

LLM - Chat sobre partidos políticos portugueses

Ana Rita Poças, Henrique Alvelos, Nelson Almeida

pg53645@alunos.uminho.pt, pg50414@alunos.uminho.pt, pg52697@alunos.uminho.pt

Abstract—Um chatbot que responde a perguntas sobre os partidos políticos portugueses tendo em conta o seu programa eleitoral das eleições legislativas de 2024. Extraímos e processamos informações dos manifestos dos partidos, principalmente disponíveis em formato PDF. Utilizando Python, convertimos esses PDFs em arquivos .txt e empregamos técnicas de processamento de linguagem natural (NLP) para categorizar as medidas políticas por tema. A preparação dos dados incluiu a divisão de documentos em secções coerentes usando o **spaCy**, seguida pela geração de embeddings com o modelo BGE através da API da HuggingFace. Esses embeddings foram armazenados no Pinecone, uma base de dados vetorial otimizada para buscas por similaridade. O mecanismo de recuperação utiliza a seleção de documentos top-k com sobreposição contextual e um reranker para melhorar a precisão. Para gerar respostas às consultas dos utilizadores, experimentamos vários modelos, selecionando o Mixtral-8x7B pela sua relação custo-desempenho. Avaliamos a eficácia do sistema utilizando a framework RAGAS, que mede precisão do contexto, recall, fidelidade e relevância. O frontend, desenvolvido com React, oferece uma interface intuitiva para consultas dos utilizadores e interação com o modelo de linguagem.

Index Terms—Chatbot, Inteligência Artificial, Large Language Model, Mineração de Dados, Partidos Políticos

I. INTRODUÇÃO

Este relatório surge no âmbito da Unidade Curricular de Mineração de Dados, do perfil de Engenharia de Conhecimento do Mestrado em Engenharia Informática da Universidade do Minho, na qual nos foi proposto o desenvolvimento de um *Large Language Model* (LLM) sobre um tópico escolhido pelo grupo. Decidimos criar um LLM que funcione como um chat sobre partidos políticos portugueses. O nosso objetivo é desenvolver um website funcional e simples onde os cidadãos possam esclarecer as suas dúvidas relativamente aos planos eleitorais dos diversos partidos.

Tendo em conta o problema do voto desinformado, e a falta de consumo de conteúdo político entre os cidadãos portugueses, quer por falta de interesse, quer por dificuldade de acesso à informação de forma simples, intuitiva e centralizada, temos grandes expectativas quanto à utilidade do nosso projeto. A nosso ver o tópico escolhido para o nosso LLM é de grande relevância visto que a informação política é essencial para a tomada de decisões informadas dos cidadãos, com especial importância durante os períodos eleitorais. Assim sendo, a nossa aplicação permite que, de forma simples e prática, os utilizadores obtenham informações verídicas e esclareedoras.

II. ESTADO DE ARTE

Nos dias de hoje, a investigação em *Large Language Models* (LLMs) tem avançado significativamente, com aplicações cada vez mais diversificadas em diferentes domínios. Os modelos

como GPT da OpenAI e BERT da Google têm demonstrado capacidades impressionantes na compreensão de linguagem natural e geração de texto coerente.

No contexto político, diversos estudos como [1] e [2] têm investigado o uso de LLMs para analisar discursos políticos, prever tendências eleitorais e facilitar a redação de documentos legislativos. Contudo, a criação de um LLM dedicado exclusivamente às políticas dos partidos portugueses ainda constitui uma lacuna tanto na literatura quanto nas aplicações práticas existentes até à data.

Assim, o nosso projeto diferencia-se ao concentrar-se na criação de uma aplicação que, não apenas responde a perguntas específicas sobre políticas partidárias, mas também integra um sistema robusto de consulta de dados estruturados e análise contextual. Com a utilização de uma solução baseada em RAG, poderemos devolver informações atualizadas sem ter de retrainar modelos de forma a que os resultados apresentados aos utilizadores sejam relevantes e atualizados.

III. FONTE DE DADOS

Os dados necessários à implementação da nossa solução concentram-se maioritariamente nos planos eleitorais dos partidos políticos portugueses. Ao escolhermos estas fontes de dados garantimos que estas informações são de facto as ideias que os partidos pretendem colocar em prática. Inicialmente pensamos ainda na utilização de scrapper das plataformas de comunicação social dos partidos tais como o Twitter, no entanto estas declarações podem não contemplar a totalidade das suas medidas.

Os dados selecionados, para efeitos de prova de conceito cingiram-se aos planos eleitorais dos diversos partidos para as Eleições Legislativas de 2024.

Para a obtenção dos dados, tivemos que lidar com o problema de que muitos planos políticos atuais apenas se encontravam disponíveis em formato PDF. Para resolver essa restrição, criamos um *script* Python que converte esses documentos PDF para o formato .txt. Este processo de conversão é essencial para a manipulação e análise dos dados textuais que estão dentro dos documentos.

Após a conversão dos PDFs em ficheiros .txt, realizamos um processo de *scraping* para extrair e organizar as informações relevantes. Este scraping foi feito de forma a separar as medidas políticas por tema, permitindo uma categorização eficiente das propostas. Desta maneira, conseguimos estruturar os dados de uma forma que facilita a sua integração no modelo de linguagem.

