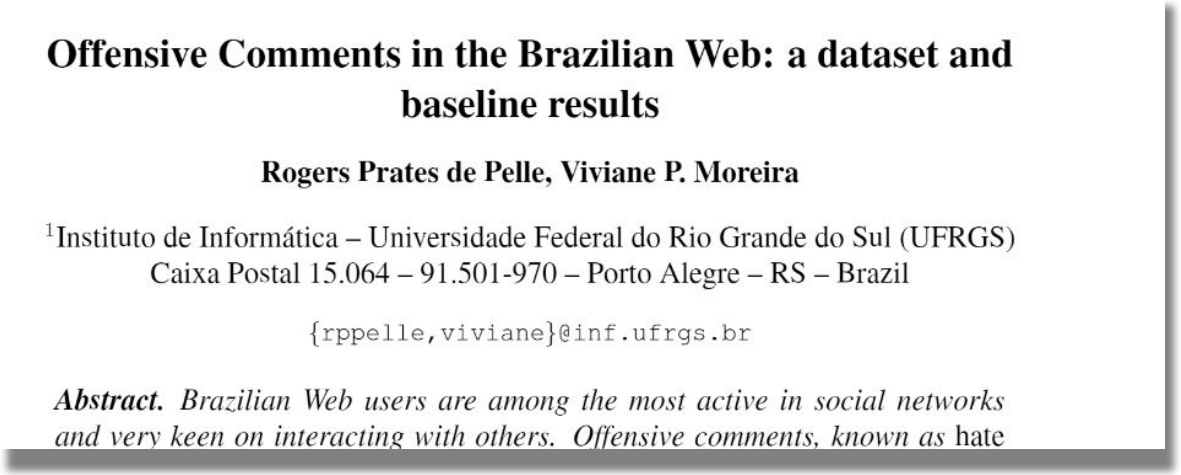
Dataset escolhido

Table 1. Prevalence of each Category in the Annotations

# Judges	Xenophobia	Homophobia	Sexism	Racism	Cursing	Religious Intolerance
1	13 (1,0%)	35 (2.8%)	14 (1,1%)	19 (1.5%)	375 (30.0%)	1 (0.1%)
2	12 (1.0%)	14 (1,1%)	8 (0.6%)	18 (1.4%)	286 (22.9%)	1 (0.1%)
3	5 (0.5%)	9 (0.9%)	4 (0.4%)	1 (0.1%)	175 (16,9%)	0 (0.0%)

Pesquisas relacionadas

- PELLE, ROGERS PRATES DE; MOREIRA, VIVIANE P. **Offensive Comments in the Brazilian Web: a dataset and baseline results**. Porto Alegre: UFRGS.



Offensive Comments in the Brazilian Web: a dataset and baseline results

Rogers Prates de Pelle, Viviane P. Moreira

[1]Instituto de Informática – Universidade Federal do Rio Grande do Sul (UFRGS)
Caixa Postal 15.064 – 91.501-970 – Porto Alegre – RS – Brazil

{rppelle,viviane}@inf.ufrgs.br

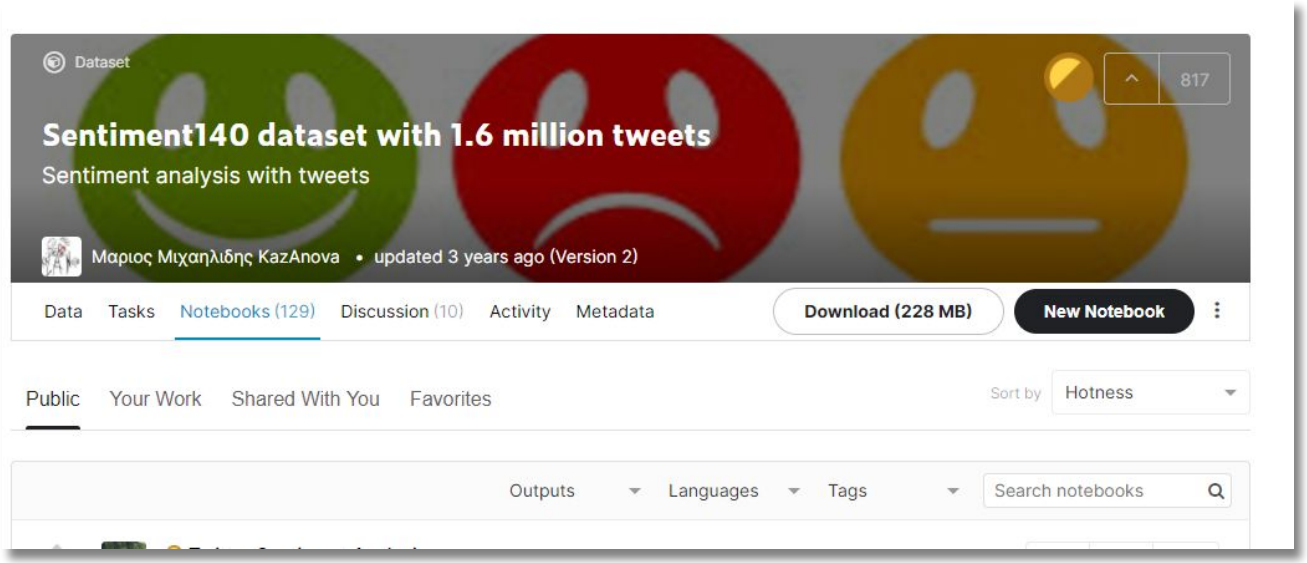
Abstract. *Brazilian Web users are among the most active in social networks and very keen on interacting with others. Offensive comments, known as* hate

Pesquisas relacionadas

- FACELI, KATTI; LORENA, ANA CAROLINA; GAMA, JOÃO; CARVALHO, ANDRÉ C. P. L. F. DE. **Inteligência Artificial**. Rio de Janeiro: LTC, 2019.

