

Mineração de Dados

Relatório do Trabalho Prático

MEI - 2023/2024

Grupo 4

Hugo Martins
A95125



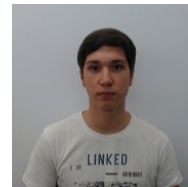
João Escudeiro
A96075



Henrique Fernandes
A95323



Alexandre Fernandes
PG53606



Universidade do Minho

1 Introdução e Objetivos

Nos últimos anos, a tecnologia tem avançado a passos largos, transformando diversos aspetos da vida quotidiana e da indústria. Uma das áreas de maior destaque nesse progresso é a inteligência artificial (IA), especialmente no que se refere aos modelos de linguagem *LLM*. Estes modelos, que são treinados com vastas quantidades de dados textuais, têm a capacidade de entender, gerar e interagir com a linguagem humana de uma forma cada vez mais sofisticada.

Este trabalho tem como objetivo explorar a implementação de uma aplicação ou conduzir um estudo que envolva a integração de um *LLM*. A escolha do modelo será baseada no desempenho no tema escolhido pelo grupo. O modelo selecionado será refinado e estendido através da recolha, processamento e análise de dados provenientes de uma ou mais fontes. O intuito do projeto é melhorar ou adaptar o modelo existente para aplicações específicas.

2 Definição do Domínio

Inicialmente, decidimos realizar uma pesquisa abrangente em diversas áreas de interesse, como desporto, alimentação e videojogos. No entanto, percebemos que muitas dessas áreas já contam com inúmeros exemplos de *LLMs* especializados. Assim, resolvemos focar na área do futebol, uma área em que todos os membros do grupo possuem interesse.

Durante a nossa pesquisa, geramos várias ideias interessantes, mas encontramos dificuldades em obter fontes de dados adequadas para alimentá-las. Após uma análise mais detalhada, descobrimos uma fonte de dados robusta que poderia suportar o nosso projeto, que será mencionada na Secção 4. Decidimos, então, criar um ChatBot especializado em prever resultados de jogos de futebol, baseado em estatísticas, resultados anteriores e desempenho atual das equipas.

No entanto, após uma conversa com nosso docente, concluímos que a melhor abordagem seria especializar um LLM em uma liga específica de nosso interesse. Optamos pela **Liga Portugal Betclic**, uma vez que é a liga de futebol do nosso país. O nosso ChatBot, com base em estatísticas, classificações, resultados, *line-ups* e a forma recente das equipas, fornecerá informações atualizadas. Sempre que for questionado sobre uma equipa, ele será capaz de fornecer estatísticas relevantes, bem como classificações das últimas temporadas.

Dessa forma, pretendemos criar uma ferramenta útil e especializada que forneça *insights* detalhados e precisos sobre a Liga Portugal Betclic.

Demos o nome de **FutProBot** ao nosso LLM, tendo em conta que esse nome reflete claramente o foco e a especialização do nosso modelo.

3 Análise do Estado da Arte

Após atingirmos uma ideia final completamente definida, decidimos procurar a fundo por ideias semelhantes, ou algo que exista relacionado com a Liga Portugal. Para iniciar, decidimos solicitar a classificação da Liga Portugal na época desportiva 2023/2024, ao mistral, um modelo Ollama que será um modelo que vamos utilizar mais à frente. O mesmo responde que não possui dados superiores à época de 2021/2022.

De seguida decidimos procurar por ferramentas especializadas nesta área na Web. Encontramos um bot chamado **FootBot.net**[1], que é uma plataforma que fornece previsões de jogos de futebol. O site oferece previsões detalhadas para várias ligas de futebol ao redor do mundo. As previsões são geradas usando o próprio sistema do Footbot.net, que analisa diversos fatores para oferecer dicas de apostas e previsões. Os utilizadores podem comprar previsões individuais por uma pequena taxa, e essas previsões incluem detalhes como resultados esperados das partidas e análises estatísticas. O site menciona a precisão e a confiabilidade das dicas, embora deixe claro que não garante precisão completa e que os seus serviços são apenas para fins de aconselhamento.

Outro exemplo era o site **SoccerSM.ai**[3], que oferece previsões de resultados de jogos de futebol baseadas em modelos de inteligência artificial. Ele utiliza dados estatísticos e algoritmos avançados para fornecer análises e previsões sobre os resultados de jogos futuros.