IV. TECNOLOGIAS UTILIZADAS

Pipeline RAG

Para o desenvolvimento do backend e da pipeline RAG utilizamos Python como a principal linguagem de programação. A sua versatilidade e a ampla gama de bibliotecas disponíveis tornam-no ideal para manipular dados, construir APIs, e integrar diferentes componentes do sistema.

Esta escolha permitiu-nos implementar a seguinte pipeline:

A. Recolha de dados

- Converter documentos PDF para o formato .txt utilizando a biblioteca PyPDF2.
- Implementar técnicas de *scraping* e de processamento de linguagem natural (NLP) para extrair informações relevantes dos planos políticos dos partidos. Criar e gerir uma API que permitiu a comunicação entre o frontend e o modelo de linguagem.
- Armazenar e estruturar os dados recolhidos de forma eficiente.

B. Preparação dos dados

De forma a gerar *embeddings* relevantes para a nossa solução, é necessário *dividir* os documentos recolhidos anteriormente em secções. Inicialmente optamos por uma abordagem de *chunking* a cada N tokens, no entanto isto trouxe problemas de *divisão* onde era perdido o contexto semântico ou sintático caso esta divisão se desse a meio de uma frase. Isto apresentou-se de extrema relevância no resultado final da nossa solução impactando diretamente o contexto dado ao modelo de linguagem. De forma a mitigar isto aplicamos uma metodologia de overlapping, isto é, a cada *divisão* de N tokens, avançamos apenas $N-K$ tokens de forma a que K tokens do contexto anterior fosse transposto para o seguinte.

Com esta abordagem obtivemos resultados mais satisfatórios, no entanto ainda era possível constatar que os *pedaços* de documento obtidos não tinham a melhor coerência semântica.

Optamos então por uma abordagem que perdura na solução final, que passa pela utilização de uma ferramenta de processamento de linguagem natural *spaCy* e utilizamos o modelo treinado em língua portuguesa para nos auxiliar no processo de *divisão*. Esta abordagem trouxe-nos os melhores resultados trazendo-nos a melhor coerência semântica nos *chunks* obtidos.

O próximo passo na nossa pipeline seria dar *embed* às nossas divisões dos documentos de forma a obtermos um vetor numérico que representasse esta mesma informação. Para esta tarefa utilizamos a API da *HuggingFace*, onde recorremos ao modelo de *embeddings* BGE.

C. Armazenamento de Dados

Para esta etapa escolhemos a base de dados *Pinecone* [3] visto ser otimizada para operações vetoriais disponibilizando diversos métodos de procura por similaridade nativamente tal como similaridade de cossenos ou distância euclidiana. A base de dados eleita trouxe-nos ainda vantagens tais como ter uma

instância *cloud* gratuita, o que nos permitiu poupar espaço de memória no ambiente de desenvolvimento e podermos aceder à informação remotamente.

Os vetores armazenados na base de dados contemplavam metadados relativamente ao partido aos quais pertenciam, esta adição tornou-se valiosa devido ao impacto no tempo necessário para a obtenção dos vetores da base de dados bem como a sua pertinência.

D. Retrieval

A seleção dos documentos a devolver pela nossa solução é ditada por dois métodos, seleção dos *top-k* documentos com mais similaridade + contexto adicional proveniente do overlap, e através da utilização um reranker, a eleição do melhor método será feita aquando da avaliação da solução.

E. Geração de respostas

Para o passo final da nossa pipeline (Generation), que seria a geração de respostas às queries do utilizador, experimentamos com diversos modelos tais como Mixtral-8x22B Instruct, Mixtral-8x7B Instruct, LLaMA-3, e ao longo das nossas análises constatamos que o modelo que nos trazia resultados mais satisfatórios e consistentes foi o modelo eleito. Para além da análise empírica proveniente apenas de resultados de um dataset de testes relativamente sucinto, a escolha do modelo passou também pela análise dos seguintes parâmetros:

Modelo	Parâmetros	Tamanho do Contexto	Preço p/ 1M Tokens
Mistral Instruct	7B	8192	\$0.20
LLaMA-3 Chat	8B	8192	\$0.20
Mixtral-8x22B Instruct	22B	65536	\$0.80

TABLE I

COMPARAÇÃO DE MODELOS BASEADO NO NÚMERO DE PARÂMETROS, TAMANHO DO CONTEXTO DISPONÍVEL E DO SEU PREÇO

Em suma, o modelo utilizado na nossa solução é o Mixtral-8x7B devido ao seu preço e resultados.

Para a utilização destes modelos optamos por consumir a API da *TogetherAI* [4] visto que providencia um vasto leque de moldes e extensa documentação de forma "gratuita".

A utilização de uma API face a modelos locais quantizados ou soluções como Ollama permitiu-nos não restringir a utilização da solução à performance da *máquina* onde está a ser executada.

O passo final da nossa pipeline RAG é a avaliação. Para avaliar o nosso modelo utilizamos duas metodologias, uma recorrendo à *framework* RAGAS, onde o modelo é testado quanto à **Context precision** - mede a relação sinal-ruído do contexto recuperado. Esta métrica é calculada utilizando a pergunta e os contextos.

