Projeto de Formatura – 2023

PCS - Departamento de Engenharia de Computação e Sistemas Digitais

Engenharia de Computação

Tema:

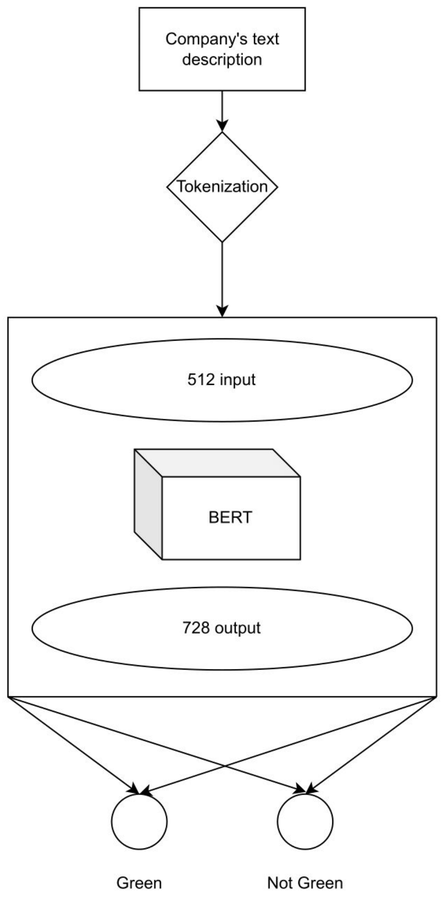
Método de classificação de empresas sustentáveis desenvolvido com NLP

Introdução

O trabalho propõe um método de classificação de empresas sustentáveis, combinando Aprendizado de Máquinas, especialmente NLP, e Empreendedorismo Verde. Essa abordagem inovadora visa avaliar o compromisso das empresas com práticas sustentáveis, respondendo à crescente importância do Aprendizado de Máquinas na análise de dados linguísticos e à necessidade do mercado por empreendimentos social e ambientalmente responsáveis.

Motivado por facilitar decisões de investidores sustentáveis, atrair talentos para empresas socialmente responsáveis e atender às expectativas dos consumidores por produtos de empresas sustentáveis, o trabalho busca contribuir para um ambiente de negócios mais ético e alinhado com as demandas contemporâneas da sociedade.

Modelo



O modelo proposto adota o BERT, um modelo pré-treinado disponibilizado online, que se destaca por sua eficácia em compreender contextos linguísticos complexos. Baseado na arquitetura Transformer, o BERT possui uma feature chamada de 'atenção', que permite ao modelo focalizar diferentes partes da entrada, capturando relações semânticas e melhorando a compreensão global do texto. Essa capacidade de atenção é fundamental para a eficácia do modelo na tarefa de classificação.

A abordagem de classificação utiliza a saída do BERT como entrada para uma rede neural que categoriza as empresas como 'verde' ou 'não verde'. O uso do BERT como extrator de características é relevante, considerando sua habilidade em compreender nuances semânticas e contextuais, contribuindo para decisões de classificação mais precisas.

Quanto ao dataset, a fonte proveniente do site Crunchbase oferece uma ampla variedade de descrições de empresas. No entanto, é importante observar que os rótulos de classificação são derivados das categorias selecionadas pelas próprias empresas durante o cadastro no Crunchbase, introduzindo a possibilidade de viés. O dataset inclui duas versões de descrição, uma longa e outra curta, e o modelo foi treinado em três configurações distintas: utilizando as 32 primeiras palavras da descrição longa, as 128 primeiras palavras da descrição longa e a descrição curta completa (67 palavras ao máximo).