Pesquisas relacionadas

- https://www.kaggle.com/kazanova/sentiment140/kernels



Materiais e métodos

- Pandas
  - Biblioteca para análise e manipulação de dados
- Numpy
  - Biblioteca para manipulação de arrays
- Joblib
  - Biblioteca para persistência de objetos em disco, cache e paralelismo
- Matplotlib
  - Biblioteca para geração de gráficos

Materiais e métodos

- Imbalanced-learn
  - Biblioteca para manipulação de conjuntos de dados desbalanceados
  - SMOTE (Synthetic Minority Oversampling Technique)
    - Cria novos dados sintéticos a partir de dados que são considerados próximos

Materiais e métodos

- NLTK
  - Biblioteca para manipulação de conjuntos de dados voltados a Processamento de linguagem natural (NLP)
  - WordCloud
  - FreqDist
  - Stopwords
    - Lista de palavras comuns que são consideradas irrelevantes para o contexto

Materiais e métodos

- NLTK
  - White Space Tokenizer
    - Divide o texto em tokens a partir de espaços e novas linhas
  - RSLP Stemmer
    - Remove sufixos de palavras

Materiais e métodos

- Scikit learn
  - Biblioteca para treinamento de modelos supervisionados e não supervisionados, inclui também ferramentas para pré processamento de dados
  - TF-IDF Vectorizer
    - Converte textos em matrizes de valores TF-IDF (term frequency–inverse document frequency)
  - Pipeline
    - Executa uma sequência de tarefas

Materiais e métodos

- Scikit learn
  - SVD
    - Aplica redução de dimensionalidade
  - Select Percentile
    - Seleciona as melhores características de acordo com percentil das melhores pontuações

Materiais e métodos

- Scikit learn (seleção de hiperparâmetros)
  - Grid Search CV
    - Procura exaustivamente pelos melhores parâmetros para o classificador
  - Randomized Search CV
    - Procura pelos melhores parâmetros de forma randômica a um máximo de parâmetros especificados

Materiais e métodos

- Scikit learn
  - Logistic Regression
  - Gaussian NB
  - Multinomial NB
  - Random Forest Classifier

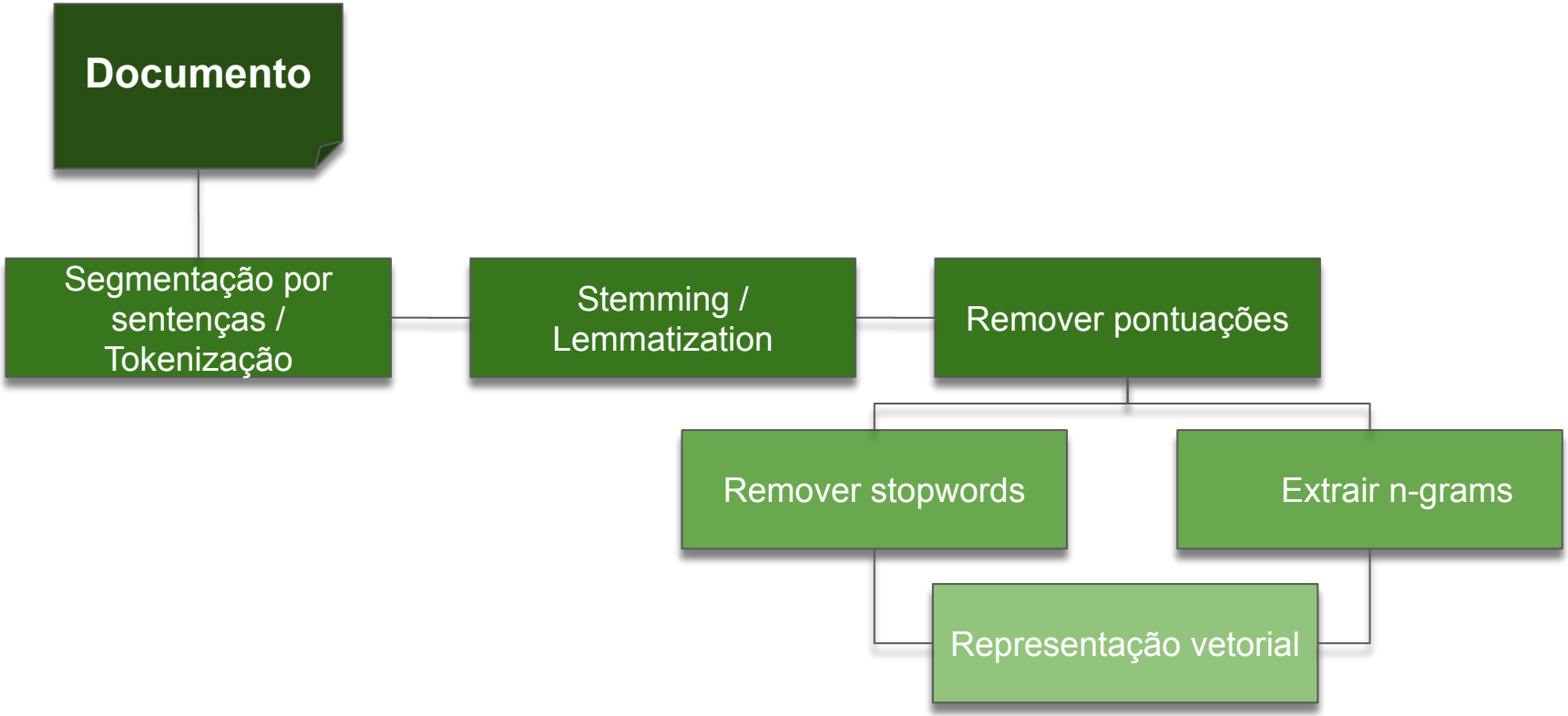
Análise dos dados

- Contagem de sentenças por classificação
- Redefinição da nomenclatura das colunas
- Redefinição da coluna de classificação (0 e 1)
- Frequência das palavras