Recall do contexto mede se toda a informação relevante necessária para responder à pergunta foi recuperada. Esta métrica é calculada com base na verdade de referência (ground truth) (esta é a única métrica no framework que depende de

rótulos de verdade de referência anotados por humanos) e os contextos.

A **fidelidade** mede a precisão factual da resposta gerada. O número de afirmações corretas dos contextos fornecidos é dividido pelo número total de afirmações na resposta gerada. Esta métrica usa a pergunta, os contextos e a resposta.

A **relevância da resposta** mede quão relevante é a resposta gerada em relação à pergunta. Esta métrica é calculada utilizando a pergunta e a resposta. Por exemplo, a resposta "A França está na Europa Ocidental." para a pergunta "Onde está a França e qual é a sua capital?" teria uma baixa relevância de resposta porque responde apenas metade da pergunta.

A outra metodologia utilizada consiste no uso de outro LLM onde através de *prompt-engineering* e *few-shot learning*, obtemos uma avaliação com as mesmas métricas do que o RAGAS sem a utilização do mesmo.

Para a utilização destas métricas necessitamos de testes a aplicar à nossa solução. Estes testes foram obtidos de duas formas, manualmente, isto é, a nossa equipa leu alguns dos documentos recolhidos e elaborou questões a aplicar ao modelo. A outra metodologia de obtenção de testes foi através da utilização da funcionalidade de geração de testes sintéticos do RAGAS, que através dos documentos presentes para cada partido gerou 40 questões de diferentes tipos (i.e raciocínio, múltiplos contextos).

F. React

React é uma biblioteca JavaScript popular para construir interfaces de utilizador (UIs) interativas e responsivas. No frontend, React foi utilizado para:

- Implementar uma interface de consulta na qual os utilizadores possam se comunicar com o nosso sistema, tendo cada um as suas conversações guardadas numa conta.
- Implementar elementos dinâmicos que permitem a inserção de perguntas e a exibição de respostas de forma intuitiva.
- Permitir que o utilizador tenha uma experiência de fluida e eficaz, tornando a navegação e a interação com a aplicação mais fáceis.
- Interagir com o backend por meio de APIs, assegurando que as perguntas dos utilizadores são enviadas para o modelo de linguagem e que as respostas sejam apresentadas de maneira clara e organizada.

V. ARQUITETURA DA SOLUÇÃO

A interação com a nossa solução dá-se da seguinte forma:

- **Utilizador (à esquerda):** O utilizador interage com o sistema através de uma interface de consulta, no caso o website desenvolvido.
- **Recolha dos Dados:** Os planos políticos dos partidos são recolhidos, geralmente disponíveis em formato PDF.
- **Conversão dos PDFs para .txt:** Um programa em Python é utilizado para converter os documentos PDF para o formato .txt. Esta conversão facilita a manipulação e a análise dos textos.

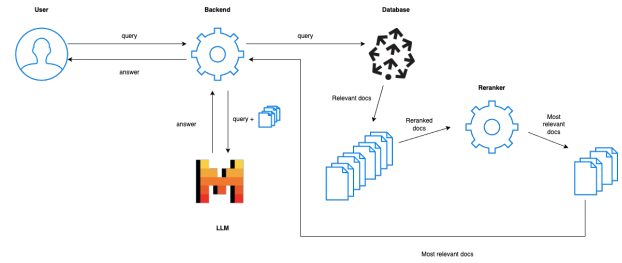


Fig. 1. Arquitetura da solução com reranker

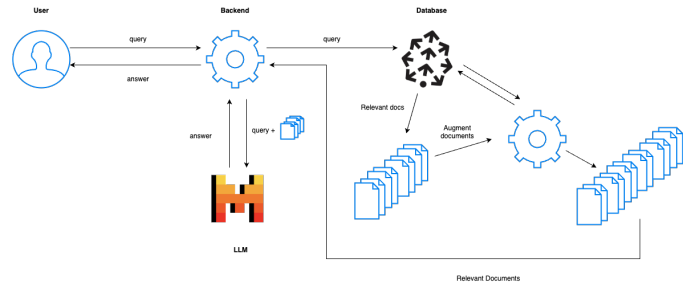


Fig. 2. Arquitetura da solução com TopK+Overlap

- **Processamento dos Dados:** Após a conversão, os dados textuais são processados. Aqui são aplicadas técnicas de scraping e processamento de linguagem natural (NLP) para extrair e categorizar as medidas políticas por tema.
- **Armazenamento Estruturado:** As informações extraídas e organizadas são armazenadas de forma estruturada. Esta estruturação permite uma fácil integração com o modelo de linguagem.
- **Treino do Modelo de Linguagem (LLM):** Utilizando os dados estruturados, o modelo de linguagem é treinado ou ajustado (fine-tuned). Este treino capacita o modelo a entender e responder a perguntas sobre as medidas políticas dos partidos.
- **Implementação da Aplicação Web:** O modelo de linguagem treinado é integrado numa aplicação web funcional. Esta aplicação permite que os cidadãos façam perguntas sobre as medidas dos partidos políticos e obtenham respostas precisas e relevantes.
- **Interação do Utilizador:** O utilizador utiliza o website para tirar dúvidas sobre as medidas dos principais partidos políticos portugueses. O LLM processa as perguntas e fornece respostas adequadas baseadas nos dados políticos recolhidos e processados.

