Projeto de Formatura – 2021

PCS - Departamento de Engenharia de Computação e Sistemas Digitais

Engenharia de Computação

Tema:

ZeroBERTo - Zero-Shot BERT based on Topic Modeling applied to the Portuguese Parliament

Introdução e Motivação

Democracias produzem grande quantidade de dados públicos e técnicas de Inteligência Artificial, em especial as técnicas de Processamento de Linguagem Natural (PLN), oferecem novas oportunidades para processar esses dados visando melhor informar cidadãos.

Apesar de PLN ser usado para interpretar grande volumes de textos e extrair informações relevantes de forma concisa, ela ainda encontra obstáculos, sobretudo quando se trabalha com uma língua de baixa disponibilidade de dados como o português, em um domínio restrito ou com textos muito longos.

Objetivo

O objetivo desse projeto é desenvolver um sistema que interprete e classifique textos longos em língua portuguesa. Os textos usados são de atas do Parlamento Português e os resultados são mostrados de forma simples e visual.

Modelo

Foram investigadas diferentes técnicas de *Topic Modeling* e a aplicabilidade do *Zero-Shot learning* com modelos *Transformers.* A partir disso, desenvolveu-se um modelo matemático que combina as etapas de *Topic Modeling*, como uma tarefa de aprendizagem de representação, e *Zero-Shot classification*, com os dados de entrada abstraídos na representação aprendida na etapa anterior, conforme ilustra a Fig.1.

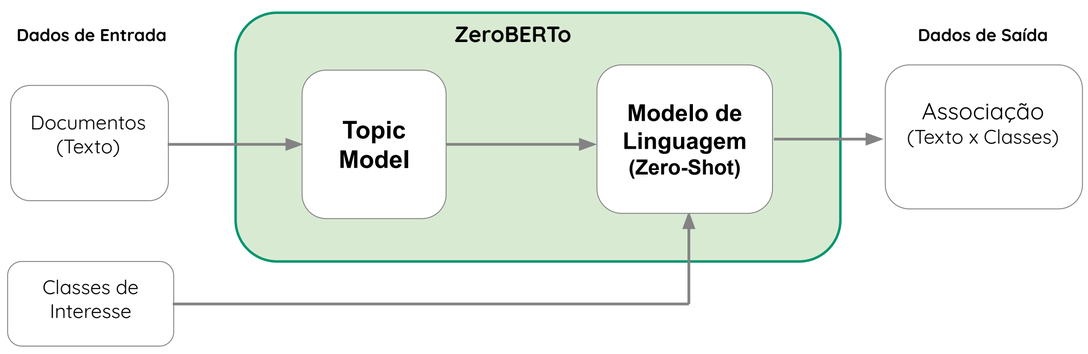

Figura 1 – Arquitetura do sistema desenvolvido, com *Topic Model* e *Zero-Shot*. Os dados de entrada são os documentos e as classes de interesse. A saída é a associação entre documentos e classes de interesse.

Base de dados: os dados foram obtidos por *web scraping* no site do Parlamento Português, preparados e estruturados em uma base de dados.

Resultados e Conclusão

Análise: o modelo ZeroBERTo foi avaliado inicialmente na base de dados folhaUOL, comparando-o com o modelo XLM-R, estado-da-arte em Inferência em PLN. Em 2 cenários de baixa disponibilidade de dados, ZeroBERTo atinge F1-score até 10% superior a XLM-R, apresentando tempo de execução até 14x menor, como mostra a tabela a seguir.

	XLM-R	ZeroBERTo	XLM-R	ZeroBERTo
P	0.47 ± 0.00	0.66 ± 0.01	0.46 ± 0.01	0.64 ± 0.01
R	0.43 ± 0.00	0.54 ± 0.01	0.43 ± 0.00	0.56 ± 0.02
F1	0.43 ± 0.00	0.54 ± 0.01	0.42 ± 0.01	0.52 ± 0.02
Time	61h30min	**9h21min**	15h22min	**1h10min**

Testes: ZeroBERTo foi então aplicado aos dados do Parlamento Português, classificando as falas dos parlamentares entre os temas de interesse público. O resultado do módulo de *Topic Modeling* está ilustrado na Fig. 2, na qual o verde mais escuro indica probabilidade maior de cada tópico (eixo horizontal) estar associado ao tema do eixo vertical.