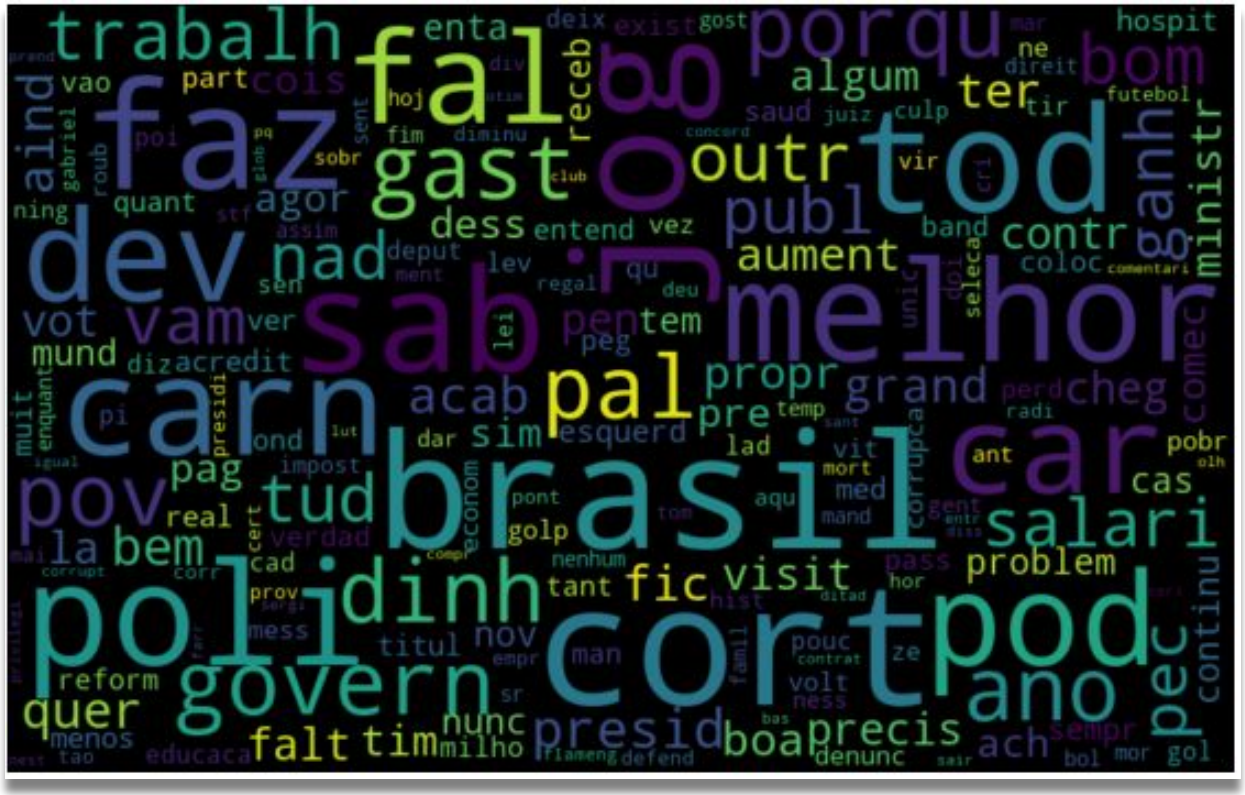
	id	@@class	document
0	1	yes	Votaram no PEZAO Agora tomem no CZAO
1	2	no	cuidado com a poupanca pessoal Lembram o que a...
2	3	no	Sabe o que eu acho engracado os nossos governa...
3	4	yes	os cariocas tem o que merecem um pessoal que s...
4	5	no	Podiam retirar dos lucros dos bancos

	Word	Frequence
32	e	640
64	de	426
13	que	419
12	o	407
8	a	316
27	nao	278
17	do	202
145	da	150
84	um	136
7	com	135