Encontramos também o **AIDA**[4], que é um chatbot avançado de futebol desenvolvido pela Soccerment, que utiliza inteligência artificial. A plataforma oferece análises detalhadas de jogadores e equipas, incluindo comparações, rankings, e métricas de desempenho. AIDA é útil para fantasy football e scouting, permitindo consultas interativas à base de dados da Soccerment para obter insights em tempo real sobre jogadores e jogos.

Incluído no **ChatGPTPRO** temos o **Football Chatter**, desenvolvido por İbrahim Demir, que utiliza modelos avançados de processamento de linguagem natural (NLP) e machine learning para analisar dados de futebol. O chatbot integra diversas fontes de dados, como estatísticas de jogos, desempenhos de jogadores, e históricos de partidas, para gerar respostas precisas e insights detalhados. O modelo de NLP usado permite uma compreensão profunda de consultas textuais e a entrega de informações em tempo real sobre resultados, previsões, e análises de desempenho.

Embora tenhamos encontrado diversos chats que oferecem resultados interessantes, nenhum deles se mostrou especializado na Liga Portuguesa, nem em estatísticas ou história relacionadas a ela. Portanto, acreditamos que a nossa ideia preenche uma lacuna no cenário atual, proporcionando uma abordagem mais aprofundada e especializada neste campo.

4 Dados

Nesta seção, vamos explorar as fontes de dados e descrever o processo completo, desde a obtenção inicial dos dados até os dados estarem na sua forma final, prontos para "alimentarem" a LLM.

4.1 Fonte

Descobrimos uma plataforma chamada SPORTMONKS[2] que possui dados atualizados a cada semana com as informações das principais ligas europeias (La Liga, Premier League, Bundesliga, Serie A, Ligue 1 e Liga Portugal).

Para a obtenção destes dados seria necessário o pagamento de uma taxa mensal. Decidimos entrar em contacto com a plataforma, mencionando que os dados seriam utilizados para fins escolares. Forneceram-nos um token gratuito com duração de 6 meses que nos permite fazer pedidos ilimitados sobre qualquer uma das ligas e das informações que constam nas bases de dados que suportam a plataforma. Para filtrar os dados (liga e informação pretendida), utilizamos valores específicos nos pedidos da API, inseridos no URL correspondente (liga, época, informação pretendida).

4.2 Processamento dos dados

O processamento dos dados foi realizado em duas etapas principais. A **recolha** dos dados e o **tratamento**. Para a **recolha** dos dados, o grupo decidiu criar vários scripts em GO responsáveis por fazerem pedidos à plataforma e armazenarem os resultados numa base de dados em nuvem, para que fosse possível posteriormente os elementos do grupo dividirem a parte do tratamento dos dados. Os scripts armazenaram a informação na plataforma MongoDBAtlas em 4 coleções:

- **Fixtures** - Contém informação sobre todos os jogos desde 2005, resultados, táticas, estatísticas do jogo, árbitros e eventos (golos, substituições e cartões).
- **Seasons** - Contém informação sobre as estatísticas globais das épocas como equipa com mais cantos, mais cartões, mais remates, mais golos sofridos e marcados.
- **Standings** - Contém a classificação de todas as épocas desde 2005, incluindo o número de pontos de cada equipa.
- **Teams** - Contém informações detalhadas sobre o plantel de cada equipa, bem como as datas de início e fim de contrato e transferências.

De seguida, e após obtermos os dados armazenados, decidimos passá-los para formato textual, construindo frases, para que a informação esteja articulada de forma a que o LLM utilizado consiga consumir a informação de uma forma mais estruturada e simples. Com a informação armazenada nas coleções mencionadas

acima, construímos scripts Python que convertessem a informação dos jogos em formato textual. De seguida são apresentados os ficheiros resultantes, bem como exemplos de frases que constituem os mesmos.

