Investigação sobre a Leitura de Documentos de Modelos de Perguntas e Respostas no Domínio Esportivo

Laura Fernandes Camargos[1], **Leonardo Mauro Pereira Moraes**[1,2], **Cristina Dutra Aguiar**[1]

[1]Departamento de Ciências de Computação, Universidade de São Paulo, São Carlos, Brasil

[2]Centro de Excelência Data & AI, Amaris Consulting, Vernier, Genebra, Suíça

{laura.camargos, leonardo.mauro}@usp.br, cdac@icmc.usp.br

Abstract. *This paper investigates document reader models in question-answering systems. These models analyze pre-selected documents using advanced natural language processing techniques to understand the context and semantics of the text, producing relevant answers. We compare the models BERT, DistilBERT, MiniLM, RoBERTa, and ELECTRA, considering their ability to answer questions related to the sports domain. The results demonstrated that the RoBERTa model provided the best performance considering Exact Match and F-Score, while the DistilBERT model provided the best execution time.*

Resumo. *Neste artigo são investigados modelos de leitor de documentos em sistemas de perguntas e respostas. Esses modelos analisam documentos pré-selecionados usando técnicas avançadas de processamento de linguagem natural para entender o contexto e a semântica do texto, produzindo respostas relevantes. São comparados os modelos BERT, DistilBERT, MiniLM, RoBERTa e ELECTRA, considerando a capacidade desses em responder perguntas referentes ao domínio esportivo. Os resultados obtidos demonstraram que o modelo RoBERTa proveu melhor desempenho para as métricas Exact Match e F-Score, e o modelo DistilBERT proveu melhor tempo de execução.*

1. Introdução

Question Answering (QA, em português, Perguntas e Respostas) é uma abordagem de Processamento de Linguagem Natural (PLN) promissora no contexto de recuperação de informação [Mishra and Jain 2016], sendo usada para os mais diferentes fins. Por exemplo, pode-se citar os sistemas de QA [Hirschman and Gaizauskas 2001], tais como ChatGPT[1] e Amazon Kendra[2]. Essa abordagem também tem sido empregada desde o desenvolvimento de assistentes virtuais até a aplicações de suporte ao cliente.

Os sistemas de QA são projetados para compreender a linguagem natural e extrair informações relevantes de grandes volumes de dados textuais, empregando modelos especializados geralmente compostos por duas etapas. Primeiramente, o *recuperador de documentos* (em inglês, *Document Retriever*) recupera os documentos, ou pedaços dos documentos tal como parágrafos, mais pertinentes que possam conter respostas a uma

[1]ChatGPT - openai.com/blog/chatgpt/

[2]Kedra - aws.amazon.com/en/kendra/

pergunta. Em seguida, o *leitor de documentos* (em inglês, *Document Reader*) examina cuidadosamente os contextos e extrai respostas a partir dos documentos recuperados.

A qualidade dos sistemas de QA depende da capacidade desses em compreender e responder de forma adequada às perguntas dos usuários. Em especial, sistemas de QA são dependentes dos dados usados como fonte, os quais fornecem conhecimento necessário sobre o domínio em questão e suas diferentes terminologias. Existe, assim, a necessidade de se investigar modelos de QA de forma que eles possam garantir melhor qualidade das respostas providas, principalmente quando se consideram domínios específicos.

Um domínio específico de grande interesse é o de esportes, uma vez que ele tem impactos sociais, emocionais, comportamentais e econômicos, além de abranger os mais variados tipos de usuários. Uma dificuldade de se investigar sistemas de QA dentro do domínio de esportes é a escassez de conjuntos de dados. Consequentemente, poucos modelos de QA são treinados e mostram-se aptos a responderem especificamente perguntas sobre esportes vislumbrando oferecer respostas mais completas e contextualizadas para uma variedade de perguntas nesse domínio.

Este artigo tem como objetivo investigar modelos de leitor de documentos, considerando o domínio esportivo. Visa-se identificar o desempenho desses modelos frente às características das perguntas realizadas pelos usuários no universo dos esportes, especialmente considerando basquete. São avaliados os seguintes modelos: BERT [Devlin et al. 2018], DistilBERT [Sanh et al. 2019], MiniLM [Wang et al. 2020], RoBERTa [Liu et al. 2019] e ELECTRA [Clark et al. 2020].

