Relatório Semestral de Atividades - Sistema Atena	
Tipo da Bolsa:	Iniciação Científica
Título do Projeto:	Chatbot Q&A multi-agente
Tópico Abordado:	PLN e Chatbots
Aluno:	Enzo Bustos da Silva
Departamento e Unidade do Aluno:	Escola Politécnica - PCS
Ano de Ingresso:	2019
Orientador:	Anna Helena Costa Reali
Departamento e Unidade do Orientador:	Escola Politécnica - PCS
Período das Atividades Desenvolvidas:	01/10/2021 a 15/03/2022

1. Principais objetivos iniciais do Plano de Pesquisa

Dentre os objetivos que estavam inicialmente pautados no Plano de Pesquisa vale pontuar principalmente os estudos da bibliografia-base em Inteligência Artificial (AI) e Aprendizado de Máquina (AM), além do aprendizado em técnicas específicas para o Processamento de Linguagem Natural (PLN) que envolvem o pré-processamento de textos, extração de informações relevantes e também o uso das ferramentas pautadas como estado-da-arte atual como os Transformers e o Hugging Face, usados em diversas aplicações, tais como Sumarização Automática e Modelagem de Tópicos.

2. Descrição das atividades de pesquisa realizadas (por mês):

    1. Realização das primeiras etapas da revisão da literatura, focando nos métodos clássicos de AI e AM e aprofundamento na teoria de PLN, como a vetorização de palavras (Embeddings) e métodos de simplificação do léxico (Lematização).
    2. Continuação da revisão bibliográfica, atentando principalmente para as publicações mais relevantes e métodos apontados como o estado-da-arte em PLN, como os Transformers, além da revisão da arquitetura de chatbots analisando Intenção, Entidade e Contexto desses agentes.
    3. Foi realizada a criação da base de dados, os textos utilizados foram extraídos do Diário da Assembleia da República Portuguesa (DAR), utilizando um programa de *web scraping* baseado em Selenium, de modo a baixar as atas em extensão .txt automaticamente. Após isso foi criada a primeira rotina para mostrar um Dashboard, que consistia de menções diretas ou indiretas à palavra “corrupção” no discurso dos deputados na DAR. Também foram desenvolvidos os primeiros métodos de segmentação da ata da DAR em blocos de discursos do mesmo assunto, que foi batizado de DEBACER, utilizando Random Forest.
    4. Foram aplicadas rotinas de pré-processamento ao córpus, como remoção de typos e stopwords, aplicação de lematização e tokenização, e finalmente, vetorização dos textos utilizando o word2vec pré-treinado. Além da otimização dos algoritmos feitos anteriormente, com a criação também de um notebook interativo.
    5. O foco principal foi dado à finalização do artigo submetido ao ENIAC 2021 (ver publicações a seguir), em especial no desenvolvimento de um teste de ablação comparando o BERTimbau com outras arquiteturas de machine learning não-neurais, no contexto do DEBACER. Além disso, foi realizada a apresentação da IC no 29° SIICUSP (link da apresentação) (site do SIICUSP).
    6. Neste último mês do semestre de IC, o esforço foi para a submissão de mais dois artigos (ver a seguir em Artigos Resultantes):
        a. Para o seminário: “INTELIGÊNCIA ARTIFICIAL: DEMOCRACIA E IMPACTOS SOCIAIS” (artigo submetido à Revista do IEA da USP);
        b. Para a Conferência Internacional de Processamento Computacional da Língua Portuguesa (artigo aceito no PROPOR 2022).

3. Principais resultados alcançados:

- Conclusão de uma ampla revisão bibliográfica.
- Aprofundamento nos conhecimentos na área de machine learning e PLN.
- Uso e familiarização de modelos de inteligência artificial para PLN.
- Divulgação Científica (será detalhada a seguir).

4. Artigos Resultantes:

- FERRAZ, Thomas Palmeira et al. DEBACER: a method for slicing moderated debates. *In*: ENCONTRO NACIONAL DE INTELIGÊNCIA ARTIFICIAL E COMPUTACIONAL (ENIAC), 18. , 2021, Evento Online. Anais [...]. Porto Alegre: Sociedade Brasileira de Computação, 2021. p. 667-678. DOI: https://doi.org/10.5753/eniac.2021.18293.

- Alcoforado, A. et al. (2022). ZeroBERTo: Leveraging Zero-Shot Text Classification by Topic Modeling. In: Computational Processing of the Portuguese Language. PROPOR 2022. Lecture Notes in Computer Science, v. 13208. Springer, Cham. DOI https://doi.org/10.1007/978-3-030-98305-5_12

- Alcoforado, A. et al. (2022). Augmented Democracy: Artificial Intelligence as a Tool to Fight Disinformation. Artigo aceito no 1° Seminário Internacional de Humanidades - Artificial Intelligence: Democracy And Social Impacts (C4AI 2021) e em avaliação para publicação na Revista de Estudos Avançados do Instituto de Estudos Avançados da Universidade de São Paulo (versão impressa ISSN 0103-4014; versão on-line ISSN 1806-9592): *em análise*.

5. Próximas atividades:

Para o próximo semestre da iniciação científica, é planejado começar a utilizar os modelos pautados como estado-da-arte em PLN, os *Transformers*, para aplicações mais específicas como os módulos de Sumarização Automática e Modelagem de Tópicos para a confecção do Chabot.