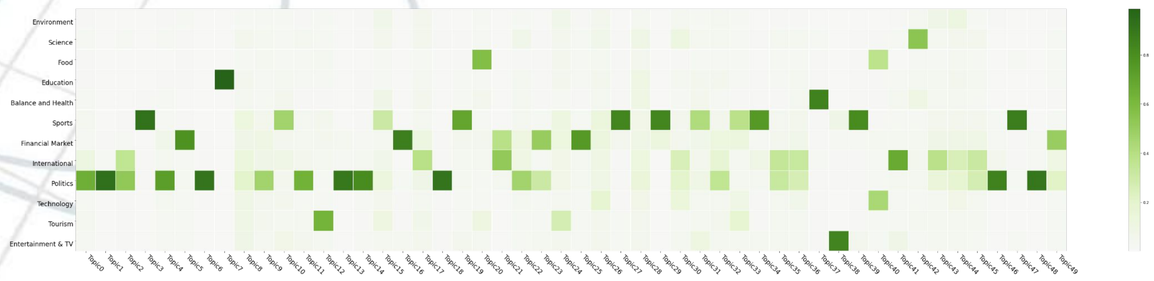

Figura 2 – Resultados ilustrativos do *Topic Model*.

Os resultados do ZeroBERTo aplicado à base de dados com as atas do Parlamento Português estão na Fig.3, com as barras de cada cor indicando os tópicos e respectivas palavras selecionadas e a classe de interesse sobre cada tópico (ex. Economia, Meio Ambiente, etc).

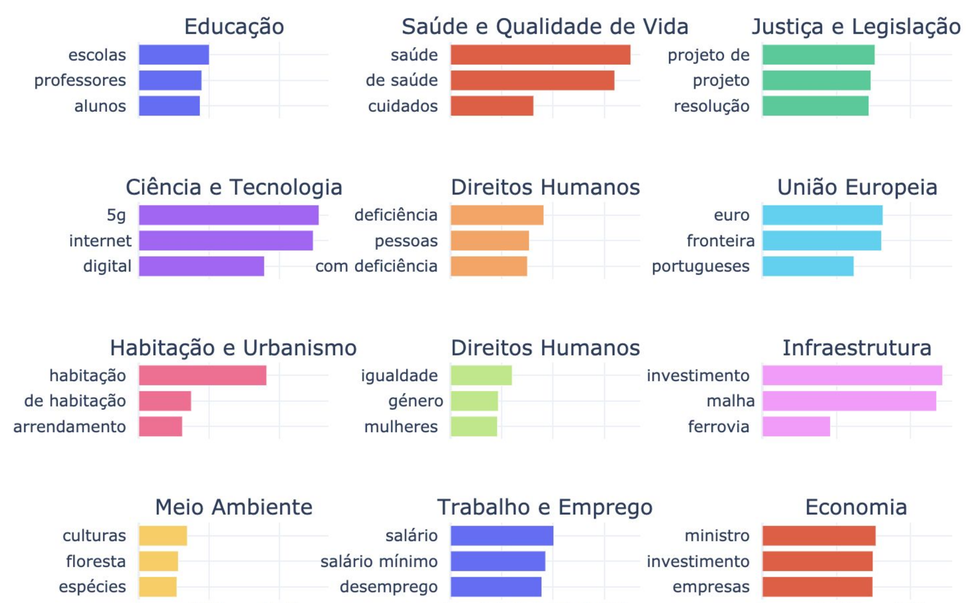

Figura 3 – Resultados ilustrativos do ZeroBERTo.

ZeroBERTo apresentou bons resultados, mostrando a associação dos documentos às classes de interesse. O modelo foi aceito e será publicado no PROPOR 2022 [1].

[1] Alcoforado, A., T. P. Ferraz, R. Gerber, E. Bustos, A. S. Oliveira, B. M. Veloso, F. L. Siqueira, and A. H. R. Costa (2022). ZeroBERTo - Leveraging Zero-Shot Text Classification by Topic Modeling. In International Conference on Computational Processing of the Portuguese Language.

Integrantes: - Alexandre Teodoro de Siqueira Guedes Alcoforado
- Rodrigo Elizardo Gerber

Professor(a) Orientador(a): - Profa. Dra. Anna Helena Reali Costa
Co-orientador(a): - Prof. Dr. Fábio Levy Siqueira