VI. BENCHMARK

Nesta secção exploramos alguns resultados empíricos obtidos ao longo do desenvolvimento da nossa solução. Este benchmark mostra a resposta *default* do modelo selecionado e da nossa solução.

A. Como pretende a AD combater a corrupção?

1) **Resposta do LLM:** A Aliança Democrática (AD) é uma coligação **política brasileira** que surgiu em 2021, composta

por partidos de diferentes espectros políticos, com o objetivo de enfrentar a crise política (...) Combate à lavagem de dinheiro: A AD defende a adoção de medidas que garantam a efetividade da persecução e punição dos crimes de lavagem de dinheiro, como a ampliação dos recursos destinados às autoridades de investigação e às unidades de combate à lavagem de dinheiro, a implementação de medidas que garantam a independência e a efetividade do Ministério Público e a adoção de medidas que garantam a transparência nos fluxos financeiros dos partidos políticos e das campanhas eleitorais.

2) *Resposta do nosso modelo:* A forma como o partido AD pretende combater a corrupção é através de três métodos principais. Primeiro, propõem aumentar a eficácia e a celeridade da investigação e punição de crimes de corrupção, o que será alcançado pela criminalização do enriquecimento ilícito, reforma dos mecanismos institucionais e processuais de combate à corrupção, e proteção de denunciantes e vítimas. Em segundo lugar, buscam aumentar a integridade e transparência da governação pública e reconstruir uma cultura de responsabilidade política, através da regulamentação do lobbying, redução de conflitos de interesse, e melhoria dos sistemas de declaração de rendimentos, património e interesses. Por último, têm como meta melhorar a posição de Portugal no Índice de Perceção da Corrupção.

B. Compara as medidas do PS e da AD para a educação

1) *Resposta do LLM:* O Partido Socialista (PS) e a Aliança Democrática (AD) têm diferentes propostas e ênfases quando se trata de medidas para a educação. Partido Socialista (PS):

- Gratuidade da educação pré-escolar: O PS propõe a gratuidade da educação pré-escolar para todas as crianças com idade entre 4 e 6 anos, a fim de garantir um melhor início no sistema educacional.
- Aumento de investimentos na educação: O PS defende um aumento dos investimentos na educação, especialmente na formação de professores e na infraestrutura escolar, com o objetivo de melhorar a qualidade do ensino e as condições de aprendizagem.
- Valorização da formação profissional: O PS busca reforçar a formação profissional e a educação técnica, oferecendo mais opções de cursos e aumentando a ligação entre a escola e o mundo do trabalho.
- Combate às desigualdades socioeconômicas: O PS busca combater as desigualdades socioeconômicas na educação por meio de políticas como bolsas de estudo, transporte escolar gratuito e apoio a escolas em áreas desfavorecidas.
- Promoção da inclusão e diversidade: O PS defende a promoção da inclusão e diversidade na educação, incluindo o ensino de direitos humanos, igualdade de gênero e educação intercultural.

Aliança Democrática (AD):

- Liberdade de escolha: A AD defende a liberdade de escolha dos pais em relação à educação de seus filhos, incluindo o apoio à escolha de escolas particulares e à educação doméstica.

- Promoção da excelência acadêmica: A AD busca promover a excelência acadêmica por meio de avaliações regulares, métricas de desempenho e incentivos para escolas e professores que obtiverem bons resultados.
- Menos burocracia e mais autonomia escolar: A AD defende a redução da burocracia no sistema educacional e a promoção da autonomia das escolas, permitindo que elas tomem decisões mais próximas das necessidades locais.
- Foco na responsabilidade individual: A AD busca promover a responsabilidade individual dos alunos, professores e escolas, incentivando o mérito e o desempenho acadêmico.
- Parcerias público-privadas: A AD defende o estabelecimento de parcerias público-privadas para melhorar a infraestrutura escolar, oferecer formação profissional e promover a inovação no ensino.

Em resumo, o PS prioriza a igualdade de oportunidades, o combate às desigualdades e a promoção da inclusão, enquanto a AD dá ênfase à liberdade de escolha, à excelência acadêmica e à responsabilidade individual. Ambos os partidos têm propostas diferentes, mas visam melhorar a qualidade e o acesso à educação no país.

2) *Resposta do nosso modelo:* O Partido Socialista (PS) e o Partido AD apresentam medidas diferentes no seu plano eleitoral para a educação. O PS enfatiza a importância da literacia motora nas primeiras idades escolares, entre os 6 e os 10 anos, e propõe aumentar o investimento no desporto escolar. Além disso, o PS quer criar um novo estatuto do dirigente desportivo benévolo para mobilizar mais cidadãos para as funções dirigentes, especialmente nos clubes de base local. O Partido Socialista também dá prioridade à proteção de crianças e jovens com deficiência, consolidando e expandindo projetos de apoio à vida independente e garantindo a avaliação do impacto e eficácia das normas de promoção da inclusão dessas pessoas no mercado de trabalho. Por outro lado, a Partido AD quer confiar mais na autonomia das escolas públicas para ensinar e se diferencia e rejeita o reforço do papel da diplomacia económica, proposta do Partido Socialista. A AD também menciona a crise na habitação e os baixos níveis de investimento e poupança como problemas a serem abordados em seu plano eleitoral, mas não fornece medidas específicas para a educação além disso. Portanto, as medidas do PS e da AD para a educação são diferentes, com o PS se concentrando em questões relacionadas à literacia motora, desporto escolar e proteção de crianças e jovens com deficiência, enquanto a AD prioriza a autonomia das escolas públicas e aborda outros problemas além da educação em seu plano eleitoral.

