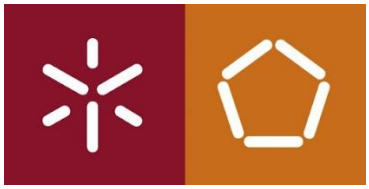
UNIVERSIDADE DO MINHO

ESCOLA DE ENGENHARIA

MESTRADO EM ENGENHARIA INFORMÁTICA

MINERAÇÃO DE DADOS



Tema: Modelo Baseado em LLM para Discutir o Programa Eleitoral da Coligação AD (Aliança Democrática)

Alunos:
PG50944 Fautisno Sachimuco - MMC,
PG50008 Marcos André Mussungo - MEI
PG52762 Lívia Pérez Bettero – MHD

Maio, 2024

Sumário

Objectivo Geral

Desenvolver e avaliar um modelo baseado em Linguagem Natural (LLM) para discutir com o usuário o as propostas de governo da colicação Aliança Democrática, com base em seu Programa Eleitoral apresentado nas eleições de 2024 - conteúdo do site oficial, vídeos do canal oficial e reportagens de sites confiáveis - de forma ética.

Fonte de Dados

A fonte de dados para este projeto foi constituída a partir das seguintes fontes: . Programa Eleitoral em PDF . Web Scrapping de conteúdo do site oficial . Web Scrapping de conteúdo de publicações em sites de notícias considerados confiáveis

Metodologia

Scrapping de Dados

Fontes de Dados, Bibliotecas e Ferramentas

Foram selecionados como fontes de dados os seguintes: ##### Site oficial da colicação Aliança Democrática, considerando textos do site e outros arquivos como o pdf com o proposta de governo Utilisou-se o wget para baixar e analisar o que se podia extrair do site inicialmnete. Como comando, foi baixado o arquivo PDF da proposta de governo e algumas poucas páginas em formato html. A extração dos textos do arquivo PDF contendo a proposta de governo do AD foi feita dentro da função *pdf_scrapping()* do código em python, com a biblioteca **PyPDF2** (https://pypi.org/project/pypdf/#description). Foram removidos caracteres como "•" e quebras de linha no meio de frases. Já os textos em formato html do site, que incluíam hino, textos de apoio e notícias, foram extraídos na função *extrai_oficial()* usando as bibliotecas **jjcli** (https://pypi.org/project/jjcli/) e **BeautifulSoup** (https://pypi.org/project/beautifulsoup4/).

Sites de notícias de grandes veículos, considerados confiáveis. Foi criada uma lista de sites indicados para extração dos arquivos. As noticias selecionadas passaram por análise humana devido á preocupação com questões éticas que envolvem o tipo de chatbot que seria treinado.

```
novo - reportagem - https://www.cnnbrasil.com.br/internacional/chega-nao-ganhou-as-eleicoes-em-portugal-mas-e-o-maior-vitorioso-da-noite/
novo - reportagem - https://www.cnnbrasil.com.br/internacional/portugal-vai-as-urnas-em-eleicao-que-ocorre-dois-anos-antes-do-previsto/
novo - reportagem - https://rr.sapo.pt/noticia/politica/2024/02/13/alianca-democratica-apresenta-programa-eleitoral-queremos-virar-a-pagina-do-
desespero/366201/
novo - reportagem - https://www.voaportugues.com/a/portugal-l%C3%ADder-da-alian%C3%A7a-democr%C3%A1tica-sem-maioria-assume-que-far%C3%A1-
governo/7522196.html
novo - reportagem - https://www.rfi.fr/pt/mundo/20240311-alian%C3%A7a-democr%C3%A1tica-ganha-por-pouco-e-chega-consegue-eleger-48-deputados
novo - reportagem - https://www.dn.pt/3040673111/projecoes-dao-vitoria-a-alianca-democratica-e-grande-crescimento-do-chega/
novo - reportagem - https://www.rtp.pt/noticias/politica/alianca-democratica-em-busca-de-uma-efetiva-mudanca-politica_es1552304
novo - reportagem - https://observador.pt/2024/02/05/alianca-democratica-vence-mas-fica-a-tres-deputados-da-maioria/
```