Para a condução dos testes de desempenho comparativos, são utilizados a arquitetura BigQA [Moraes et al. 2023] e o conjunto de dados QASports [Jardim et al. 2023], os quais representam o estado da arte na proposta de arquitetura de *Big Data Question Answering* e de conjunto de dados esportivos, respectivamente. Cada um dos modelos foi avaliado considerando uma variedade de perguntas relacionadas ao basquete. Os resultados obtidos mostraram que o modelo RoBERTa proveu o melhor resultado para as métricas *Exact Match* e *F-Score* e o modelo DistilBERT para a métrica *Wall Time*.

O artigo está organizado como segue. Na seção 2 são resumidos trabalhos relacionados. Na seção 3 são descritos as arquitetura BigQA e o conjunto de dados QASports. Na seção 4 é detalhado o desenvolvimento do trabalho. Na seção 5 são descritos os experimentos e as análises realizadas. Na seção 6 são feitas as conclusões do artigo.

2. Trabalhos Relacionados

Os trabalhos relacionados são agrupados em duas classes. A classe 1 engloba estudos que propõem novos modelos de leitor de documentos. Usualmente quando um novo modelo é proposto, são realizadas comparações com outros modelos já existentes para mostrar as vantagens da nova abordagem proposta. O modelo utilizado como *baseline* nessas comparações é o BERT [Devlin et al. 2018], desde que ele revolucionou a área de perguntas e respostas ao introduzir um novo modelo de linguagem. A classe 2 contém estudos que investigam modelos de leitor de documentos sem a proposta de um novo modelo e consideram um domínio específico, assim como o presente trabalho.

Como observado na Tabela 1, o presente trabalho investiga mais modelos de leitor de documentos. Outro diferencial é que, neste trabalho, define-se um ambiente de teste

Tabela 1. Comparação de Trabalhos Relacionados.

Classe	Estudo	Modelos Comparados	Domínio
1	DistilBERT [Sanh et al. 2019]	BERT, ELMo	Geral
	MiniLM [Wang et al. 2020]	BERT, DistilBERT, TinyBERT	Geral
	RoBERTa [Liu et al. 2019]	BERT, XLNet, DistilBERT	Geral
	ELECTRA [Clark et al. 2020]	BERT, RoBERTa, XLNet, GPT	Geral
2	DrQA [Chen et al. 2017]	AttentiveReader, TF-IDF	Wikipédia
	[Aurpa et al. 2022]	BERT, ELECTRA	Língua Bangla
	Presente Trabalho	BERT, DistilBERT, MiniLM, RoBERTa, ELECTRA	Esportes

de forma que os modelos são comparados considerando a mesma configuração, conjunto de dados e parâmetros de testes, garantindo homogeneidade nas comparações realizadas. Adicionalmente, o presente trabalho é o primeiro a considerar o domínio de esportes.

3. Fundamentação Teórica

A seção 3.1 detalha a arquitetura de perguntas e respostas *BigQA*. Enquanto, a seção 3.2 descreve as características do conjunto de dados esportivo QASports.

3.1. Arquitetura BigQA

BigQA [Moraes et al. 2023] é uma arquitetura projetada para apoiar o desenvolvimento de sistemas de perguntas e respostas escaláveis, considerando as características de volume, velocidade e variedade de *big data* [Chebbi et al. 2015]. Esses princípios consideram quesitos: (i) negócio, o fato do usuário ter que recuperar uma resposta apropriada à sua pergunta e poder acessar somente documentos permitidos; (ii) dados, como o fato do sistema ter que persistir os documentos e trabalhar com documentos de fontes heterogêneas; e (iii) técnicos, tais como modularidade e flexibilidade, segurança e análise.

Na Figura 1 é ilustrada a visão geral de BigQA, suas camadas e respectivos componentes. Por meio das camadas *Input* e *Big Data Storage*, BigQA é capaz de processar e armazenar documentos em diferentes formatos e referentes a domínios variados, os quais encontram-se espalhados em diversas bases de conhecimento como *data lakes* e *document stores*. A camada *Big Querying* provê uma maneira unificada e precisa de recuperar documentos e gerar respostas às perguntas dos usuários, as quais podem ser feitas por meio da camada *Communication*. Os artefatos de segurança são de responsabilidade da camada *Security*. Por fim, a camada *insights* é responsável por prover suporte à análise de dados por meio de modelos de inteligência artificial, de consultas analíticas realizadas em ambientes de *data warehousing* e de ferramentas geradoras de relatórios.