C. Compara as medidas do PS e da AD para a habitação

1) *Resposta do LLM:* O Partido Socialista (PS) e a Aliança Democrática (AD) têm propostas diferentes em relação à habitação. O Partido Socialista tem como prioridade a construção de moradias acessíveis e de qualidade, com foco em áreas centrais das cidades e em transporte público. O PS também defende a regulamentação do mercado imobiliário, a promoção do aluguer social e a criação de um

Fundo Nacional de Habitação. Além disso, o partido propõe a revitalização de bairros e a criação de novas zonas habitacionais, com o objetivo de combater a especulação imobiliária e a gentrificação. Por outro lado, a Aliança Democrática defende uma abordagem mais liberal em relação à habitação, com menos intervenção estatal. A AD propõe a simplificação dos processos de licenciamento e a redução de impostos para incentivar a construção de moradias, especialmente em áreas periféricas das cidades. A aliança também defende a promoção do aluguer privado e a criação de um mercado de aluguer flexível, com o objetivo de aumentar a oferta de moradias e reduzir os preços. Em resumo, o Partido Socialista tem uma abordagem mais regulada e intervencionista em relação à habitação, enquanto a Aliança Democrática defende uma abordagem mais liberal e desregulada. As diferenças entre as duas propostas refletem as diferenças ideológicas entre os dois partidos e suas diferentes visões sobre o papel do Estado na economia e na sociedade.

2) *Resposta do nosso modelo:* As medidas do Partido Socialista (PS) e da Aliança Democrática (AD) para a habitação apresentam algumas diferenças significativas. O PS enfatiza a regulamentação dos preços do aluguer como forma de abordar a crise de acessibilidade habitacional, através do controlo administrativo de preços, por congelamento de rendas ou limitação dentro de bandas estreitas. No entanto, o contexto menciona que medidas semelhantes adotadas em Portugal e no mundo fora tiveram resultados negativos, com aumentos de 30% nas rendas dos novos contratos e uma redução do investimento e da renovação e manutenção dos imóveis. Por outro lado, a AD defende uma abordagem diferente, concentrando-se na mobilização de toda a sociedade para um efetivo estímulo à oferta de habitações acessíveis, tanto no mercado de arrendamento como de aquisição. A AD propõe flexibilizar as limitações de ocupação dos solos, densidades urbanísticas e exigências construtivas, além de injetar no mercado imóveis e solos públicos devolutos ou subutilizados. Ao contrário do PS, a AD não defende o controlo administrativo de preços, mas sim medidas que incentivem a oferta de habitação. Em resumo, o PS defende o controlo administrativo de preços no mercado de aluguer, enquanto a AD se concentra em medidas que incentivem a oferta de habitação acessível, flexibilizando as limitações e injetando no mercado imóveis e solos públicos devolutos ou subutilizados.

VII. VALIDAÇÃO DOS RESULTADOS OBTIDOS

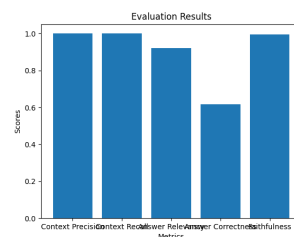
De forma a validar a solução implementada, utilizamos um conjunto de questões geradas manualmente pelo grupo e um *testset sintético* gerado através do RAGAS.

Algumas das questões contempladas neste testset são as seguintes:

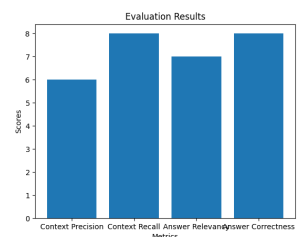
- 1) Como a retenção de talentos pode beneficiar a sociedade?
- 2) Quais são as iniciativas da Rede Portuguesa de Arte Contemporânea para promover a arte e a inclusão social dos idosos?

- 3) Quais medidas específicas visam melhorar a mobilidade dos cidadãos em territórios de baixa densidade e promover a conectividade digital no país?
- 4) Quais são as propostas para promover a valorização da produção agrícola nacional?
- 5) Quais são as medidas da AD para melhorar a oferta habitacional de forma sustentável?
- 6) Por que é importante ter centros de apoio para sobreviventes de violência sexual e como isso se relaciona com a criação de um canal permanente de denúncia de abusos sexuais em instituições?
- 7) Quais as soluções apresentadas pela AD para a emigração de jovens?
- 8) Quais são as medidas propostas para digitalizar os processos nos tribunais e garantir o registo em áudio e vídeo com transcrição automática?
- 9) De que forma as propostas de educação, habitação e benefícios fiscais no plano visam combater o desemprego jovem em Portugal?
- 10) Como implementar programas habitacionais em áreas metropolitanas inspirados em modelos de planeamento urbano bem-sucedidos, como o Parque Expo, através da reformulação da fiscalidade na habitação?
- 11) Como garantir a celeridade da justiça fiscal?
- 12) Quais as metas da aliança democrática para a educação?
- 13) Quais medidas podem ser tomadas para resolver atrasos nos Tribunais Administrativos e Fiscais e melhorar as condições nas prisões em Portugal?
- 14) Quais são as iniciativas da Rede Portuguesa de Arte Contemporânea para promover a circulação da arte e como se relacionam com a inclusão social dos idosos?
- 15) Como implementar um programa de reforço de psicólogos nas escolas e instituições de ensino superior para atender às necessidades identificadas?
- 16) Como fortalecer a presença de psicólogos nas escolas públicas do ensino básico e secundário, considerando a burocracia, exames nacionais e progressão na carreira docente?