O código para extração dos textos seguiu o mesmo formato utilizado no site oficial mas, por se tratar de sites com estruturas difertenes, foi necessário criar um novo arquivo contendo as referências de que tags deveriam ser extraídas e que tags deveriam ser desconsideradas pelo script para cada site diferente (CNN, SAPO, etc) Os textos passaram também por uma

limpeza para remoção de linhas em branco e caracteres especiais e pequenos blocos de texto indesejados como “Leia Mais”, “&quot” e outros.

Vídeos oficiais da colicação disponíveis no canal do Youtube Para os vídeos oficiais disponíveis no youtube, inicialmente cogitou-se fazer download dos audios dos videos com a API do Youtube, seguida da transcrição com a API do Speech to Text do Google, mas ao final, a ferramenta Download Youtube Subtitles (https://www.downloadyoutubesubtitles.com/) resultou mais ágil para o volume de textos que se precisava baixar. Os textos foram gravados em arquivos individuais e postos em uma pasta para tratamento via script onde foram remividas quebras de linhas. Ao fim do scarapping, os textos extraídos foram gravados no mesmo arquiv, “SAIDA.txt”

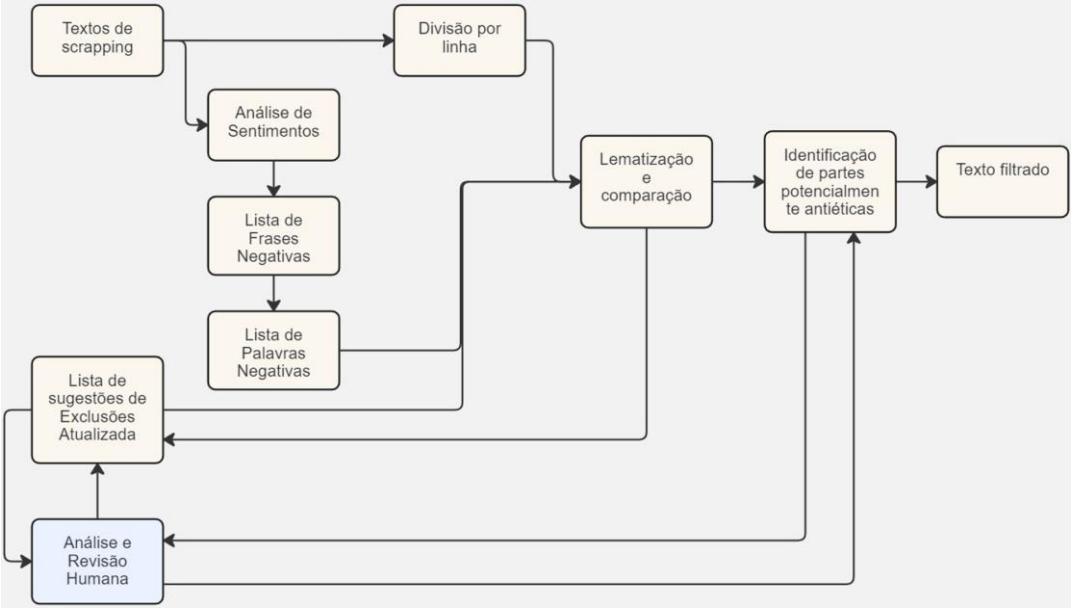


Figura 1 Fluxo de scapping completo com filtro ético

Filtragem de Conteúdo

Mesmo nos textos extraídos de sites confiáveis e de publições oficiais da colicação AD, foi possível identificar partes de discursos que precisariam ser filtradas por questões éticas, principalmente nas transcrições de vídeos de candidatos do AD no youtube. Analisando os textos, foram encontradas falas como: - *“Eu sei que o líder do PS adora automóveis de luxo – agora parece que os esconde (...)”*

