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

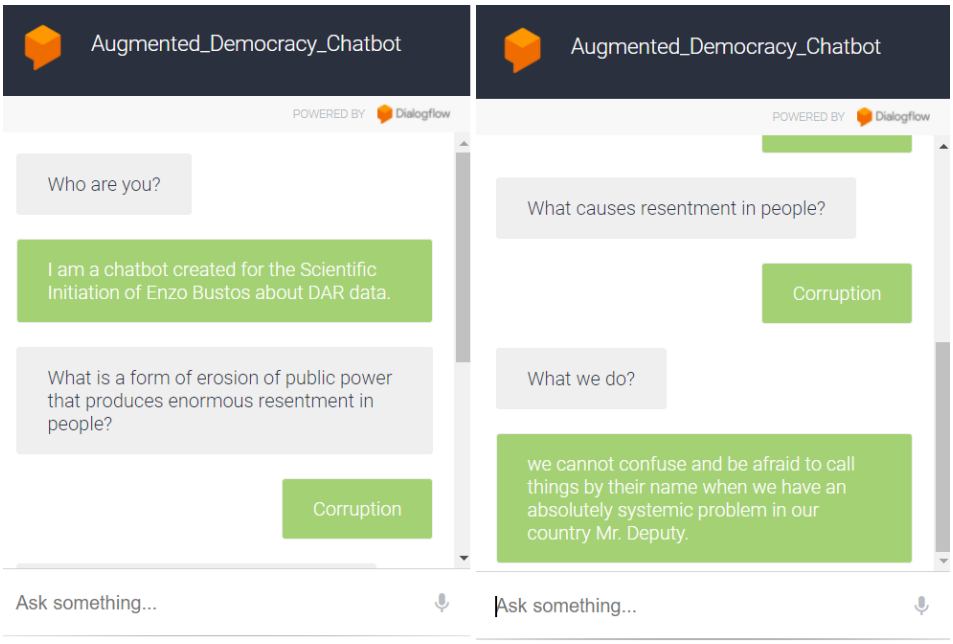
09/2021

Dados da Bolsa
Tipo de Bolsa: (X) IC ( ) TCC ( ) PqEP ( ) ME
Nome do/a Orientador/a: Anna Helena Reali Costa
Nome do Projeto: Chatbot Q&A multi-agente
Período da Bolsa: 01/02/2021 a 01/02/2022
Relatório: (X) Final ( ) Parcial
Período do Relatório: 25/12/2021 a **25/01/2022**
Descrição das Atividades de Pesquisa do Projeto
Descrição das atividades acadêmicas: Período de férias acadêmicas no mês de janeiro de 2022.
Descrição das atividades planejadas para o relatório (repetir do relatório anterior): Para este relatório estava planejado finalizar o Chatbot para concretizar o encerramento da bolsa
Descrição das atividades de pesquisa realizadas (por mês): 1. Realização das primeiras etapas da revisão da literatura, focando na teoria básica de NLP, como *Lematização* e *Embeddings* 2. Continuação da revisão literária, com foco especial para a aplicação em *chatbots*, com estudo sobre Intenção, Entidade e Contexto. Submissão do artigo intitulado "NLP, Data Analytics and Dashboards: People's Tools to Influence Anticorruption Political Agendas", submetido ao evento Data Analytics for Anticorruption in Public Administration, organizado pelo World Bank (este artigo não foi aceito para publicação). 3. Criação da base de dados dos textos do Diário da Assembleia da República Portuguesa (DAR) e criação do primeiro esquema de *Dashboard* com citações diretas e indiretas à corrupção. Primeiros testes com *Random Forest* do DEBACER. 4. Aplicação de rotinas de pré-processamento ao córpus (typos, tokenização, stopwords, lematização e word2vec) e criação do notebook interativo.

5. Foco na finalização do artigo submetido ao Bracis 2021, com foco em um teste de ablação comparando o BERTimbau com arquiteturas de *machine learning* não-neurais
6. Foco na submissão de artigo para o seminário “INTELIGÊNCIA ARTIFICIAL: DEMOCRACIA E IMPACTOS SOCIAIS”, além do DEBACER foi adicionada a ideia do SISTEMA de Democracia Aumentada que reúne diversas abordagens de NLP para extrair dados relevantes das atas do DAR
7. Estudo e testes com Sumarização Automática, testando diversas arquiteturas como BertExt, BART e PEGASUS. Além da submissão do artigo "DEBACER: a method for slicing moderated debates" para o ENIAC 2021.
8. Definição dos módulos para o SISTEMA de Democracia Aumentada, com Sumarização Automática, Modelagem de Tópicos e Análises Quantitativas. Além disso, houve a apresentação da IC no 29° SIICUSP (Apresentação).
9. Início dos estudos para a criação de um aplicativo *web* para a confecção do SISTEMA de Democracia Aumentada com as funcionalidades já citadas. Submissão do artigo “ZeroBERTo - Leveraging Zero-Shot Text Classification by Topic Modeling” para o PROPOR 2022
10. Continuação dos estudos com Django para a criação da Interface Web. (Será continuada durante a próxima bolsa de IC, com início em fevereiro de 2022)
11. Retomada da construção do *chatbot*, definindo o uso da plataforma DialogFlow e criação de uma base de dados paralela que foi usada para construção das Requisições e Respostas do agente.

