UNIVERSIDADE DE SÃO PAULO
ESCOLA POLITÉCNICA DA UNIVERSIDADE DE SÃO PAULO
DEPARTAMENTO DE ENGENHARIA DE COMPUTAÇÃO E SISTEMAS DIGITAIS
CENTRO DE CIÊNCIA DE DADOS (C^2D)

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 (finalização antecipada)
Relatório: (X) Final ( ) Parcial
Período do Relatório: 25/07/2022 a **31/08/2022**
Descrição das Atividades de Pesquisa do Projeto
Descrição das atividades acadêmicas: O Aluno encontra-se em período de férias acadêmicas devido à realização de intercâmbio para a faculdade RTWH Aachen University que se inicia no Winter Semester 2022 (Outubro)
Descrição das atividades planejadas para o relatório (repetir do relatório anterior): Definir ambos os módulos que estarão presentes na interface do hugging face e início na implementação do módulo online
Descrição das atividades de pesquisa realizadas (por mês): 1. Por se tratar de uma continuação da IC realizada em 2021, o início desta nova Iniciação Científica começou focado no projeto Democracia Aumentada, o qual visava utilizar os dados do Diário Da Assembleia da República (DAR) Portuguesa como base para avaliar modelos e também com uma revisão bibliográfica mais aprofundada principalmente em técnicas de Processamento de Linguagem Natural mais avançadas, as quais utilizavam modelos de transformers (dentre eles vale citar técnicas de Robustness, Ataques em textos, Self-Training e Co-Training que foram estudadas pelo aluno). 2. A partir do segundo mês o projeto do Democracia Aumentada ficou instável, então buscou-se novas opções para a continuação da pesquisa, dessa forma o segundo mês continuou sendo uma revisão bibliográfica, porém dessa vez averiguando diferentes tipos de projetos e dados para continuar os rumos da pesquisa. 3. No terceiro mês o aluno foi aprovado para realizar intercâmbio para a RTWH Aachen University, então visou-se usar dados mais facilmente disponíveis para a finalização do prometido no plano de pesquisa da IC. Passou-se a focar no desenvolvimento de uma plataforma com módulos de machine learning,

utilizando a ferramenta Spaces do HuggingFace que permite a criação de páginas web de forma mais fácil do que utilizando o Django e criando uma ferramenta do zero. Neste mês também foi proposto um novo cronograma para a IC, com as ligeiras alterações sendo descritas nos relatórios de IC do aluno.

4. No quarto mês foi definido um novo escopo para o projeto, que não mais utilizaria os dados do DAR e sim dados de textos financeiros extraídos dos sites da Exame, InfoMoney e Valor Econômico (revistas do mercado financeiro). Com essa base de dados em mãos também foram definidos os módulos para apresentação na plataforma (vide Figura 1), que consistiram de: (1) Tradução Automática, (2) Análise de Sentimento e (3) Classificação ZeroShot dos Temas. A base de dados financeiros completa tem 1.526.914 linhas de dados.
5. O quinto mês serviu para a implementação dos módulos e testes exaustivos sobre quais arquiteturas melhor encaixariam na interface do HuggingFace. Para classificação em setores econômicos foi testado o antigo ZeroBERTo e uma abordagem heurística que utiliza uma citação explícita a uma empresa dentro do texto. Para a análise de sentimentos foi testado também o ZeroBERTo, além de pacotes como o Leia e VADER. Porém os testes não foram muito promissores com a base de dados que existia até o momento.
6. No sexto mês foi definida uma sub-base de dados (Figura 2), com 50.000 linhas que contém os textos traduzidos para língua inglesa (utilizando-se TextBlob) e, dessa forma, os textos poderiam ser processados por muitos outros tipos de transformers, que serão discutidos no desenvolvimento realizado este mês.