Técnicas de pré-processamento

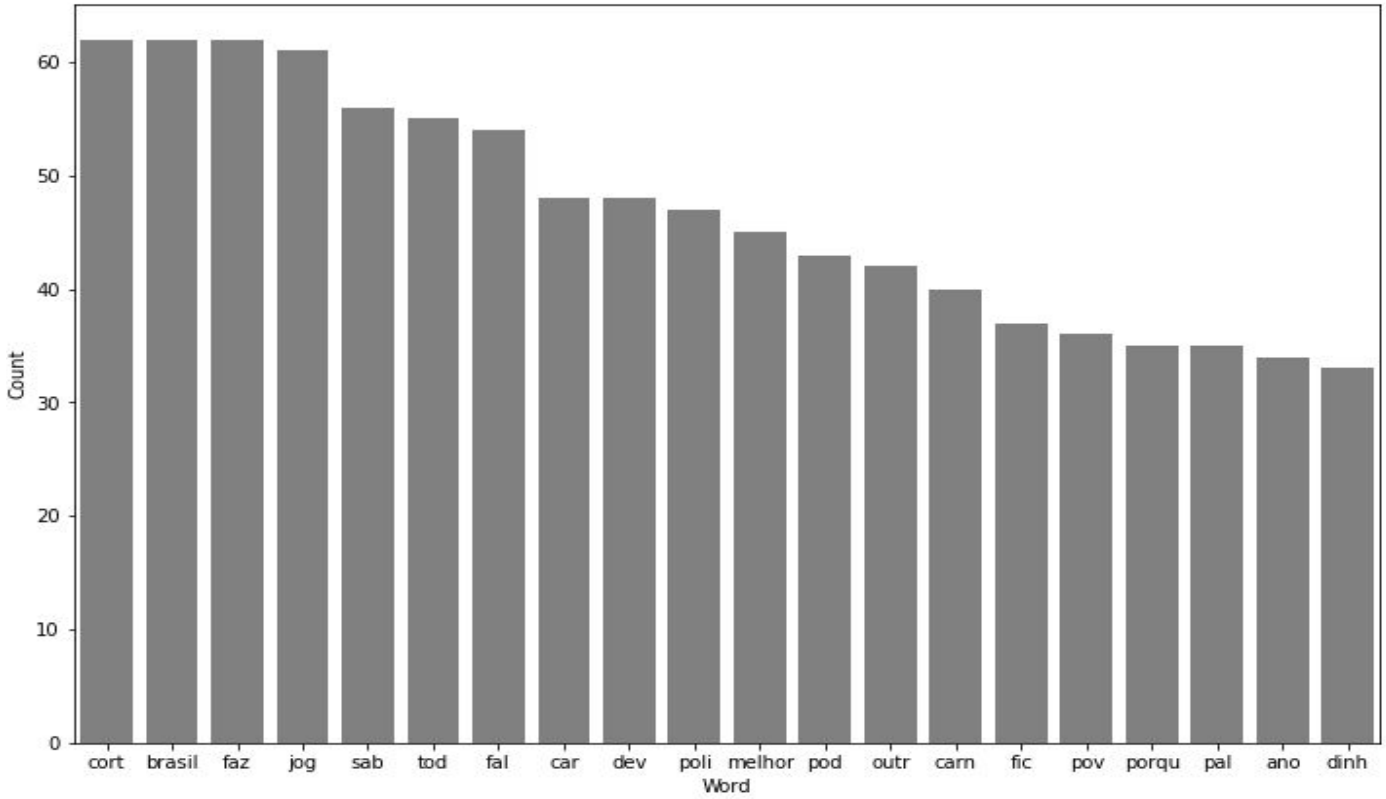


Visualizações

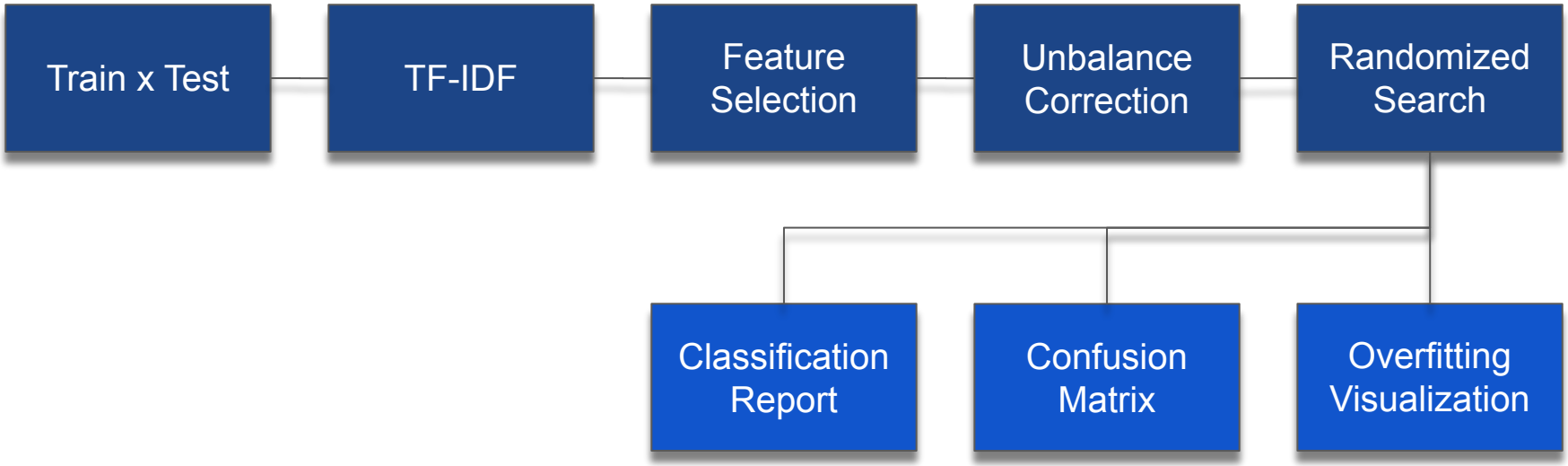


Visualizações

Visualizações



Configuração (para o processamento)



