UNIVERSIDADE DE SÃO PAULO
ESCOLA POLITÉCNICA DA UNIVERSIDADE DE SÃO PAULO
DEPARTAMENTO DE ENGENHARIA DE COMPUTAÇÃO E SISTEMAS DIGITAIS
CENTRO DE CIÊNCIA DE DADOS (C²D)

NOVO MODELO DE RELATÓRIO DOS BOLSISTAS

09/2021

Dados da Bolsa
Tipo de Bolsa: (X) IC ( ) TCC ( ) PqEP ( ) ME
Nome do/a Orientador/a: Anna Helena Reali Costa
Nome do Projeto: Democracia Aumentada
Período da Bolsa: 02/02/2022 a 01/02/2023
Relatório: ( ) Final (X) Parcial
Período do Relatório: 25/05/2022 a **25/06/2022**
Descrição das Atividades de Pesquisa do Projeto
Descrição das atividades acadêmicas: 2° Semestre Letivo, em que o aluno está cursando as seguintes matérias: ● PEA3306 - Conversão Eletromecânica de Energia ● PEA3301 - Introdução aos Sistemas de Potência ● PEA3311 - Laboratório de Conversão Eletromecânica de Energia ● PTC3307 - Sistemas e Sinais ● PCS3335 - Laboratório Digital A ● PSI3213 - Circuitos Elétricos II ● PEA3100 - Energia, Meio Ambiente e Sustentabilidade ● PMT3100 - Fundamentos de Ciência e Engenharia dos Materiais ● PMT3131 - Química dos Materiais Aplicada à Engenharia Elétrica
Descrição das atividades planejadas para o relatório (repetir do relatório anterior): Rotulagem da base de dados entre as subáreas do mercado financeiro e análise de sentimentos utilizando ZeroBERTo (ou modelo conveniente)
Descrição das atividades de pesquisa realizadas: Durante este mês o bolsista realizou a rotulagem de dados das bases financeiras, segue então o resultado dos testes realizados: - Para classificação em setores econômicos:

- Foi testada a utilização do ZeroBERTo, porém pelo número de rótulos o modelo se confundia gerando distribuições diferentes a cada vez que era executado (hora com mais dados em Comunicação, hora mais dados em Bens Industriais, não chegando a uma estabilização na classificação). A estratégia adotada para tentar sanar isso foi utilizando várias frases de *template* (observe que o ZeroBERTo é em parte um modelo de Zero-Shot Classification), porém mesmo assim os resultados continuaram sendo não-ótimos.
- A abordagem utilizada foi então através de REGEX, ou seja, basicamente foi criado um código que vê as citações diretas a empresas de um determinado ramo econômico e classifica a frase com base nisso; por exemplo, se tivermos uma frase como "Raia Drogasil perdeu 10% dos seus investimentos" anotamos o setor econômico como Saúde, pois Raia Drogasil é uma empresa de saúde. Dessa forma cerca de ⅓ dos dados conseguiram ser rotulados (~300.000) adequadamente.

- Para classificação em sentimentos:
  - Foi testada primeiramente a utilização do ZeroBERTo para análise de sentimentos, porém novamente não foi bem-sucedida, dessa vez pelo fato do modelo entrar em *looping* infinito com o número de dados de entrada mapeando um cluster de palavras nele mesmo na etapa que utiliza o BERTopic.
  - Uma segunda abordagem foi tentando utilizar o LEIA, que é um léxico para análise de sentimentos em português; porém, a biblioteca não é estruturada para *imports* em *pip* (package installer for Python) e o bolsista não conseguiu utilizar a biblioteca adequadamente
  - A terceira abordagem que funcionou, foi utilizando o TextBlob para a tradução dos textos de português para inglês e, em seguida, utilizando o VADER que é um léxico para análise de sentimentos em inglês. Apesar de funcional, a demora para executar a etapa de tradução automática limitou a rotulagem de dados em apenas 20.000 dados, que serão utilizados para a etapa futura do projeto.
  - Ainda existe margem para outros testes em análise de sentimentos, porém não foram encontradas boas referências em português para a tarefa, o que limitou bastante o processo.

Descrição das próximas atividades:

Para a próxima etapa, planeja-se quais modelos terão dentro da interface, criação de modelos para predizer o setor com base nos dados utilizados, e utilização do VADER para análise de sentimentos da etapa futura.

Houve alteração significativa no tema ou prazo: ( ) Sim (X) Não

Apreciação Circunstanciada do/a Orientador/a sobre as Atividades da/o Bolsista

Etapa cumprida no relatório: ( ) Ótimo (X) Bom ( ) Regular ( ) Fraco

Programação para a próxima etapa: ( ) Ótimo (X) Bom ( ) Regular ( ) Fraco

Resultados em relação às expectativas iniciais: ( ) Acima (X) Dentro ( ) Abaixo ( ) Muito abaixo

Previsão de conclusão no prazo: (X) Sim ( ) Não **Justifique em caso negativo:**
APRECIAÇÃO: O bolsista tem desempenhado de forma adequada, visando terminar o projeto até setembro, quando irá estudar no exterior. Tem buscado ferramentas para concluir o projeto com o prazo reduzido, como explicado em relatórios anteriores, sem, entretanto, prejudicar os resultados da pesquisa.
Protocolo
Data: 25/06/2022
Nome Completo da/o Bolsista: Enzo Bustos Da Silva