Neste último mês da bolsa de IC foram finalizados os modelos que seriam efetivamente usados na plataforma do HuggingFace, esses modelos são baseados nos transformers que têm um custo computacional mais elevado. A plataforma, ilustrada na Figura 1, funciona da seguinte maneira:

1. Um texto de cunho financeiro é inserido na área delimitada para entrada do usuário, espera-se que seja inserido um texto com uma proposta financeira e em português.
2. Em seguida o usuário deve selecionar qual modelo deseja utilizar para a tarefa de Tradução Automática, Análise de Sentimentos e Classificação ZeroShot.
3. Após clicar no botão Gerar Análises! é iniciado o programa.
4. O texto em português é convertido para uma variável agora em língua inglesa, essa variável além de ser uma das saídas do sistema é também usada futuramente nos outros módulos. Essa conversão é feita utilizando ou o TextBlob (por padrão) ou o modelo transformer capaz de realizar a Tradução Automática.
5. Após isso é realizada a tarefa de Análise de Sentimentos. Essa tarefa visa classificar se a polaridade do texto é Positiva, Negativa ou Neutra; dessa forma é usado ou o VaderSentiment (por padrão) ou um modelo de transformer que foi pré-treinado ou passou por fine-tuning na tarefa de análise de sentimentos.
6. Para a Classificação em qual área financeira o texto está inserido, primeiramente passa-se o texto por um método heurístico que analisa se existe alguma menção direta a uma empresa; se houver é atribuído o Tema do texto à área de atuação da empresa -- por exemplo, se houver menção à Raia Drogasil (empresa do ramo farmacêutico) o texto é então classificado como pertencendo ao Tema de

Saúde, pois Raia Drogasil é uma empresa desse ramo. As empresas e suas classificações foram retiradas de uma planilha de classificação setorial da B3.

7. Por fim, caso não seja achada uma menção explícita de empresa, seja por falta de uma ou por aparição de uma empresa não listada no banco de dados, é usado um modelo de Classificação ZeroShot (definido qual deve ser usado pelo usuário) que tentará classificar qual a classificação setorial do texto de entrada entre: "Bens Industriais", "Comunicações", "Consumo Cíclico", "Consumo Não-Cíclico", "Finanças", "Materiais Básicos", "Óleo, Gás e Biocombustíveis", "Saúde", "Tecnologia da Informação" ou "Utilidade Pública".
8. Assim que todas essas etapas forem processadas a saída para o usuário aparece.
9. Adicionalmente existe na ferramenta uma visualização em WordCloud (vide Figura 3) do dataset usado para validação dos modelos, em que, analogamente, o usuário seleciona qual filtro de Tema ele deseja e aparece em sua tela a WordCloud correspondente às palavras mais vistas dependendo do tema (para a geração dessas WordClouds, os textos do dataset foram filtrados por tema, removidas as stopwords e em seguida gerada previamente a imagem da figura mostrada ao usuário).

Explicando mais a fundo cada um dos modelos por trás das funcionalidades, temos:

- Para Tradução Automática: Esta tarefa visa traduzir um texto de entrada de uma língua para outra, no caso a língua de partida é o português e a língua de destino é o inglês. Os modelos selecionados são:
  - TextBlob: Biblioteca para processamento textual capaz de traduzir textos, sem custo computacional muito elevado.
  - M2M100: Modelo transformer multilingual para a tarefa de sequence-to-sequence, capaz de traduzir mais de 100 línguas entre si.
  - OPUS: Modelo que utiliza transformers para traduzir várias línguas diretamente para o inglês.
  - T5: Modelo transformer da Unicamp capaz de realizar a tradução PT→EN utilizando um hardware modesto.
  - mBART: Modelo transformer especializado na tradução PT→EN criado por finetuning do mBART original (que é multilingual).
- Para Análise de Sentimentos: Tarefa que visa classificar o texto de acordo com o sentimento que ele transmite em Positivo, Negativo ou Neutro.
  - VADER: Biblioteca python que infere o sentimento baseado no léxico presente no texto.
  - FinBERT: Modelo transformer pré-treinado para analisar o sentimento de textos financeiros, construído através do finetuning do BERT para um domínio financeiro.
  - DistilBERT: Modelo transformer especializado na tarefa de análise de sentimentos, desenvolvido pelo próprio HuggingFace Transformers e usado como modelo padrão para esta tarefa.
  - BERT: Modelo transformer do BERT que passou pelo finetuning em reviews de produtos, para aferir uma análise de sentimento desses textos.
- Para Classificação ZeroShot: Tarefa em que o modelo não é treinado com os dados que ele deve classificar, porém é fornecido um contexto adicional para o modelo poder inferir corretamente a classe.