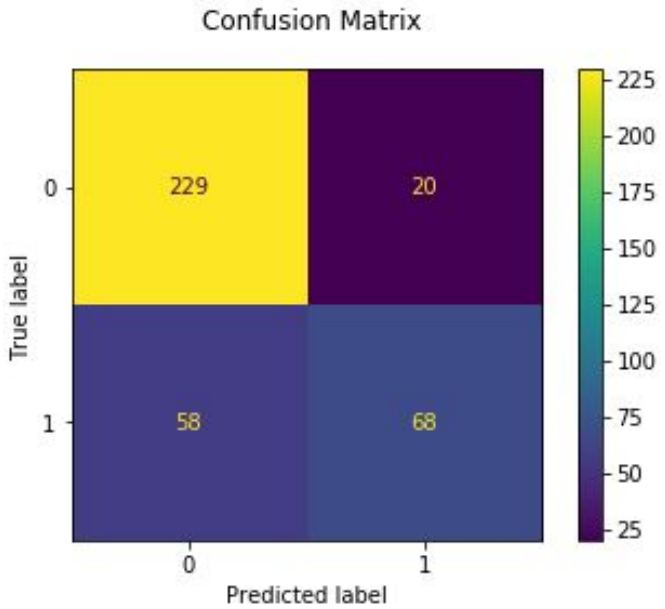
Resultados e discussões

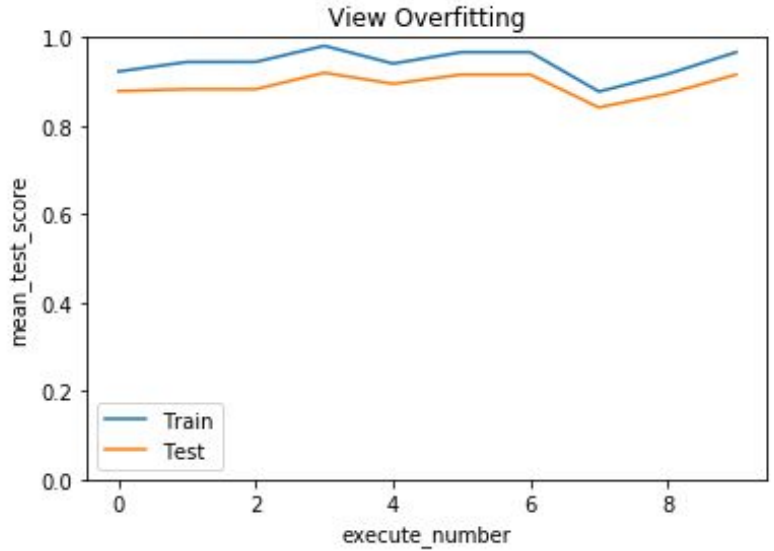
Regression Logistic: **Select Percentile**

```
F Medida: 0.64

Melhor configuração:  LogisticRegression(C=100, fit_intercept=False, solver='liblinear')

Melhor Score (F1):  0.9204498134756675


Classification Reports:
              precision    recall  f1-score   support

           0       0.80      0.92      0.85       249
           1       0.77      0.54      0.64       126

    accuracy                           0.79       375
   macro avg       0.79      0.73      0.74       375
weighted avg       0.79      0.79      0.78       375
```

Resultados e discussões

Regression Logistic: **Select Percentile**





Resultados e discussões

Regression Logistic: **Truncated SVD + Pipeline**

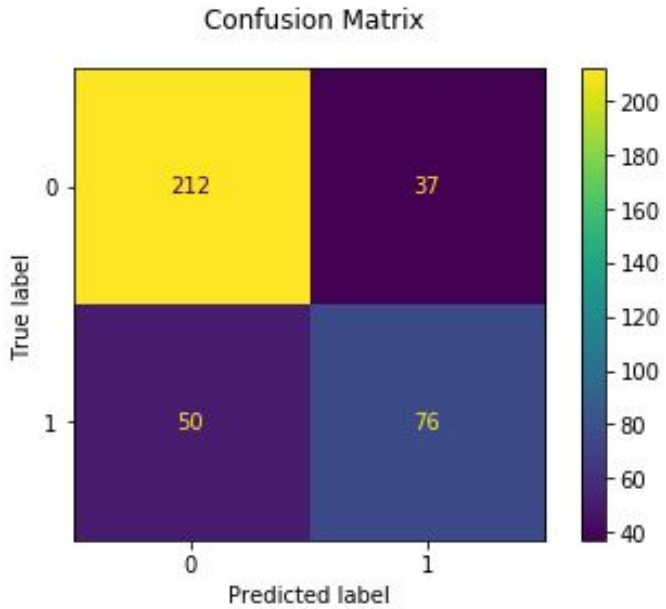
```
F Medida: 0.64

Melhor configuração:  Pipeline(steps=[('smt', SMOTE(sampling_strategy='minority')),
                ('svd', <__main__.SVDDimSelect object at 0x00000226B3823E88>),
                ('clf',
                 LogisticRegression(C=10, fit_intercept=False,
                                    intercept_scaling=4, max_iter=500,
                                    solver='liblinear'))])

Melhor Score (F1):  0.6240395634687447
```

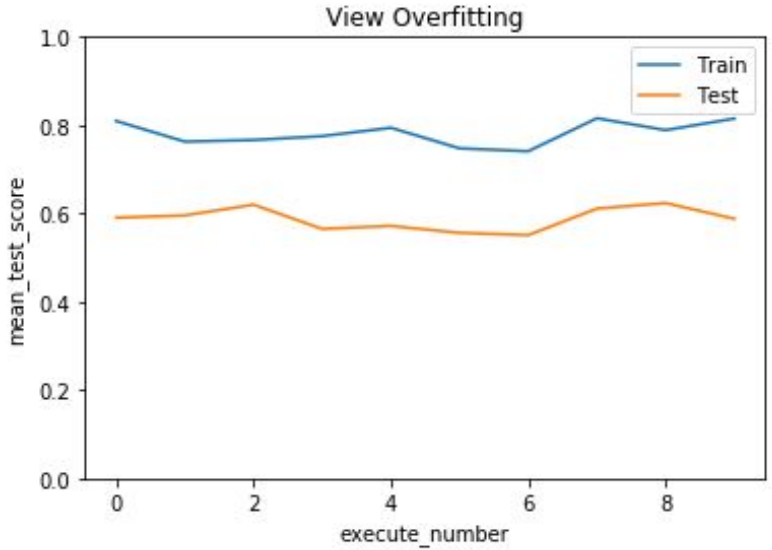
Classification Reports:

	precision	recall	f1-score	support
0	0.81	0.85	0.83	249
1	0.67	0.60	0.64	126
accuracy			0.77	375
macro avg	0.74	0.73	0.73	375
weighted avg	0.76	0.77	0.76	375

Resultados e discussões

Regression Logistic: **Truncated SVD + Pipeline**





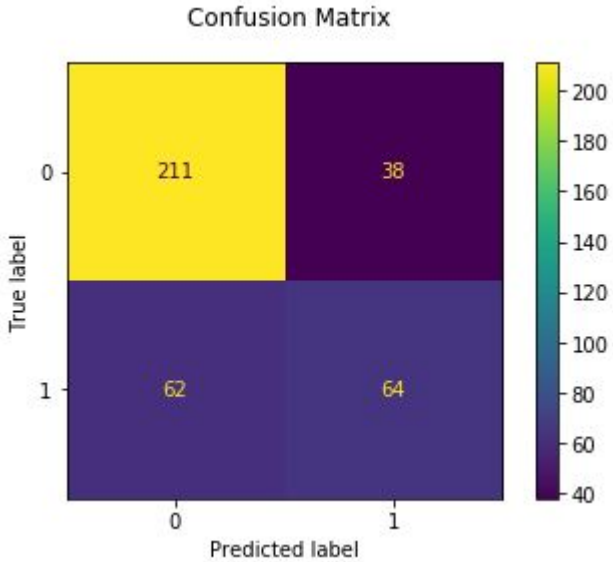
Resultados e discussões

Gaussian Naive Bayes: **Select Percentile**

F Medida: 0.56

Classification Reports:

	precision	recall	f1-score	support
0	0.77	0.85	0.81	249
1	0.63	0.51	0.56	126
accuracy			0.73	375
macro avg	0.70	0.68	0.68	375
weighted avg	0.72	0.73	0.73	375