NLTK (Natural Language Toolkit)

Embora não seja necessariamente uma regra, sabe-se que a partir da sinalização da polaridade de palavras do texto pode auxiliar na identificação de linguagem negativa associada a discursos de ódio, assédio ou criscriminatórios e, consequentemente, a conteúdo anti-ético. No caso dos textos analisados, algumas palavras poderiam acusar falsos negativos ou positivos. Palavras como “Socialista” ou “Pedro Nuno”, por exemplo, tomam peso negativo nos textos devido ao contexto, visto que representam os opositores da colicação AD nas

eleições de 2024. Assim, foi criada uma lista contendo termos que devem ser considerados negativos, independente da polaridade sinalizada com a biblioteca NLTK.

A partir da lista de palavras que obrigarotiamente devem ser consideradas negativas e da sinalização da polaridade das demais palavras, definiu-se o limite mínimo de três palavras negativas para que determinada frase pudesse ser considerada potencialmente anti-ética, como po exemplo: A frase: "O texto acusa ainda a governação socialista se ter caracterizado pela intromissão na gestão e relações acionistas de empresas privadas e até pelo enfraquecimento e tentativa de dominação das instituições independentes de regulação económica e de justiça." é claramente o tipo de discurso que se deseja evitar que o chatbot aprenda, no entanto, teve score de sentimento = 0.0 (neutro) ao utilizar apenas a bilbioteca NLTK.

Uma vez verificada a primeira classificação das frases, optou-se por criar uma lista de palavras que deveriam ser consideradas, obrigatoriamente negativas dentro do tema do projeto. Para isto, incluiu-se na lista, prinripalmente palavras e nomes que se associassem aa colicação de oposção do AD, como "Pedro Nuno", [Governo] "anterior", "socialista", "socialismo", etc. Neste ponto, entende-se fundamental ter conhecimento dos textos e do contexto político Português atual.

Spacy (Industrial-Strength Natural Language Processing)

A lematização de palavras foi adotada para facilitar a identificação dos termos da lista de palavras obrigatoriamente negativas em cada frase analisada. Após alguns testes considerando a contagem de lemas negativos por frase, definiu-se um limite mínimo de três lemas negativos combinados em uma única frase para considerá-la potencialmente entiética, portanto, filtrável do texto original. Com esta função, foram identificadas e removidas, frases como: - *"(...) comparar o doutor Nuno Santos com o professor Cavaco Silva é comparar um Ferrari com um calhambeque encostado numa garagem (...)"* - *"(...) a colicação socialista e Pedro Nuno Santos que enquanto Ministro mostrou tudo menos preparação."* - *"(...) a experiência do líder da colicação Socialista está marcada por exemplos dessa cultura de informalidade (...)"*

O texto filtrado foi gravado em um arquivo *TEXTO_SAIDA_FILTRADO.txt* e o processo de Scrapping, definido da seguinte forma:

Construção do Modelo

Tokenização

Tokenização é o processo de dividir um texto em unidades menores chamadas "tokens". Esses tokens podem ser palavras, subpalavras, caracteres ou até símbolos específicos, dependendo do esquema de tokenização utilizado. A tokenização é um passo fundamental no processamento de texto porque facilita a manipulação e a análise do texto em um formato estruturado.

tokenizer = tiktoken.get_encoding("cl100k_base")

usamos a função get_encoding da biblioteca tiktoken com paramêtro (cl100k_base) como esquema de codificação ou modelo de tokenização.

cl100k_base

Este esquema de codificação possui uma base de cerca de 100.000 tokens. Esses tokens podem incluir palavras inteiras, subpalavras, caracteres individuais, ou mesmo sequências de caracteres comuns, E é baseado no mêtodo de codificação de pares de bytes (BPE). Ele é reversível e sem perdas de informações, oque significa que os tokens podem ser convertidos de volta para o texto original, comprime o texto, tornando a sequência de tokens mais curta do que os bytes correspondentes ao texto original.