Estes testes foram aplicados à solução com a utilização de *Top-K* e da utilização de um *Reranker* de forma a aclarar qual seria a arquitetura mais benéfica para a nossa solução.



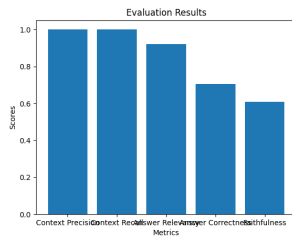
Resultados RAGAS



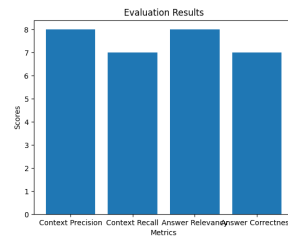
Resultados modelo crítico

Fig. 3. Resultados com a utilização de Top-K

1) Resultados Test_1:

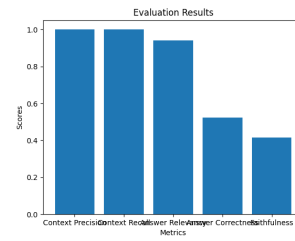


Resultados RAGAS

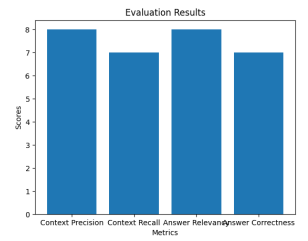


Resultados modelo crítico

Fig. 4. Resultados com a utilização de Reranker

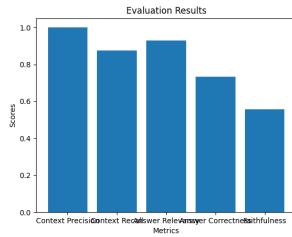


Resultados RAGAS

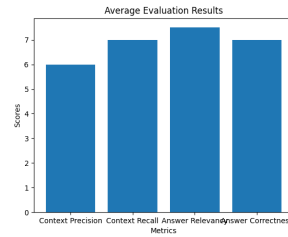


Resultados modelo crítico

Fig. 8. Resultados com a utilização de Reranker



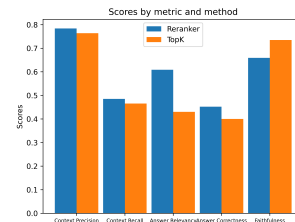
Resultados RAGAS



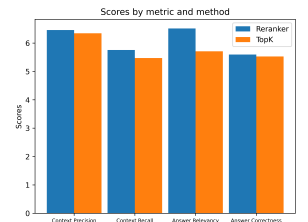
Resultados modelo crítico

Fig. 5. Resultados com a utilização de Top-K

A. Média dos testes

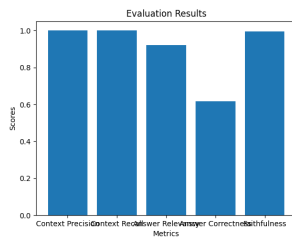


Resultados RAGAS

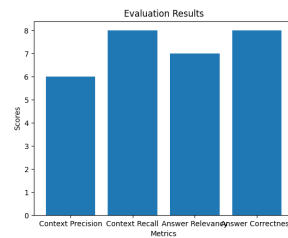


Resultados modelo crítico

Fig. 9. Resultados RAGAS vs Modelo Crítico



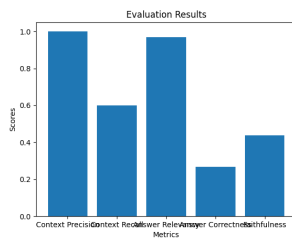
Resultados RAGAS



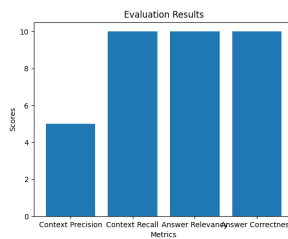
Resultados modelo crítico

Fig. 6. Resultados com a utilização de Reranker

2) Resultados Test_7:



Resultados RAGAS



Resultados modelo crítico

Fig. 7. Resultados com a utilização de Top-K

3) Resultados Test_12:

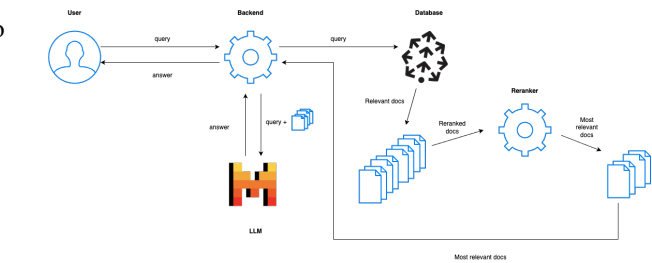


Fig. 10. Arquitetura da solução com reranker

VIII. DEMONSTRAÇÃO DO FUNCIONAMENTO DO WEBSITE

Como referido anteriormente, usamos o React para o front-end. A nossa ideia seria um website que fosse parecido, em termos de funcionalidades, com o chatGPT e que tivesse as cores da bandeira nacional, tendo em conta o contexto do projeto. Assim, foram feitas 4 páginas:

- **Página Inicial:** Página que contém uma breve apresentação
- **Página de login:** Página onde cada utilizador pode entrar e, posteriormente, acessar aos seus chats;
- **Página de registo:** Página com a função de registar novos utilizadores;

- **Página dos Chats:** Página que contém uma lista de conversações ocorridas com o PolitAI. O utilizador pode continuar a falar nesses chats ou pode criar um novo. Por fim, só resta a funcionalidade de logout.

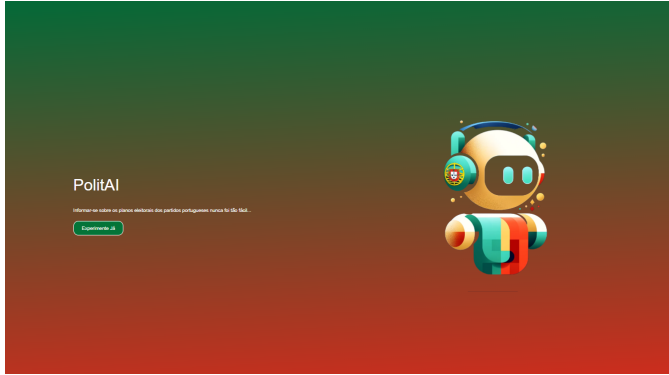


Fig. 11. Página Inicial

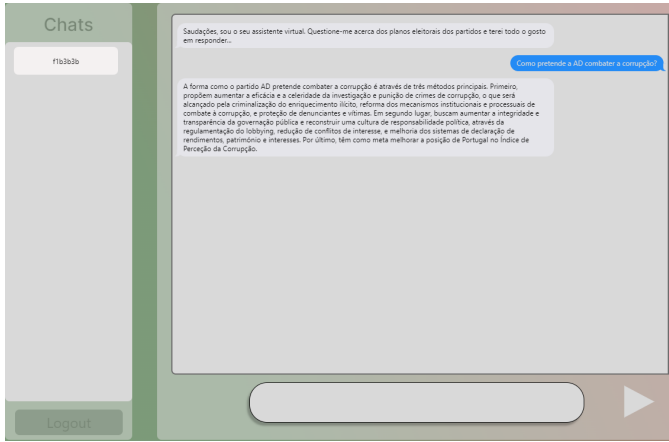


Fig. 12. Página do Chat

IX. TRABALHO FUTURO

Futuramente, experimentar outras formas de validação (*i.e. Galileo Luna*) seria certamente um passo fulcral visto o RAGAS e os modelos críticos terem custos avultados para a avaliação sistemática da solução.

Para aprimorar ainda mais a precisão e a relevância das respostas fornecidas pelo nosso modelo de linguagem, consideramos a implementação de uma solução multiagente. Esta abordagem poderia ser realizada de duas formas principais: um **agente dedicado para cada partido**, onde cada partido político teria um agente dedicado. Cada agente seria especializado nas políticas e medidas específicas de seu partido ou um **agente parafraseador** e **agente sumarizador** que teria a função de parafrasear a consulta original do utilizador para melhorar a compreensão da pergunta pelo sistema. Ao reformular a pergunta, o agente pode identificar melhor o contexto e as nuances, resultando numa pesquisa mais eficaz das informações relevantes e ainda um agente sumarizador,