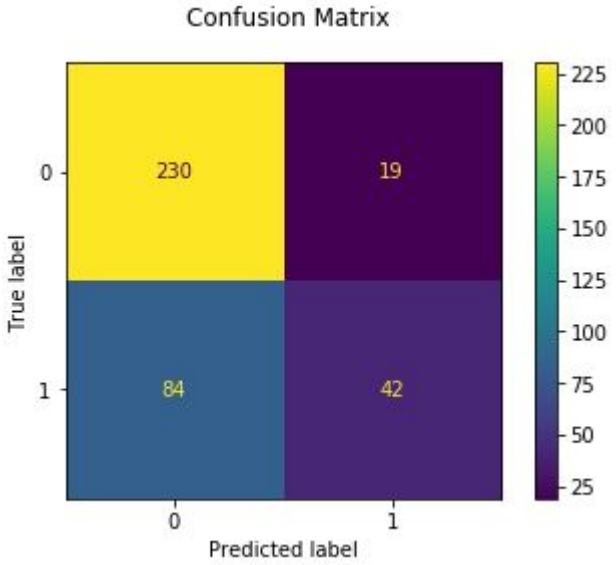
Resultados e discussões

Gaussian Naive Bayes: **Truncated SVD + Pipeline**

F Medida: 0.45

Classification Reports:

	precision	recall	f1-score	support
0	0.73	0.92	0.82	249
1	0.69	0.33	0.45	126
accuracy			0.73	375
macro avg	0.71	0.63	0.63	375
weighted avg	0.72	0.73	0.69	375



Resultados e discussões

Multinomial Naive Bayes: **Select Percentile**

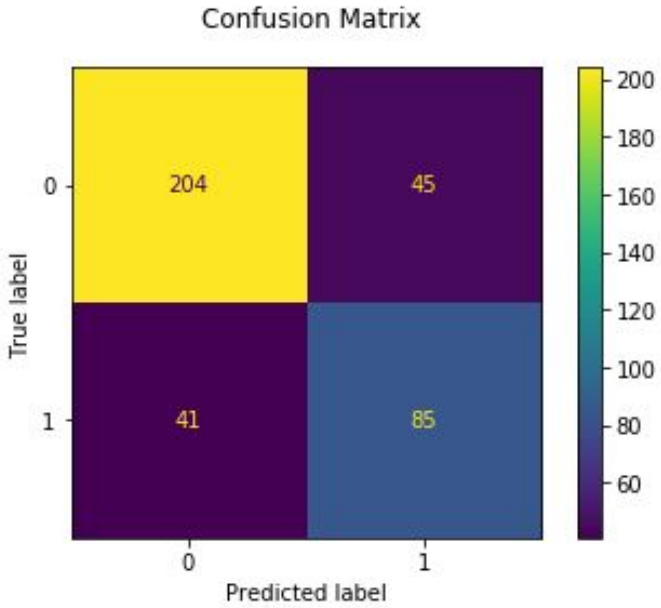
```
F Medida: 0.66

Melhor configuração:  MultinomialNB(alpha=0, fit_prior=False)

Melhor Score (F1):  0.9256734215693319
```

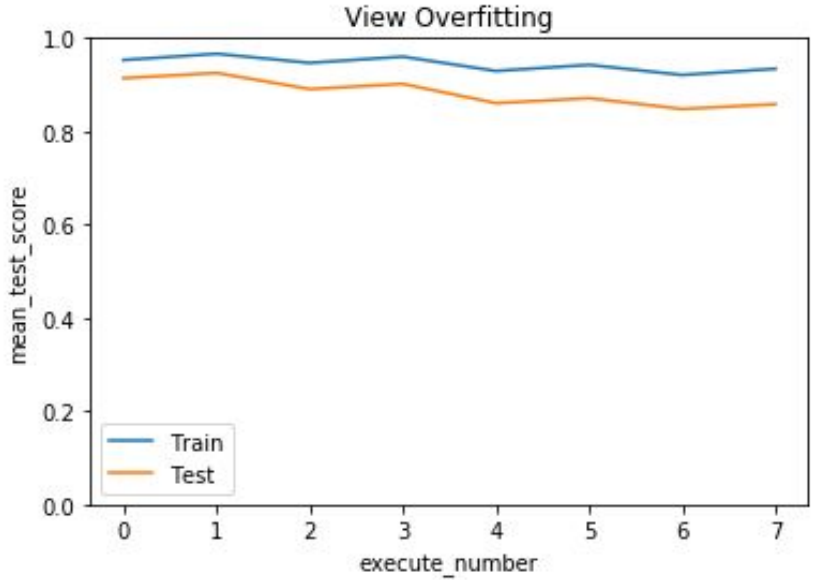
Classification Reports:

	precision	recall	f1-score	support
0	0.83	0.82	0.83	249
1	0.65	0.67	0.66	126
accuracy			0.77	375
macro avg	0.74	0.75	0.74	375
weighted avg	0.77	0.77	0.77	375