Função para leitura da nossa API

Para leitura da API da OpenAI, foi criada a função abaixo. Esta é uma API pessoal que, para qualquer teste dos modelos treinados pela OpenAI, é obrigatório o uso de API para requisições de acesso aos endpoints do modelo.

```
def read_openai_api_key():
with open('openai_chave.txt', 'r') as file:
    api_key = file.read().strip()
return api_key
my_api_key = read_openai_api_key()
```

Embeddinging

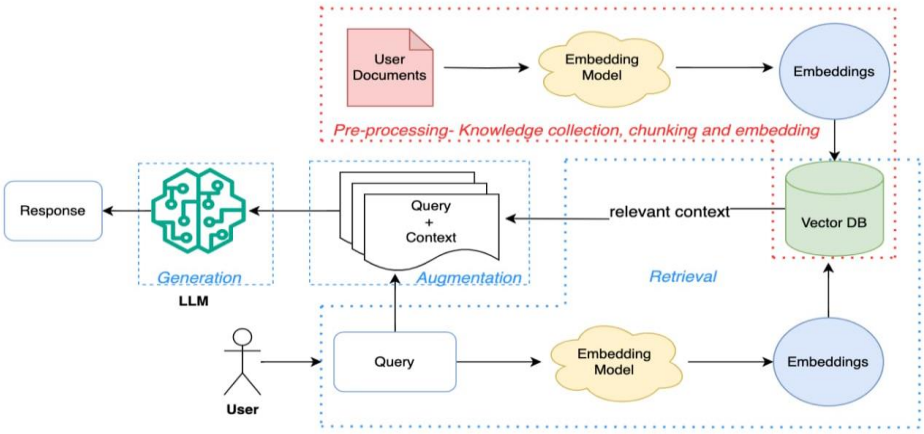


Figura 2 RAG Architecture. Fonte: Mindful Matrix

Embeddings são vetores ou matrizes de números que representam o significado e o contexto dos tokens processados pelo modelo, são usados para codificar e decodificar os textos de entrada e saída, usamos o modelo "text-embedding-3-small" , por ser menor e muito eficiente, sendo uma versão melhorada do modelo "text-embedding-ada-002".

Função Resposta_Pergunta

Usamos o modelo GPT-3.5 Turbo Instruct que é uma versão melhorada do GPT-3 (Generative Pre-trained Transformer 3) segue o estilo do InstructGPT, o que significa que é otimizado para seguir instruções específicas, ao contrário de alguns modelos maiores, o GPT-3.5 Turbo Instruct suporta apenas uma janela de contexto de 4.000 tokens. Isso significa que ele considera apenas os últimos 4.000 tokens do texto para gerar suas respostas.

O custo do GPT-3.5 Turbo Instruct é de USD 1,50 por 1 milhão de tokens para entrada e USD 2,00 por 1 milhão de tokens para saída.

paramêtros:

Temperatura: é um parâmetro que controla a aleatoriedade na geração de texto pelo modelo. Valores mais altos de temperatura (por exemplo, 0,8) tornam as saídas mais aleatórias e criativas, enquanto valores mais baixos (por exemplo, 0,2) tornam as saídas mais determinísticas e focadas.

Top-p (Penalização de Probabilidade):

O parâmetro "top-p" (também conhecido como "nucleus sampling") controla a probabilidade cumulativa das palavras geradas. Valores mais altos de "top-p" (por exemplo, 0,9) incluirão mais palavras no conjunto de saída, enquanto valores mais baixos (por exemplo, 0,3) restringirão a saída a um conjunto menor de palavras mais prováveis.