Neste artigo, é explorada a camada *Big Querying*, desde que ela incorpora o recuperador e o leitor de documentos dos modelos de QA. Especificamente, é investigado o leitor de documentos, o qual analisa os documentos selecionados pelo recuperador de documentos, aplica modelos de compreensão de leitura a segmentos de texto para extração de respostas e retorna essas resultados aos usuários.

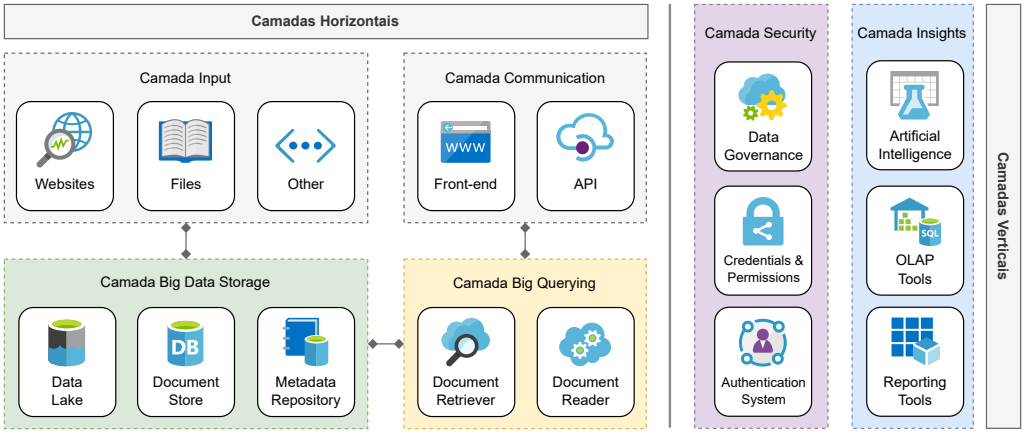

Figura 1. Visão geral da arquitetura BigQA. Adaptado de [Moraes et al. 2023].

3.2. Conjunto de Dados QASports

QASports [Jardim et al. 2023][3] é um conjunto de dados pioneiro no campo de perguntas e respostas esportivas. Ele oferece uma variedade abrangente de tópicos e informações sobre os três esportes mais populares do mundo, a saber: futebol, futebol americano e basquete. O conjunto de dados compreende mais de 1,5 milhão de perguntas e respostas extraídas de 54 mil páginas wiki previamente processadas.

São armazenadas as seguintes coleções de dados em QASports: (i) páginas no formato JSON; (ii) arquivos contexto no formato CSV; e (iii) perguntas e respostas no formato de triplas contexto-pergunta-resposta. Adicionalmente, QASports é organizado em diferentes divisões e subconjuntos de dados. As divisões dividem o conjunto de dados em partes específicas para treinamento, validação e teste, enquanto os subconjuntos segmentam os dados em categorias de basquete, futebol americano e futebol. Essa organização facilita a seleção de subconjuntos específicos para diferentes tarefas.

Características distintivas de QASports referem-se ao grande volume de perguntas e respostas, à diversidade de exemplos e à quantidade de amostras linguísticas esportivas. Como resultado, esse conjunto de dados proporciona diversos desafios e vantagens para investigar leitores de documentos dos modelos de QA.

4. Investigação dos Modelos de Leitor de Documentos

Foram implementados cinco modelos baseados em redes neurais para instanciar o leitor de documentos da camada *Big Querying* da arquitetura BigQA: BERT [Devlin et al. 2018], DistilBERT [Sanh et al. 2019], MiniLM [Wang et al. 2020], RoBERTa [Liu et al. 2019] e ELECTRA [Clark et al. 2020]. Esses modelos são pré-treinados em perguntas de domínio geral, e destacam-se por empregar técnicas avançadas de aprendizado profundo e serem capazes de aprender representações semânticas complexas.

Para avaliar modelos de leitor de documentos, é recomendado utilizar computadores com processamento paralelo e GPUs. Portanto, foi usado o Google Colab, uma plataforma de *Jupyter Notebooks* baseada em computação em nuvem que oferece suporte a GPUs e TPUs. Foi empregada uma GPU NVIDIA Tesla T4, acompanhada por 16 GB

[3]QASports - huggingface.co/datasets/PedroCJardim/QASports

de memória RAM. Google Colab fornece um ambiente de execução para a linguagem de programação Python, além de diversas bibliotecas e pacotes pré-configurados, necessários para executar tarefas como a avaliação de modelos investigados.