Resultados e discussões

Multinomial Naive Bayes: **Select Percentile**





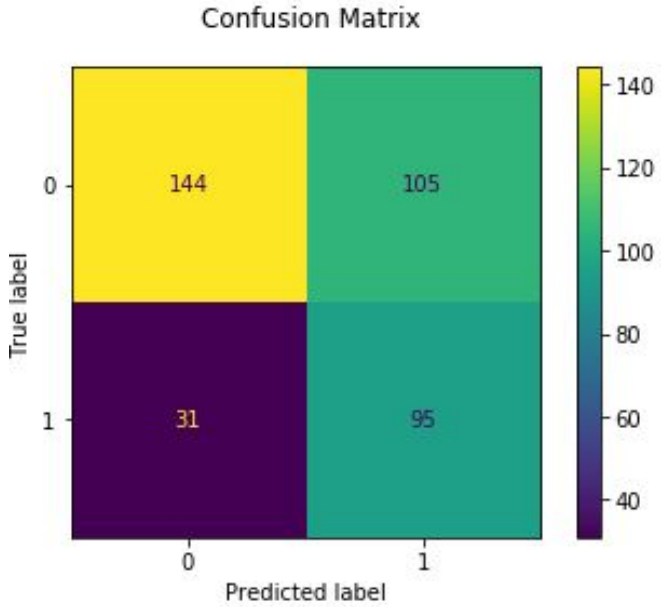
Resultados e discussões

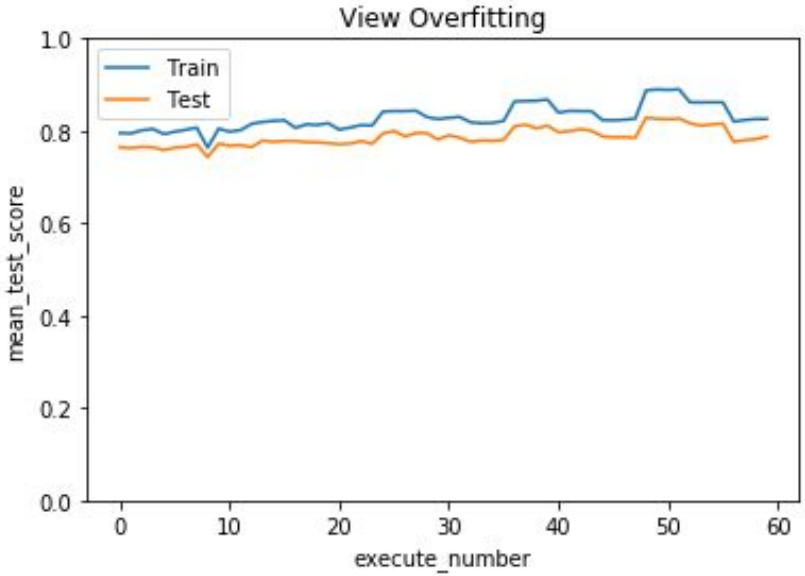
Random Forest Classifier: **Select Percentile**

```
F Medida: 0.58

Melhor configuração:  RandomForestClassifier(max_depth=40)

Melhor Score (F1):  0.8288952390183001


Classification Reports:
              precision    recall  f1-score   support

           0       0.82      0.58      0.68       249
           1       0.47      0.75      0.58       126

    accuracy                           0.64       375
   macro avg       0.65      0.67      0.63       375
weighted avg       0.71      0.64      0.65       375
```

Resultados e discussões

Random Forest Classifier: **Select Percentile**





Resultados e discussões

Random Forest Classifier: **Truncated SVD + Pipeline**

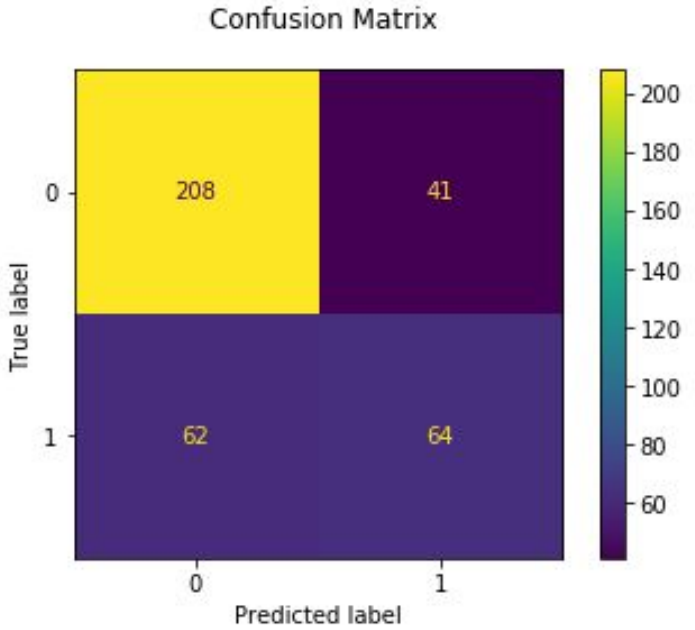
```
F Medida: 0.55

Melhor configuração:  Pipeline(steps=[('smt', SMOTE(sampling_strategy='minority')),
                ('svd', <__main__.SVDDimSelect object at 0x00000226B244D0C8>),
                ('clf',
                 RandomForestClassifier(max_depth=40, n_estimators=500))])

Melhor Score (F1):  0.6141733188088948
```

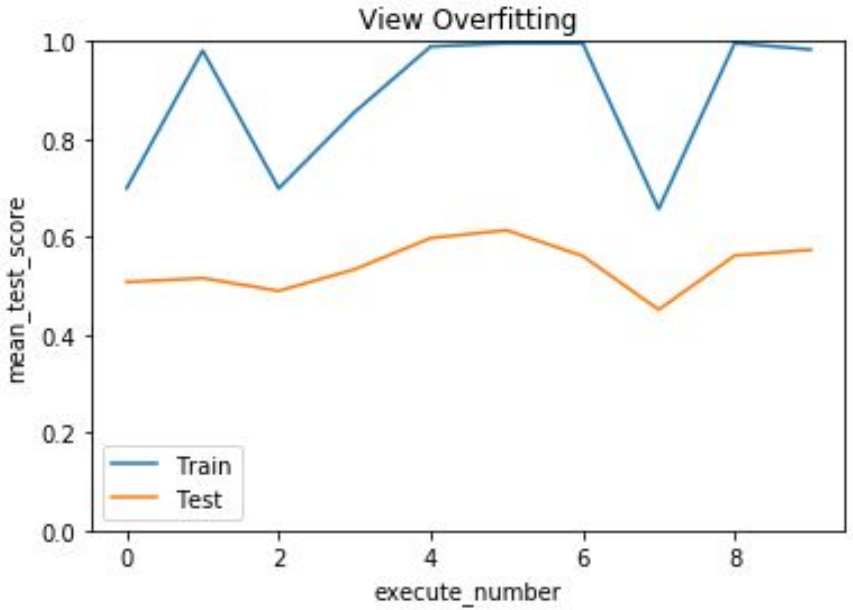
Classification Reports:

	precision	recall	f1-score	support
0	0.77	0.84	0.80	249
1	0.61	0.51	0.55	126
accuracy			0.73	375
macro avg	0.69	0.67	0.68	375
weighted avg	0.72	0.73	0.72	375

