Classificador de prompts para banir manipulação de LLMs por meio de comandos maliciosos

Nicole Sarvasi Alves da Costa (INSPER, São Paulo, Brasil)

Introdução

Esse produto resolve o problema de manipulação para mau uso de Large Language Models (LLMs), como o GPT 3.5. Essa camada de segurança previne que usuários de LLMs, por meio de *prompt engineering*, consigam desbloquear respostas de tópicos tais como atividades ilegais e conteúdo adulto.

Ele atende a todo o público que deseja disponibilizar alguma aplicação segura que faça uso de alguma LLM, por exemplo aplicações para escolas, até mesmo para desenvolvimento de chatbots.

Utilizando-se de técnicas de NLP e machine learning, o modelo vai aprender, por meio de prompts tabelados quais tem caráter malicioso e quais não, por meio da análise do conteúdo de seus textos.

Metodologia

Com o escopo do produto definido, foi o momento de implementar um processo de desenvolvimento. Começando com a construção da base de dados que, devido a atualidade do tema, ainda não tinham modelos disponíveis na web, portanto foi feita uma curadoria de prompts maliciosos e não maliciosos, totalizando num *database*, composto por 2 colunas, uma com o conteúdo do prompt e outra com a sua classe, de aproximadamente 230 linhas, número que se provou suficiente pela variedade de prompts maliciosos.

A próxima etapa foi baixar a base de dados e fazer um pré-processamento básico para retirar os valores Nan e conferir o balanceamento do database. Com isso feito foram feitas 4 estratégias de machine learning com o intuito de serem comparadas conforme sua acurácia e observando as seguintes características: tempo de inferência, tamanho do modelo e memória RAM usada por cada um a ser carregado na memória.

As quatro abordagens variam de modelos mais tradicionais até modelos de zero-shot learning. As estratégias utilizadas foram: Count Vectorizer + Logistic Regression, Embedding Softmax Neural Network, Universal Sentence Encoder Multilingual Large + Logistic Regression e Zero-Shot Bart Large MNLI.

Resultados

Por meio das análises de acurácia, tempo de inferência e memória feitas com base nos Gráficos 1 e 2, o modelo baseline de regressão logística seria considerado a melhor pedida, uma vez que tem a segunda melhor acurácia e esta não passa de um desvio, e também ocupa e consume menos memória

Porém considerando que é um produto de segurança, o modelo que utiliza um *embedding* pré-treinado e uma regressão logística se destaca por abranger sua compreensão em 16 línguas. Além de ter a melhor acurácia e um tempo de inferência razoável.

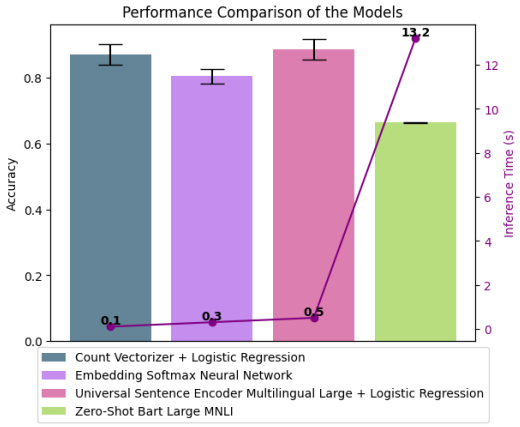

Gráfico 1

Algo interessante de se observar em relação ao tamanho dos modelos e memória RAM utilizada por cada modelo é que os modelos que utilizam algum tipo de *transfer learning* são muito maiores. Mas entre os dois, o modelo que possui apenas o pré-processamento pré-treinado utiliza muito menos RAM.

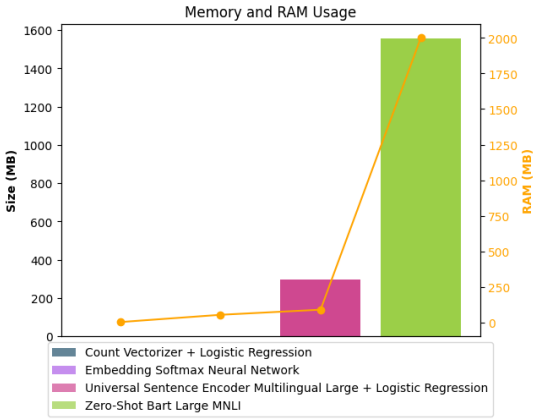

Gráfico 2

Adicionalmente foram mapeados os bigramas mais recorrentes em predições corretas e incorretas. Com estes valores é possível compreender melhor o funcionamento do modelo e futuramente melhorá-lo.

Portanto, foi possível elaborar, desenvolver, avaliar e entregar um produto de segurança de manipulação de LLMs baseado em IA e NLP. Uma demonstração da API foi disponibilizada para demonstração por meio de uma página HTML/Javascript.

Link do repositório: https://github.com/nicolecosta/nlpnicole/tree/main/AF_NicoleSarvasi