Dentre os dados esportivos disponíveis no conjunto de dados QASports, foram escolhidos os dados de basquete porque a versão gratuita do Google Colab não foi capaz de oferecer suporte para processar o grande volume de perguntas e respostas presente em QASports. Portanto, os modelos implementados foram avaliados considerando esse subconjunto de dados, representado aproximadamente 23.200 documentos de validação.

A avaliação dos modelos investigados foi realizada para medir a eficiência desses em termos das métricas *Exact Match*, *F-Score* e *Wall Time* [Rajpurkar et al. 2016]. *Exact Match* mede a porcentagem de respostas corretas preditas pelo modelo. *F-score* calcula a sobreposição média entre a previsão realizada pelo modelo e a resposta correta definida como *ground truth*. O *Wall Time*, utilizado apenas como uma noção geral de tempo de execução e não uma comparação exata, reflete o tempo total gasto pelo modelo para processar o conjunto de dados e gerar as respostas. O código está disponível no GitHub[4].

5. Resultados

Como ilustrado na Figura 2, a investigação dos leitores de documentos envolveu os seguintes passos: (1) pré-processamento dos documentos, e criação dos rótulos de avaliação (*labels*), para garantir a consistência e padronização dos dados; (2) configuração dos modelos de leitura de documentos para extração das respostas; (3) avaliação utilizando uma *pipeline* configurada para avaliar as etiquetas e os documentos correspondentes em termos das métricas Exact Match, F-Score e Wall Time.

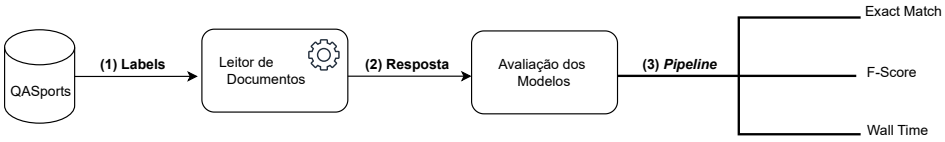

Figura 2. Etapas para a Investigação dos Modelos de Leitor de Documentos.

5.1. Análise Comparativa

A tabela 2 apresenta os resultados dos experimentos realizados sob os modelos BERT, DistilBERT, ELECTRA, MiniLM e RoBERTa utilizando o subconjunto de dados de basquete de QASports. O modelo RoBERTa destacou-se como o mais eficaz em termos de Exact Match e F-Score, demonstrando maior precisão em comparação com os outros modelos. No entanto, seu tempo de execução foi significativamente maior do que os demais. Enquanto, o MiniLM apresentou métricas competitivas de F-Score e Exact Match, mas seu elevado tempo de execução pode limitar sua aplicabilidade.

Demonstra-se, portanto, um *trade-off* entre os resultados providos pelas métricas Exact Match e F-Score e os resultados de Wall Time. Na leitura de documentos, o modelo deve processar a pergunta e o documento, bem como retornar uma resposta adequada. Assim, os modelos de redes neurais mais complexos apresentam melhores resultados para

[4]Código - github.com/leomaurodesenv/big-qa-architecture

Tabela 2. Avaliação dos Modelos Investigados.

	Exact Match	F-Score	Wall Time
BERT	0.576	0.604	57min 11s
DistilBERT	0.549	0.582	56min 32s
ELECTRA	0.547	0.586	59min 16s
MiniLM	0.579	0.606	4h 2min 58s
RoBERTa	0.656	0.656	1h 24min 47s

Exact Match e F-Score. Contudo, à medida que o modelo se torna mais complexo, o seu custo de inferência aumenta, requerendo maior tempo para sua execução.

5.2. Ajuste Fino do Modelo DistilBERT

Analisando-se o desempenho do modelo DistilBERT (Tabela 2), pode-se observar que ele proveu o menor tempo de execução. Assim, optou-se por realizar um ajuste fino de desempenho neste modelo, usando os 5.000 primeiros dados de treinamento de basquete. Com isto, o DistilBERT alcançou um Exact Match de 0.697 e um F-Score de 0.763, superando os demais modelos pré-treinados e apresentados na Tabela 2.