- **champions**-Campeões da Liga portuguesa. "In the 2000/2001 season the champion was Boavista."
- **events**- Eventos durante um jogo. "The game Vitória SC vs Sporting Braga of the season 2023/2024 at minute 90, Rony Lopes scored a goal for team Sporting Braga."
- **formations**-Esquemas tácticos das equipas num jogo. "In the game Sporting CP vs Portimonense of the season 2023/2024 the away team was Portimonense and played in a 4-3-3 formation and the home team was Sporting CP and played in a 3-4-3 formation."
- **intro**- Informação sobre um jogo. "The game between Benfica and Vitória SC of the season 2019/2020 took place in 2020-07-14 and started at 20:30:00 and Benfica won after full-time."
- **cup_winners**-Vencedores da taça de Portugal. "In the 2005/2006 season, the Portuguese Cup winner was Porto."
- **referees**-Árbitros de um jogo. "The game Benfica vs Arouca of the season 2023/2024 was refereed by Fábio Oliveira Melo, André Filipe Nogueira Dias and António Ricardo de Mesquita Moreira."
- **seasons**-Estatísticas da época. "In Season 2023/2024 the top goalscorer was Viktor Gyökeres with 29 goals."
- **standings**-Classificações finais. "In Season 2019/2020 Porto was champion with 82 points."
- **statistics**-Estatísticas de um jogo. "In game Porto vs Boavista of the season 2023/2024 Boavista have 27 Throw-ins"
- **teams**-Informação sobre os plantéis das equipas. "Marcos Leonardo Santos Almeida, 177.0 tall, is a centre forward for Benfica with the 36 jersey number."

Para finalizar, fizemos a conversão dos ficheiros de texto para PDF com o objetivo de otimizar a alimentação do modelo de linguagem, facilitando o processamento subsequente dos documentos.

5 Modo de Funcionamento

Nesta secção será abordado o modo de funcionamento do **FutProBot**, bem como todas as etapas ultrapassadas até atingir o produto final.

Para criar um modelo de linguagem (LLM) eficiente e bem alimentado, o grupo optou por um processo estruturado que envolveu várias etapas, desde a conversão de documentos até a interação com o modelo final. O processo seguido foi o seguinte :

1. **Carregamento dos ficheiros de dados:** Tal como foi referido nas secções anteriores deste documento, toda a informação recolhida pela a nossa equipa de desenvolvimento encontra-se armazenada em ficheiros PDF. Isto deve-se principalmente à fácil integração que este tipo de ficheiros apresentam em cenários semelhantes àqueles apresentados no nosso processo de conspeção do chat bot.

Tendo isto em mente, esta primeira etapa teve como principal objetivo o carregamento dos dados recolhidos para um formato que fosse compatível com o seguinte processo no pipeline de desenvolvimento do FutProBot, sendo para isso utilizado bibliotecas como o PyPDFLoader, em ordem a atingir este objetivo.

2. **Configuração do Divisor do texto:** Após terem sido carregados os dados com sucesso, foi configurado um separador recursivo de texto *RecursiveCharacterTextSplitter*, com um *chunksizes* de 2000 caracteres e uma sobreposição de 200 caracteres. No entanto, é também importante salientar que este processo passou também por várias fases de iteração distintas, de modo a ser alcançada uma configuração capaz de apresentar melhores resultados quando em ambiente de execução.

Com isto, a grupo garante que esta configuração permite dividir os documentos em partes menores (*chunks*), de um modo que permite ao nosso modelo a criação de embeddings e a indexação dos dados de maneira a que o processo de busca e filtragem dos dados seja simples e eficaz.

3. **Criação dos Embeddings:** O passo seguinte para a completude da nossa pipeline de desenvolvimento passou pela criação dos embeddings dos repetitivos chunks de informação produzidos na etapa anterior. Para conseguirmos atingir este objetivo, o grupo de trabalho utilizou a biblioteca do *Langchain* que disponibiliza um modelo de embeddings já pré-configurado e adaptado para LLMs do mesmo tipo daquele que estamos a utilizar (*OllamaEmbeddings*), ou seja, LLMs da família **Ollama** como é o caso do **Mistral**.
4. **Armazenamento dos ficheiros numa Base de Dados vetorial:** Tendo já os nossos PDFs com todos os dados foram de seguida armazenados numa base de dados vetorial ChromaDB, que foi escolhida devido à sua capacidade de indexação e recuperação dos dados de maneira eficiente, essencial para lidar com grandes volumes de dados textuais.