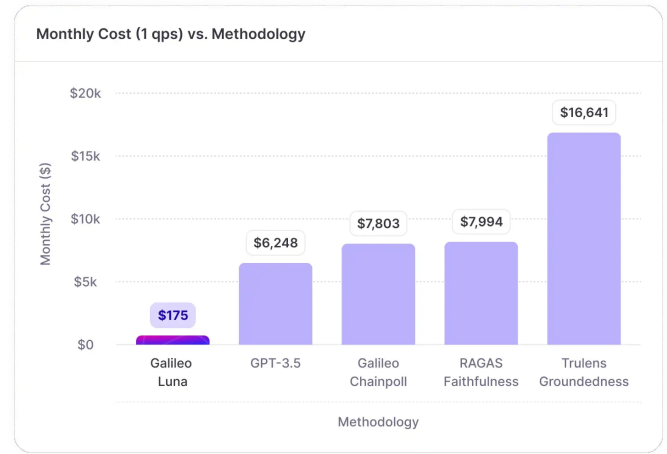


Fig. 13. Galileo Luna vs outros métodos de avaliação

Method	QUESTION ANSWERING			DATA-TO-TEXT WRITING			SUMMARIZATION			OVERALL		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
Prompt _{GPT-3.5 turbo} [†]	18.8	84.4	30.8	65.1	95.5	77.4	23.4	89.2	37.1	37.1	92.3	52.9
Prompt _{GPT-4 turbo} [†]	33.2	90.6	45.6	64.3	100.0	78.3	31.5	97.6	47.6	46.9	97.9	63.4
SelfCheckGPT _{GPT-3.5 turbo} [†]	35.0	58.0	43.7	68.2	82.8	74.8	31.1	56.5	40.1	49.7	71.9	58.8
LMvLM _{GPT-4 turbo} [†]	18.7	76.9	30.1	68.0	76.7	72.1	23.2	81.9	36.2	36.2	77.8	49.4
Finetuned Llama-2-13B [†]	61.6	76.3	68.2	85.4	91.0	88.1	64.0	54.9	59.1	76.9	80.7	78.7
ChainPoll _{GPT-3.5 turbo}	33.5	51.3	40.5	84.6	35.1	49.6	45.8	48.0	46.9	54.8	40.6	46.7
RAGAS Faithfulness	31.2	41.9	35.7	79.2	50.8	61.9	64.2	29.9	40.8	62.0	44.8	52.0
Trulens Groundedness	22.8	92.5	36.6	66.9	96.5	79.0	40.2	50.0	44.5	46.5	85.8	60.4
Luna	37.8	80.0	51.3	64.9	91.2	75.9	40.0	76.5	52.5	52.7	86.1	65.4

Fig. 14. Benchmark Galileo Luna

que após a obtenção das respostas dos agentes especializados ou do banco de dados, o agente sumarizador teria a tarefa de condensar e estruturar a informação de maneira clara e concisa para o utilizador.

Ao longo da nossa pesquisa por métodos de procura por similaridade, visto ser um passo crucial (Retrieval) da pipeline RAG, tomamos conhecimento da biblioteca *FAISS* [5] open-source desenvolvida pela *Meta*, que disponibiliza métodos mais avançados e mais rápidos de procura por similaridade.

Gostaríamos, também, de experimentar o modelo Zephyr-7b-alpha [6], um fine-tune para assistentes virtuais, para avaliar a diferença de performance em relação à nossa solução atual. Gostávamos de o comparar com o nosso modelo de forma a averiguar se oferece respostas mais precisas e naturais, embora possa exigir maior capacidade computacional e ajustes adicionais.

No futuro, gostaríamos de abranger ainda mais partidos portugueses na nossa aplicação, com ou sem assento parlamentar. Poderíamos, ainda, escalar a nossa aplicação para que consiga comparar com partidos de outros países, tendo também em conta também os grupos parlamentares europeus.

X. CONCLUSÃO

Este relatório demonstrou o desenvolvimento bem-sucedido de uma solução robusta que atua como um chatbot interativo sobre os partidos políticos portugueses. Através de uma abordagem metodológica estruturada, abordamos a conversão e processamento de dados políticos, a implementação de uma

pipeline RAG eficiente e a integração de tecnologias avançadas para melhorar a precisão e relevância das respostas geradas. Com este trabalho conseguimos uma *proof of concept* que demonstra a utilidade do nosso chatbot e que pode vir a mitigar os problemas que incitaram o desenvolvimento do mesmo. Os *benchmarks* realizados evidenciaram que a nossa abordagem, especialmente com a utilização de técnicas de *NLP* e modelos de linguagem como o *Mixtral-8x7B*, superou soluções padrão em termos de precisão e coerência semântica. A escolha de tecnologias como *Pinecone* para armazenamento vetorial e *React* para o *frontend* garantiu uma interface de usuário intuitiva e responsiva, melhorando a experiência geral dos utilizadores. Para aprimorar ainda mais nossa solução, consideramos a implementação de uma abordagem multiagente, utilizando agentes dedicados para cada partido político e agentes parafraseadores e sumarizadores. Além disso, planeamos explorar métodos mais avançados de procura por similaridade, como a biblioteca *FAISS*, e testar modelos de linguagem adicionais, como o *Zephyr-7b-alpha*. Em suma, o trabalho desenvolvido não apenas alcançou os objetivos propostos pela Unidade Curricular de Mineração de Dados, mas também estabeleceu uma base sólida para futuras expansões e melhorias. Acreditamos que, com o contínuo aperfeiçoamento, a nossa aplicação pode se tornar numa ferramenta essencial para a democratização da informação política e o fortalecimento da participação cidadã em Portugal e, potencialmente, noutros países.

REFERENCES

- [1] F. Moghimifar, Y.-F. Li, R. Thomson, and G. Haffari, “Modelling political coalition negotiations using llm-based agents,” *arXiv preprint arXiv:2402.11712*, 2024.
- [2] K. Kato, A. Purnomo, C. Cochrane, and R. Saqr, “L (u) pin: Llm-based political ideology nowcasting,” *arXiv preprint arXiv:2405.07320*, 2024.
- [3] Pinecone, “Pinecone database,” 2024. [Online]. Available: <https://docs.pinecone.io/home>
- [4] “Together ai,” 2024. [Online]. Available: <https://www.together.ai/>
- [5] Meta, “The faiss library,” 2024. [Online]. Available: <https://github.com/facebookresearch/faiss>
- [6] HuggingFaceH4, “Huggingfaceh4/zephyr-7b-alpha,” 2024. [Online]. Available: <https://huggingface.co/HuggingFaceH4/zephyr-7b-alpha>