12. Neste último mês foi finalizado o *chatbot.* A partir da base de dados foi utilizado um gerador de pares de perguntas e respostas (Question Generation) que utiliza o modelo T5, porém como o modelo só funciona para a língua inglesa a base teve que ser traduzida. A partir destes pares foi construída uma base de dados de pares Pergunta e Resposta (Dataset Original), este dataset teve que ser limpado, com a remoção de alguns destes pares por conta principalmente de alucinação do modelo, por não ter sido usada uma passagem de texto independente e sim um conjunto de falas que tem um contexto. Dessa forma, foram removidas perguntas do tipo “Quem”, que em sua maioria constavam de alucinação do modelo, por conta da perda de contexto e remoção de referências sobre quem está com a palavra (e a quem está respondendo), criando-se a nova base de dados (Dataset Limpo). Esta base foi encaminhada ao Knowledge Bases do DialogFlow para que o *chatbot* pudesse ser montado.

Considerações: A construção do chatbot foi um sucesso, porém pelo fato do modelo usado não ter sido treinado para textos sequenciais (como é o caso dos textos das atas do DAR, usadas neste trabalho), existem muitos pares que não fazem sentido algum, além da aplicação apenas em inglês por conta do Knowledge Bases que é uma *feature beta* do DialogFlow.



Na imagem temos algumas aplicações, observe que na primeira temos um caso que é bem respondido e também, além de também ser generalizado (ambas as perguntas têm o mesmo teor), porém no último exemplo percebemos uma alucinação em que o par de pergunta e resposta foi criado de uma fala que não faz sentido sem o seu contexto.

O *Chatbot* pode ser testado através deste link

Objetivos alcançados:

- Conclusão de uma boa revisão bibliográfica que serviu para aprofundar os conhecimentos na área de *machine learning* e NLP
- Uso e familiarização com modelos pautados como o estado-da-arte atual.
- Uso da arquitetura de *transformers* para extrair *features* mais relevantes dos dados
- Divulgação Científica (será detalhada abaixo)
- Criação do *chatbot* FAQ sobre as atas do DAR.

Artigos Aprovados:

- Publicação de artigo aceita no XVIII Encontro Nacional de Inteligência Artificial e Computacional (ENIAC 2021): Link
- Publicação de artigo aceita na Conferência Internacional de Processamento Computacional da Língua Portuguesa (PROPOR 2022): Link
- Publicação de artigo aceita no 1° Seminário Internacional de Humanidades - Artificial Intelligence: Democracy And Social Impacts (C4AI 2021): Ainda não publicado

Descrição das próximas atividades: Planeja-se utilizar o próximo período de bolsa para aprimorar o que foi desenvolvido este ano, além de finalizar a Interface Web do SISTEMA de Democracia Aumentada.
Houve alteração significativa no tema ou prazo: ( ) Sim (x) Não **Justifique em caso positivo:**
Apreciação Circunstanciada do/a Orientador/a sobre as Atividades da/o Bolsista
Etapa cumprida no relatório: (x) Ótimo ( ) Bom ( ) Regular ( ) Fraco
Programação para a próxima etapa: (x) Ótimo ( ) Bom ( ) Regular ( ) Fraco
Resultados em relação às expectativas iniciais: (x) Acima ( ) Dentro ( ) Abaixo ( ) Muito abaixo
Previsão de conclusão no prazo: (x) Sim ( ) Não **Justifique em caso negativo:**
Apreciação da orientadora: O aluno de IC amadureceu muito no período da bolsa, aprendendo e trabalhando com diversas técnicas de NLP. Fez um ótimo trabalho, participando de vários artigos e apresentando seu trabalho no SIICUSP 2021. Ele foi selecionado para um novo período de bolsa de IC, dando continuidade à sua pesquisa no tema.
Protocolo
Data: 25/01/2022
Nome Completo da/o Bolsista: Enzo Bustos Da Silva