- Inferência por Companhia presente no texto: Além dos modelos transfomers (descritos a seguir), para otimizar a inferência da classificação é utilizado um modelo Heurístico que averigua a presença de uma menção explícita a uma companhia no texto e, caso essa companhia esteja listada na base de dados, o texto é então classificado conforme a modalidade da empresa (como área de saúde, financeira, etc)
- RoBERTa: Modelo transformer multilingual que passou pelo finetuning do xlm-roberta usando um dataset de natural language inference (NLI)
- mDeBERTa: Modelo transformer pré-treinado pela Microsoft e finetuned para a tarefa de NLI
- DistilroBERTa: Modelo transformer que utiliza codificador cruzado para inferência de linguagem natural (NLI).

A plataforma visual foi desenvolvida utilizando os Spaces do próprio Hugging Face, que consiste num sistema de versionamento git capaz de comportar um app que, no caso dessa IC, foi desenvolvido utilizando a biblioteca StreamLit. A plataforma pode ser acessada através deste link.

Conclusão:

A plataforma gerada foi um sucesso pelo tempo hábil para implementação e mudanças de planos no meio do caminho da iniciação científica, sendo funcional e atendendo seu propósito de ser um ambiente onde dados textuais financeiros podem ser analisados de forma categórica utilizando diversos modelos. Da forma em que a plataforma foi implementada, os módulos são facilmente escaláveis caso novas aparições de transformers mais eficazes para as tarefas propostas, aparecerem, bastando alterar poucas linhas de código para adicioná-los. A interface final também ficou bastante amigável e intuitiva sobre como o usuário pode acessá-la, podendo ser usada para gerar múltiplas análises de textos financeiros de uma vez só no painel de inferências. Além do painel de inferências, é possível gerar WordClouds para cada uma das temáticas que foram classificadas no dataset de 50.000 linhas.

O maior problema da ferramenta é o fato de utilizar modelos pesados e que utilizam muito hardware para serem processados. Dessa forma, a execução de inferências quando não for utilizado o TextBlob ou o Vader (que não são transformers) e o texto não tiver nenhuma menção às empresas listadas demora cerca de 2 a 3 minutos, algo muito além de um tempo razoável de execução.