5. **Integração com o Modelo:** Tendo agora toda a nossa informação armazenada na base de dados vetorial, passamos agora à integração com o modelo de linguagem em concreto.

Para isso, passamos primeiro por voltar a carregar os embeddings referentes ao **Mistral**, assim como os documentos armazenados na base de dados **ChromaDB**, de modo a ser possível inicializar o LLM com os nossos documentos, adicionando-os assim à informação que este já dispunha anteriormente. Outro aspeto que também é importante referir é o facto de que os chunks que serão recolhidos da nossa base de dados serão aqueles que apresentam um grau de similaridade mais elevada com uma dada questão imposta pelo o utilizador da nossa aplicação, podendo cada recolha apresentar até 5 dos chunks mais relevantes para um dado contexto.

Outra adição feita à inicialização do modelo passa também pela disponibilização de um template de resposta do modelo, em que são dadas algumas orientações ao LLM no modo como este deve responder a todas as questões impostas por um dado utilizador, aliado ao contexto referente a essa mesma pergunta e também algumas indicações de formatação das suas respostas.

Posteriormente, é adicionado um buffer de memória a cada uma das conversas mantidas com o modelo, de modo a que este seja capaz de utilizar respostas e contextos anteriores na geração de novas respostas, sem que seja necessário ao utilizador especificar todo e qualquer detalhe em cada uma das suas interrogações.

6. **Criação do FrontEnd:** Por fim, criamos um frontend simples, como está demonstrado na Figura 2, para permitir aos utilizadores interagir com nosso modelo de linguagem e explorar os resultados dos comentários de futebol fornecidas pelo nosso ChatBot. O FrontEnd foi implementado utilizando *Streamlit*.

Na Figura 1, podemos observar um diagrama de como funciona o nosso Bot. Embora seja um pouco abstrato, a imagem é capaz de transmitir/representar os principais pilares que constituem a aplicação de chat desenvolvida, passando desde o carregamento de ficheiros, até à sua partição e armazenamento na base de dados vetorial, assim como ao processo interativo entre um dado cliente e/ou utilizador do sistema, e consequente ações tomadas pelo modelo para a realização/resposta de pedidos.

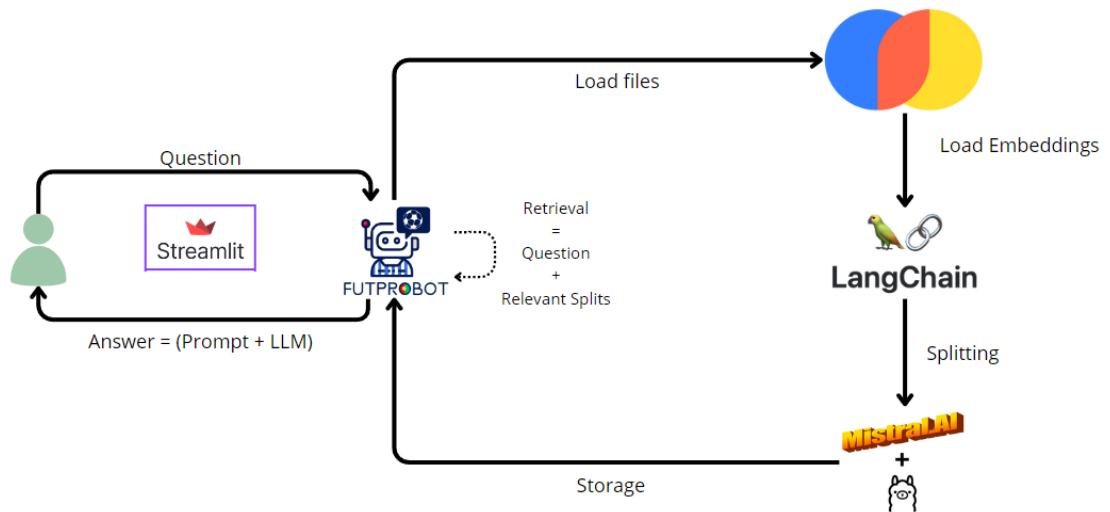


Figure 1: Exemplo do modo de funcionamento do FutProBot

6 Resultados Obtidos

Na figura 2 está presente um exemplo do produto final. O frontend foi concebido utilizando a biblioteca *streamlit*, que facilitou e muito na implementação do mesmo. Optamos por manter a simplicidade, em que as perguntas feitas pelo utilizador aparecem alinhadas à direita, enquanto que as respostas do **FutProBot** aparecem alinhadas à esquerda. Em baixo, existe uma barra de pesquisa, onde são feitas as perguntas ao chatbot. Na aba esquerda é exibido o logótipo criado pelo grupo. Nesta secção seria também apresentado o histórico de conversas, que não foi implementado devido à escassez de tempo.