Resultados e discussões

Random Forest Classifier: **Truncated SVD + Pipeline**

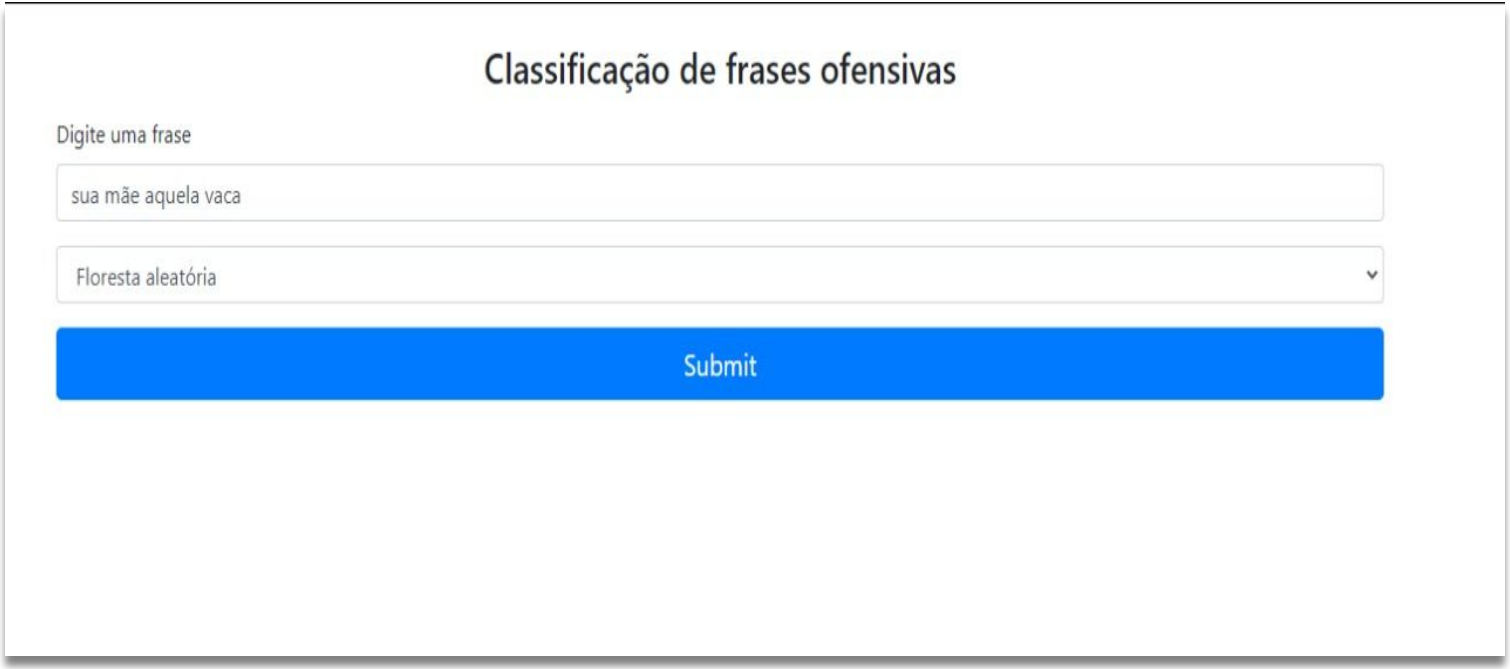




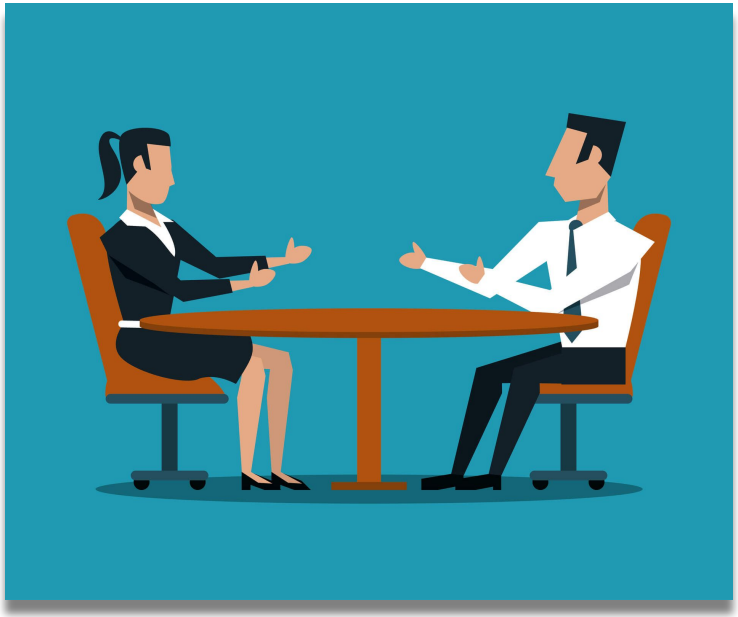
Conclusão e considerações finais

- Utilizados classificadores e técnicas acessíveis ao conjunto de dados utilizado.
- Recursos heurísticos apresentaram melhores resultados.
- A estratégia de aplicação dos recursos fez-se eficiente na apresentação de resultados justos, considerando a F Medida - assim como experimentos passados demonstraram o contrário.
- A quantidade de dados presentes no dataset (bem como seu histórico de criação) contribuíram para as justificativas finais.
- Maiores quantidades de dados e assuntos diversos poderiam contribuir mais com o uso de outros classificadores / redes neurais.

Deploy do Modelo



Feedback



https://github.com/adrianovss/hate-speeches-facens