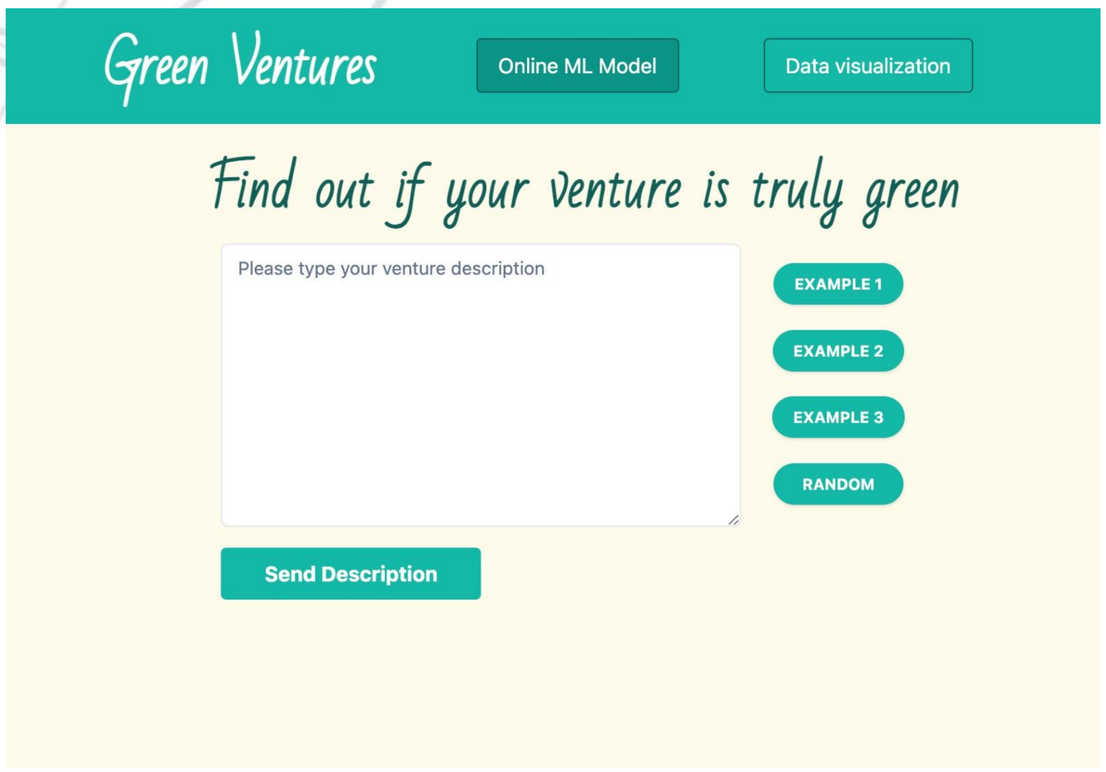
Essa abordagem visa explorar diferentes resultados do modelo e comparar para chegar no melhor resultado.

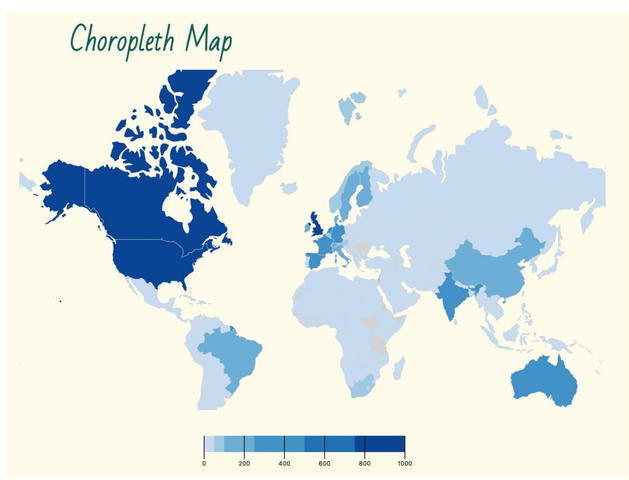
Model	Accuracy	Training Time
Big Description (128 words maximum length)	96%	1h25min
Big Description (32 words maximum length)	91%	27min
Short Description (67 words maximum length)	94%	54min

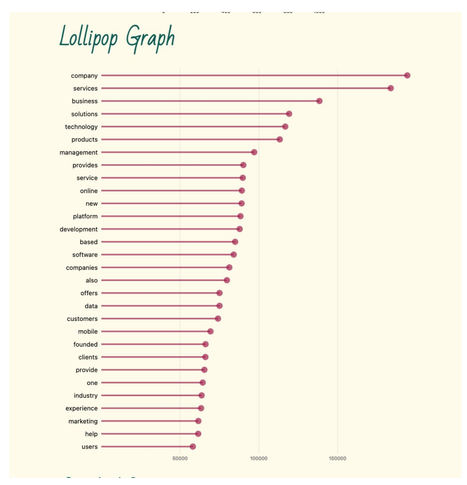
O melhor modelo, como esperado, foi o de 128 palavras com 96% de acurácia. Quanto mais dados, melhor o resultado em troca de tempo de treino.

Interface e Visualização

Para que se possa utilizar tal modelo, foi desenvolvido também uma interface para qual o usuário pode digitar prompts para o modelo. Nesta, o usuário pode escolher exemplos pré determinados, ou inserir a descrição de sua própria empresa. Além disso, foram adicionados visualizações do dataset, fornecendo ao usuário informação sobre sustentabilidade.







Integrantes: Marcelo Dias de Oliveira Fernandes
Rafael Rosa Rahal

Professor(a) Orientador(a): Prof. Dr. Pedro Corrêa