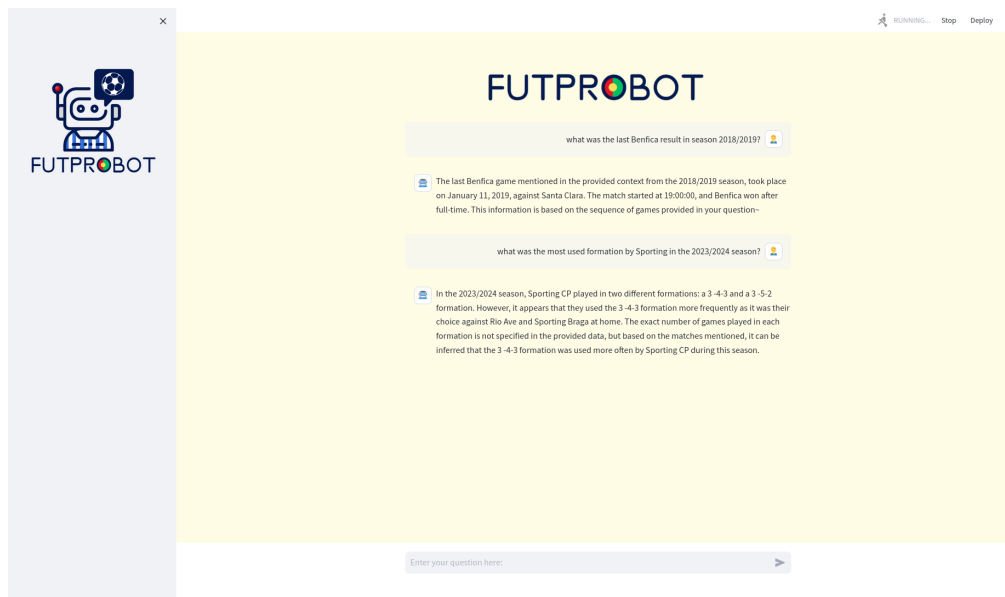


Figure 2: Exemplo do modo de funcionamento do FutProBot

7 Avaliação de Desempenho

Para avaliar o desempenho do resultado final decidimos realizar uma comparação entre as respostas fornecidas pelo modelo e o resultado esperado das mesmas. De seguida são apresentadas algumas perguntas, respostas e accuracy das respostas.

1. **Pergunta:** What was the most used formation by Sporting in the 2023/2024 season?

Resposta: In the 2023/2024 season, Sporting CP played in two different formations: a 3 -4-3 and a 3 -5-2 formation. However, it appears that they used the 3 -4-3 formation more frequently as it was their choice against Rio Ave and Sporting Braga at home. The exact number of games played in each formation is not specified in the provided data, but based on the matches mentioned, it can be inferred that the 3 -4-3 formation was used more often by Sporting CP during this season.

Observação: Resposta Correta

2. **Pergunta:** Who was the champion for the 2023/2024 season?

Resposta: The champion of the 2023/2024 season was Sporting CP, as they finished with a total of 87 points. This achievement marked their success in that particular season, surpassing other teams such as Benfica (second place with 79 points) and Porto (third place with 69 points). The remaining top five teams were Sporting Braga (fourth place with 68

points), Vitória SC (fifth place with 60 points), Moreirense (sixth place with 52 points), Arouca (seventh place with 46 points), Famalicão (eighth place with 42 points), Farense (ninth place with 37 points), Gil Vicente (tenth place with 36 points), Rio Ave (eleventh place with 36 points), Casa Pia (twelfth place with 35 points), Estoril (thirteenth place with 33 points), Boavista (fourteenth place with 31 points), Estrela Amadora (fifteenth place with 30 points), Portimonense (sixteenth place with 29 points), Vizela (penultimate place with 25 points), and Chaves (last place with 23 points).