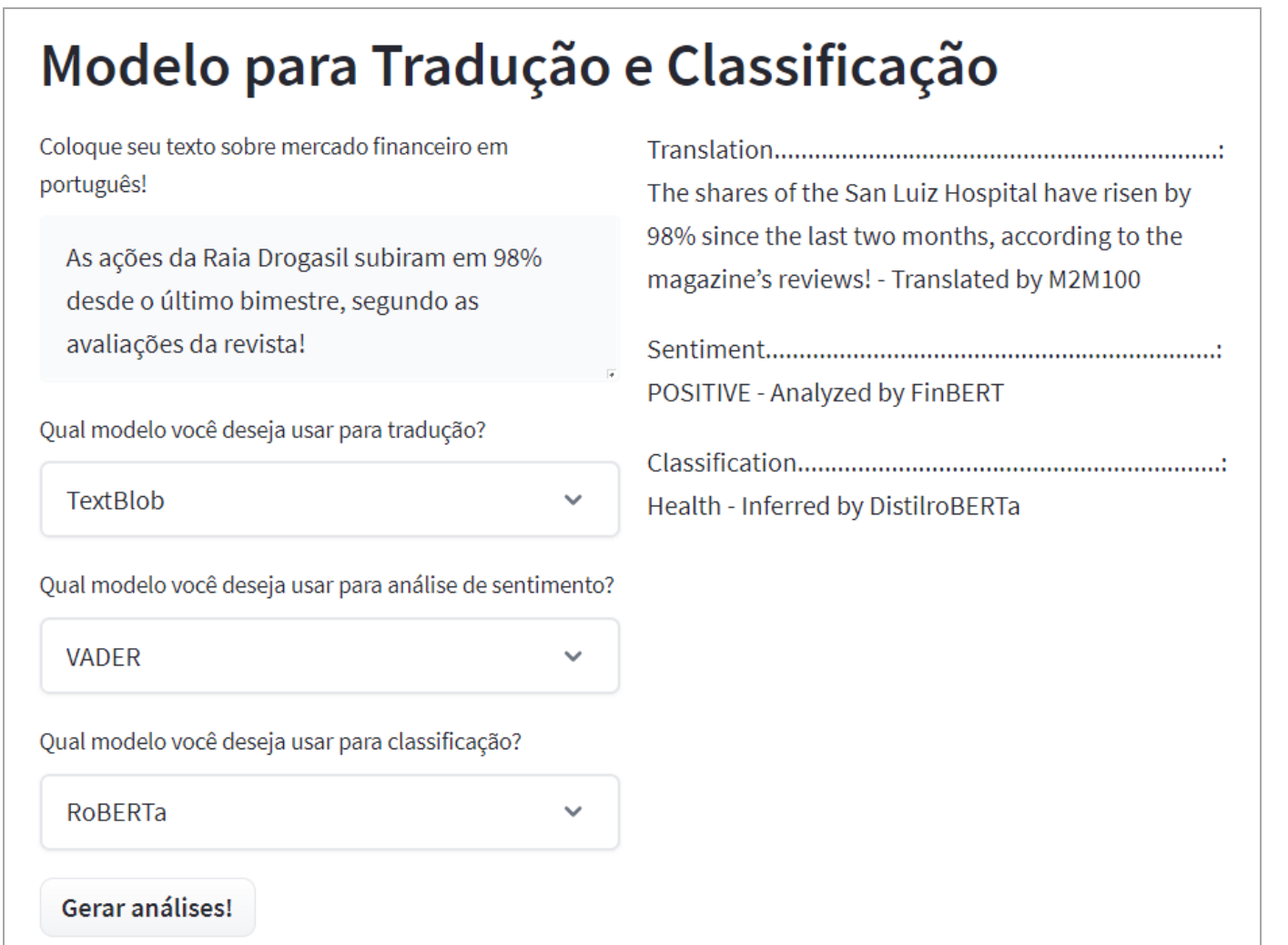

Figura 1 - Painel de classificação dos textos, que contém os diversos transformers e uma caixa para inserção do texto.

Figura 2 - Painel para a visualização do dataset utilizado.

Figura 3 - Painel para a geração de WordClouds.

Descrição das próximas atividades: Não Aplicável
Houve alteração significativa no tema ou prazo: (x) Sim ( ) Não Como relatado em relatórios anteriores, a base de dados utilizada não mais foi dos Diários da Assembleia da República Portuguesa, mas dados de textos financeiros. Além disso, como o aluno está saindo para um período de estudos no exterior, seu prazo final foi antecipado de 01/02/2023 para 31/08/2022. Essas mudanças, entretanto, não penalizaram a realização com total sucesso da pesquisa, que era desenvolver uma plataforma com diferentes ferramentas para processamento de textos em língua portuguesa.
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: Não se aplica (relatório final)
Resultados em relação às expectativas iniciais: (x) Acima ( ) Dentro ( ) Abaixo ( ) Muito abaixo
Previsão de conclusão no prazo: ( ) Sim (x) Não **Justifique em caso negativo:** vide comentários acima (o prazo de conclusão foi antecipado).
Protocolo
Data: 25/08/2022
Nome Completo da/o Bolsista: Enzo Bustos Da Silva