Penalização de Frequência e Penalização de Presença:

A penalização de frequência (frequency penalty) controla a repetição excessiva de palavras. Um valor maior reduz a repetição. A penalização de presença (presence penalty) controla a diversidade das palavras usadas. Um valor maior aumenta a diversidade.

Testes e Validação

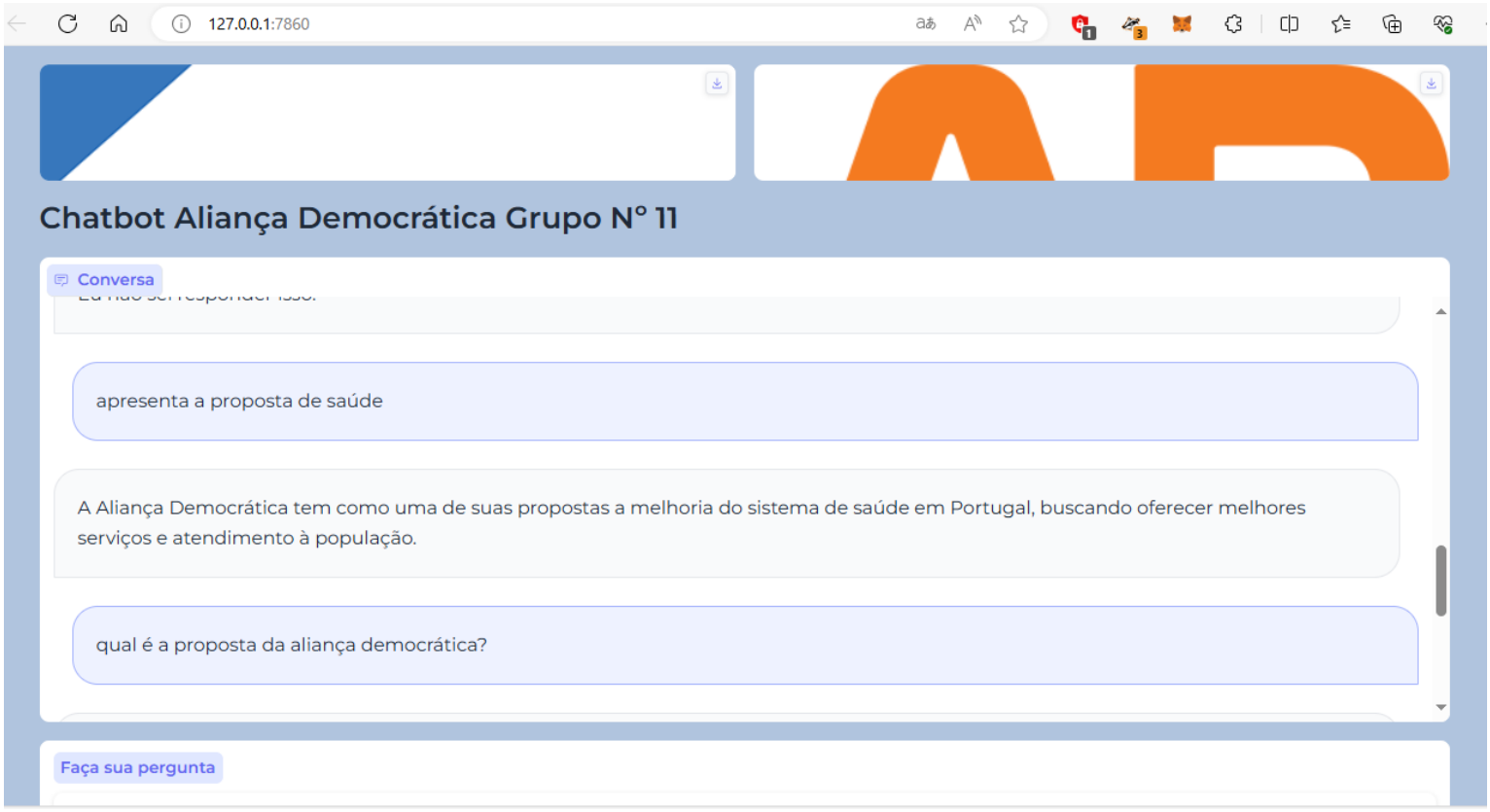


Figura 3 Resposta do chatbot sobre a proposta da AD para saúde

Figura 4 Resposta do chatbot sobre a proposta geral da AD

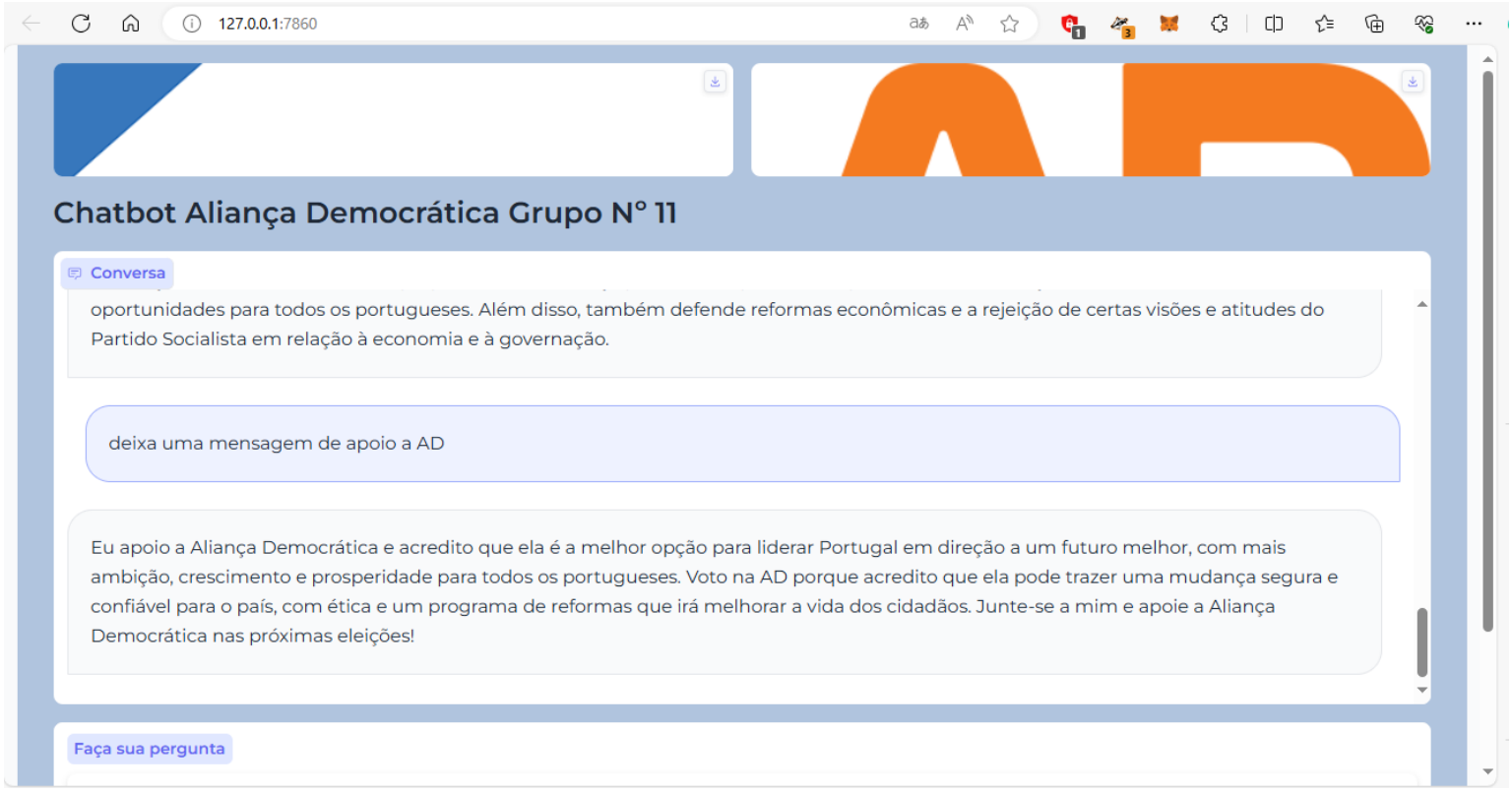


Figura 5 Mensagem de apoio do chatbot à AD

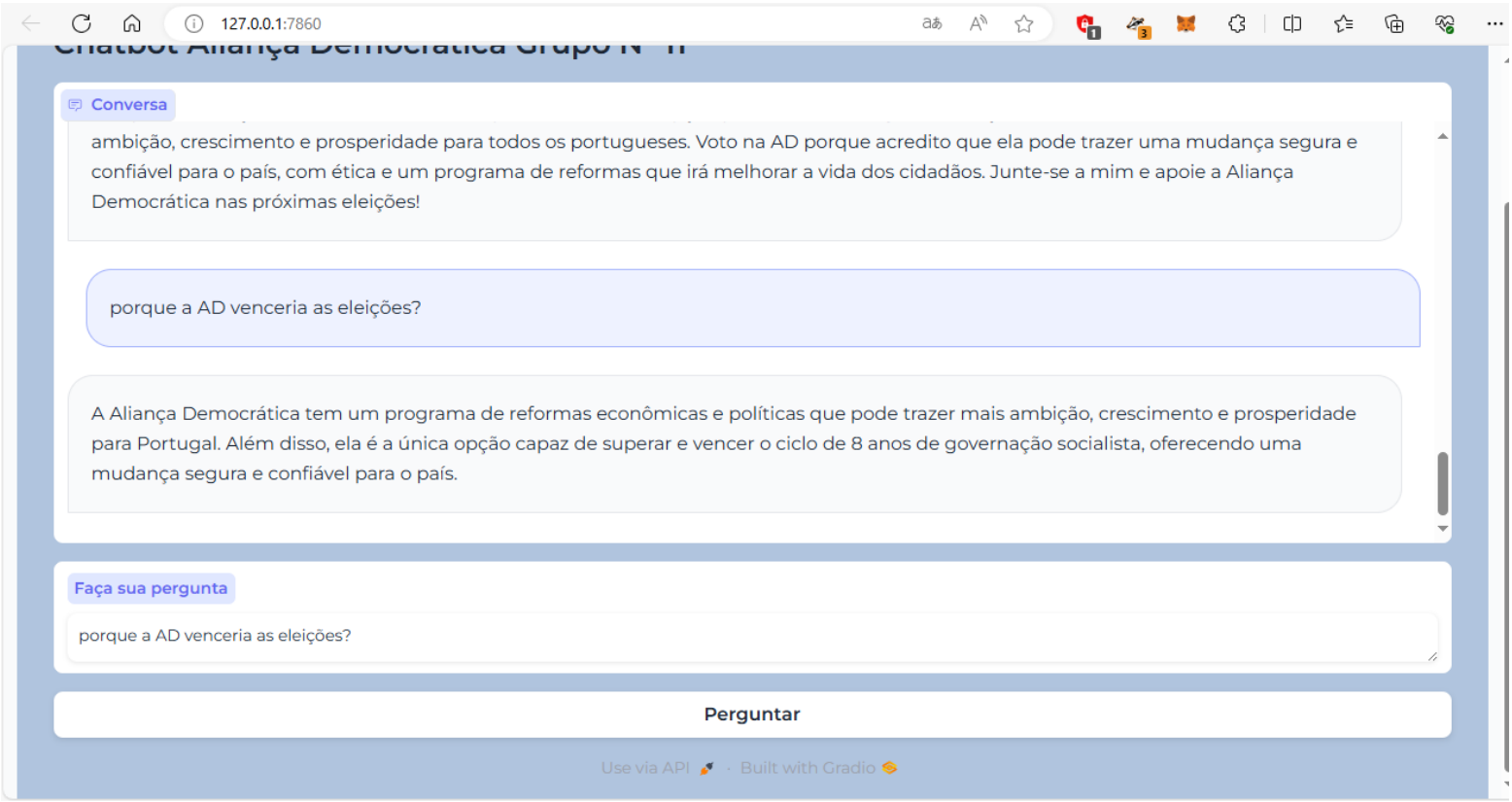


Figura 6 Resposta do chatbot sobre quem venceria as eleições

Futuros Desafios

Gerenciamento de Contexto Dinâmico:

Lidar com o contexto dinâmico é um desafio, especialmente em conversações longas ou em situações onde o contexto muda rapidamente. Desenvolver estratégias para manter e atualizar o contexto de forma eficiente melhora a capacidade do modelo de compreender e responder adequadamente às perguntas.

Avaliação de Usuários e Retroalimentação:

Coletar feedback dos usuários e realizar avaliações de usabilidade para entender como o modelo está sendo utilizado na prática e identificar áreas de melhoria. Incorporar mecanismos de feedback e análise de métricas de desempenho.

Referências Bibliográficas

https://ad2024.pt
https://ad2024.pt/pdf/ad-programa-eleitoral.pdf
https://www.nltk.org/
https://spacy.io/

CNN BRASIL. Aliança Democrática deve vencer eleições legislativas, aponta projeção da CNN. Disponível em: https://www.cnnbrasil.com.br/internacional/alianca-democratica-deve-vencer-eleicoes-legislativas-aponta-projecao-da-cnn/. Acesso em: 21 maio 2024.

DIÁRIO DE NOTÍCIAS. Projeções dão vitória à Aliança Democrática e grande crescimento do Chega. Disponível em: https://www.dn.pt/3040673111/projecoes-dao-vitoria-a-alianca-democratica-e-grande-crescimento-do-chega/. Acesso em: 21 maio 2024.

RÁDIO FRANÇA INTERNACIONAL. Aliança Democrática ganha por pouco e Chega consegue eleger 48 deputados. Disponível em: https://www.rfi.fr/pt/mundo/20240311-alianca-democratica-ganha-por-pouco-e-chega-consegue-eleger-48-deputados. Acesso em: 21 maio 2024.

RTP. Aliança Democrática em busca de uma efetiva mudança política. Disponível em: https://www.rtp.pt/noticias/politica/alianca-democratica-em-busca-de-uma-efetiva-mudanca-politica_es1552304. Acesso em: 21 maio 2024.

SIC NOTÍCIAS. Aliança Democrática 2.0: reescrever o passado e eclipsar o presente. Disponível em: https://sicnoticias.pt/pais/2024-01-10-Alianca-Democratica-2.0-reescrever-o-passado-e-eclipsar-o-presente-7e81edb2. Acesso em: 21 maio 2024.

VISÃO. AD 3.0: as histórias e memórias da Aliança Democrática que juntou os aliados mais fiáveis. Disponível em: https://visao.pt/atualidade/politica/2023-12-15-ad-3-0-as-historias-e-memorias-da-alianca-democratica-que-juntou-os-aliados-mais-fiaveis/. Acesso em: 21 maio 2024.

VOZ DA AMÉRICA. Portugal: Líder da Aliança Democrática sem maioria assume que fará governo. Disponível em: https://www.voaportugues.com/a/portugal-l%C3%ADder-da-alian%C3%A7a-democr%C3%A1tica-sem-maioria-assume-que-far%C3%A1-governo/7522196.html. Acesso em: 21 maio 2024.