Durante o ajuste fino, o modelo foi inicializado com pesos pré-treinados e configurado para a tarefa de perguntas e respostas extrativas. Também foram definidos o número de épocas e o tamanho dos lotes de treinamento e avaliação. No treinamento, foram monitoradas as métricas "Epoch", "Training Loss" e "Validation Loss" para avaliar o progresso e o desempenho de DistilBERT. Esse modelo está disponível no HuggingFace[5].

6. Conclusão

Neste trabalho foram investigados diferentes leitores de documentos (BERT, DistilBERT, MinLM, RoBERTa, ELECTRA) presentes em sistemas de perguntas e respostas. Visando prover homogeneidade nas comparações, foi utilizada a arquitetura BigQA e o conjunto de dados esportivos QASports. O processo envolveu a construção de uma *pipeline* na qual diversas perguntas foram feitas a uma série de documentos de basquete, visando encontrar respostas relevantes. De acordo com os resultados obtidos, o modelo RoBERTa proveu melhor desempenho para as métricas F-Score e Exact Match, enquanto que o modelo DistilBERT proveu melhor desempenho para a métrica Wall Time. No entanto, observa-se que, apesar do desempenho geral, esses modelos são pré-treinados em dados de domínios gerais. Desta forma, o ajuste fino é fundamental para adaptar os modelos ao domínio esportivo, afim de melhorar significativamente seu desempenho nesse contexto.

Como primeiro trabalho futuro pretende-se investigar os modelos considerando o conjunto de dados completo de QASports. Desde que a versão do Google Colab não oferece suporte para a manipulação do grande volume de dados armazenado em QASports, esse trabalho demandará novas decisões de projeto e adaptação das investigações realizadas. Outro trabalho futuro é o ajuste fino de desempenho dos modelos investigados. Os resultados promissores obtidos para o modelo DistilBERT demonstram que o ajuste fino também deve ser realizado para os demais modelos.

[5]Fine-tuning DistilBERT - distilbert-qasports-basket-small

Agradecimentos

Agradecemos à Amaris Consulting, à Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), à Agência Federal de Pesquisa CNPq e à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, Brasil (CAPES) [Código Financeiro 001] pelo apoio a este trabalho. L. F. Camargos foi apoiada pelo Programa Unificado de Bolsas/USP. C. D. Aguiar foi apoiada pela bolsa #2018/22277-8, FAPESP. L. M. P. Moraes foi apoiado pela Amaris Consulting.

Referências

Aurpa, T. T., Rifat, R. K., Ahmed, M. S., Anwar, M. M., and Ali, A. B. M. S. (2022). Reading comprehension based question answering system in Bangla language with transformer-based learning. *Heliyon*, 8(10):e11052.

Chebbi, I., Boulila, W., and Farah, I. R. (2015). Big data: Concepts, challenges and applications. In *Computational Collective Intelligence. Lecture Notes in Computer Science*, volume 9330, pages 638–647.

Chen, D., Fisch, A., Weston, J., and Bordes, A. (2017). Reading Wikipedia to answer open-domain questions. *CoRR*, abs/1704.00051.

Clark, K., Luong, M., Le, Q. V., and Manning, C. D. (2020). ELECTRA: pre-training text encoders as discriminators rather than generators. *CoRR*, abs/2003.10555.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Hirschman, L. and Gaizauskas, R. (2001). Natural language question answering: the view from here. *Natural Language Engineering*, 7(4):275–300.

Jardim, P. C., Moraes, L. M. P., and Aguiar, C. D. (2023). QASports: A question answering dataset about sports. In *Proceedings of the Brazilian Symposium on Databases: Dataset Showcase Workshop*, pages 1–12.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Mishra, A. and Jain, S. K. (2016). A survey on question answering systems with classification. *Journal of King Saud University - Computer and Information Sciences*, 28(3):345–361.

Moraes, L. M. P., Jardim, P., and Aguiar, C. D. (2023). Design principles and a software reference architecture for big data question answering systems. In *Proc. of the 25th International Conference on Enterprise Information Systems*, pages 57–67.

Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. (2016). Squad: 100,000+ questions for machine comprehension of text. *CoRR*, abs/1606.05250.

Sanh, V., Debut, L., Chaumond, J., and Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108.

Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., and Zhou, M. (2020). Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Advances in Neural Information Processing Systems*, volume 33, pages 5776–5788.