Observação: Resposta correta (os dados fora retirados no fim da penúltima jornada).

3. **Pergunta:** Can you give me the league standings for the 2022/2023 season?

Resposta: In the 2022/2023 season, the Portuguese Primeira Liga standings were as follows:

1. Benfica - 87 points; 2. Porto - 85 points; 3. Sporting Braga - 78 points; 4. Sporting CP - 74 points; 5. Arouca - 54 points; 6. Vitória SC - 53 points; 7. Chaves - 46 points; 8. Boavista - 44 points; 9. Famalicão - 44 points; 10. Casa Pia - 41 points; 11. Rio Ave - 40 points; 12. Vizela - 40 points; 13. Gil Vicente - 37 points; 14. Estoril - 35 points; 15. Portimonense - 34 points; 16. Marítimo - 26 points; 17. Paços de Ferreira - 23 points; 18. Santa Clara - 22 points; 19. Belenenses - Not available (they were last in the previous season with 26 points);

Observação: Resposta parcialmente correta, erro na inclusão do Belenenses, que está a mais

4. **Pergunta:** Who is the team with the most cup wins?

Resposta: The team with the most Portuguese Cup wins, based on the data provided for seasons up to 1989/1990 and from Season 2023/2024, is Benfica. They have won the Portuguese Cup a total of 10 times: in the 1969/1970, 1970/1971, 1971/1972, 1980/1981, 1981/1982, 1984/1985, 1985/1986, 1986/1987, and twice in the 2023/2024 season. Porto follows closely with 7 victories: in the 1967/1968, 1976/1977, 1988/1989, 1990/1991, 1991/1992, and twice in the 2023/2024 season. Sporting has won the Portuguese Cup 5 times: in the 1973/1974, 1977/1978, 1982/1983, and twice in the 2023/2024 season. Boavista has also won the cup twice, in the 1974/1975 and 1992/1993 seasons.

Observação: Resposta errada

8 Trabalho futuro

Embora tenhamos alcançado bons resultados no projeto, existem possíveis melhorias a desenvolver num trabalho futuro.

Uma possível melhoria seria a integração de novas fontes de dados, como dados de jornais desportivos online, estatísticas de jogadores detalhadas e análises das redes sociais, para enriquecer ainda mais as análises e comentários do modelo.

Outro aspeto a melhorar seria a atualização dos dados em tempo real. Tendo em conta que os dados da plataforma são atualizados semanalmente, uma possível integração atualizada semanalmente tornaria o modelo ainda mais robusto e com a possibilidade de realizar comentários cada vez mais pertinentes e atuais.

Um aspeto que poderia ser melhorado seria a possibilidade da integração de histórico de mensagens na interface, que melhoraria a experiência do utilizador, podendo este navegar pelo seu histórico de mensagens.

9 Conclusão e considerações finais

Através da elaboração deste trabalho prático, foi-nos possível aplicar e consolidar todo o conteúdo lecionado ao longo do semestre.

Os objetivos propostos foram cumpridos na íntegra, e o Bot elaborado consegue dar resposta a perguntas relacionadas pela temática da Liga Portuguesa.

O grupo considera que a maior dificuldade ultrapassada foi a implementação inicial, visto que após possuímos os dados no formato textual, não sabíamos muito bem por onde começar a implementação. Esta dificuldade foi ultrapassada através da observação dos tutoriais sugeridos pelo docente, bem como da leitura da documentação das bibliotecas utilizadas.

Concluindo, consideramos que o trabalho desenvolvido é positivo, estando cientes que no futuro existiriam aspetos que já foram mencionados, que poderiam ser alvo de melhoria.

10 Referências

- [1] Predictions - FOOTBOT.NET.www.footbot.net/predictions/.
- [2] “Fast and Reliable Sports Data for Your Web App.” Sportmonks. www.sportmonks.com/.
- [3] Soccersm,www.soccersm.ai/predictions.
- [4] “Unlock the Power of Soccer Analytics. Anytime, Anywhere.” Soccerment, xvalue.ai/aida.