Universidade de São Paulo
Faculdade de Filosofia, Letras e Ciências Humanas
Departamento de Ciência Política

FLP 0478 - Processamento de Língua Natural Aplicada para Ciência Política e Análise de Políticas Públicas (2024)

2° semestre / 2024

Laboratório 12 - Aula 13 (18/11/2024)

Professora Lorena G. Barberia

Prazo: 02 de dezembro, 2024

A tarefa deste laboratório será a de montar o seu script em Python do zero, utilizando **pelo menos 05** modelos de classificação (exceto modelos de *Deep Learning*), classificando o posicionamento dos tweets do Corpus "Mapping Political Elites COVID-19 Vaccine Tweets in Brazilian Portuguese in 2020, 2021 and 2022, Versão 2.0" (todas as classes) com relação a amostra de tweets identificados como relevantes para 2020.

Um objetivo importante deste trabalho é analisar quais modelos são melhores por aumentar a capacidade de classificar tweets com um posicionamento desfavorável e de poder explicar o aumento ou ganhos relativos ao um *baseline* ou *benchmark dentro do mesmo modelo e comparando os modelos*.

O trabalho deve ter um texto com máximo de 5.000 palavras com 4 figuras e 4 tabelas. O trabalho deve estar dividido nas seguintes seções:

1. Introdução
    - Apresentação e Justificativa dos modelos experimentais
2. Definindo a *baseline*
    - Divisão Treino/Teste (Quantos % para teste? Por quê?)
    - Bag-of-Words ou TF-IDF? Qual o n-gram? Como o texto foi pré-processado?
    - Faça uma classificação com K-fold Cross-Validation (K = 10) utilizando o **Multinomial Logit**.
    - Quais foram os resultados de **Validação** no K-Fold? (e.g. Precision, Recall, F1-score)
    - Como o modelo performa no banco de **Teste**? Accuracy, Confusion Matrix, Classification Report.
3. Modelos e Resultados
    a. Faça um Grid Search (ou Random Search) com os outros modelos. Ao final, faça uma tabela no relatório mostrando os melhores hiperparâmetros em cada modelo. (Seja criativo, está liberado modelos ensemble).
    b. Faça um K Fold (10 folds) e apresente os resultados de **validação** dos modelos. De preferência, faça *boxplots* mostrando accuracy, precision, recall e f1-score para todos os modelos. Tente apresentar os resultados de todos os modelos em uma métrica juntos.
    c. Resultados de **Teste** de cada modelo - Accuracy, Confusion Matrix, Classification Report.

4. Conclusão
    a. Como podemos avaliar os modelos e sua capacidade preditiva? Quais modelos são mais eficientes? Quais modelos têm ganhos mais significativos relativos ao baseline (multinomial logit) ou benchmark após o grid search?
    b. O que aprendemos de overfitting (Resultados de Validação vs. Teste)
    c. O que aprendemos comparando os modelos?

Vocês devem entregar um relatório de em .doc ou .pdf, junto do seu script em Python (Formato **.ipynb**), no Moodle na caixa do "Laboratório Aula 12" de seu respectivo turno. O prazo para a entrega é na última aula do curso, no dia 02/12/2024.