PROCESSAMENTO DE LINGUAGEM NATURAL APLICADO À CLASSIFICAÇÃO DE MQT EM CONSTRUÇÃO

Luís Jacques de Sousa [(1)], João Poças Martins [(2)], Luís Sanhudo [(3)]

(1) CONSTRUCT, Faculdade de Engenharia / BUILT CoLAB - The Collaborative Laboratory for the Built Environment of the Future, 0000-0002-0789-9368
(2) CONSTRUCT, Faculdade de Engenharia, Universidade do Porto, 0000-0001-9878-3792
(3) BUILT CoLAB - The Collaborative Laboratory for the Built Environment of the Future, 0000-0002-2578-6981

Resumo
Nos concursos públicos de Construção, os Mapas de Quantidades de Trabalhos (MQT) descrevem e quantificam as tarefas a executar. A classificação destas descrições consiste em interpretar o texto e associá-lo às categorias técnicas reconhecidas pelos concorrentes, permitindo atribuir preços e elaborar o orçamento do concurso. Este processo é tradicionalmente realizado manualmente por especialistas, dependendo da sua experiência individual e sendo moroso e suscetível a erros. Neste estudo, propomos uma abordagem baseada em Processamento de Linguagem Natural (NLP) para acelerar a classificação das descrições de MQT. O objetivo é apoiar o especialista, permitindo-lhe focar-se na validação das classificações e aumentando a eficiência do processo.
Foram comparados métodos tradicionais, como TF-IDF e redes LSTM, com arquiteturas baseadas em transformadores, nomeadamente BERT-HAN, LLaMA treinado e GPT-4o-mini. O conjunto de dados utilizado inclui mais de 20 000 descrições anotadas hierarquicamente (Capítulos, Subcapítulos e Itens) de trabalhos provenientes de projectos de contratação pública em Portugal. Os resultados demonstram que modelos de grande escala, baseados em transformadores, superam as abordagens convencionais, apresentando maior capacidade de generalização em tarefas complexas e específicas do domínio da Construção. O modelo LLaMA obteve o melhor desempenho global no conjunto de validação, enquanto o BERT-HAN registou a maior *recall* nos níveis mais profundos da hierarquia.
Este trabalho evidencia o potencial de soluções de inteligência artificial para melhorar a eficiência na classificação de MQTs, contribuindo para acelerar o processamento de dados e documentos e promover maior consistência na estimativa de custos e nos fluxos de trabalho de contratação na Construção.

1. Introdução

A contratação na construção é uma fase crítica para os empreiteiros, pois determina a sua capacidade de obter contratos e gerar receita. Para competir eficazmente, os empreiteiros devem

submeter propostas detalhadas, o que implica analisar as exigências do dono de obra e alinhá-las com uma base de dados interna de tarefas e preços. Esta informação é organizada num MQT, um documento fundamental que descreve trabalhos, quantidades e custos [1].
Tradicionalmente, a classificação de MQTs é um processo manual, moroso e fortemente dependente da experiência dos técnicos, com reduzido apoio em dados [2]. Esta abordagem, aliada a restrições de tempo, aumenta o risco de erros, conduzindo a ineficiências, desvios orçamentais e potenciais conflitos contratuais. Durante a fase de construção, MQTs precisos são essenciais para o controlo financeiro, planeamento e alocação de recursos, sendo determinantes para o cumprimento de prazos e custos [3].
A automatização da classificação de texto na construção tem-se revelado desafiante devido à natureza não estruturada dos contratos, à escassez de dados rotulados e à ambiguidade da linguagem técnica do setor [4]. No entanto, avanços recentes na disponibilidade de dados anotados e na geração de dados sintéticos com recurso à inteligência artificial generativa abriram novas oportunidades para o desenvolvimento de modelos de classificação neste domínio [1].
Modelos baseados em *transformers*, como o BERT e os Large Language Models (LLMs), têm demonstrado elevado desempenho em tarefas de classificação multi-rótulo, nas quais uma descrição pode pertencer a várias categorias [5]. Estas técnicas já foram aplicadas com sucesso em áreas como análise contratual, gestão de energia e avaliação de risco [6]. Contudo, os estudos existentes sobre classificação de MQTs focam-se sobretudo em categorizações simples, ignorando a sua estrutura hierárquica em capítulos e subcapítulos [7]. Adicionalmente, não existem trabalhos que explorem modelos *transformer* para classificação de MQTs, nem estudos aplicados a documentos em língua portuguesa.
Este artigo aborda diretamente esta lacuna, propondo uma metodologia inovadora baseada em algoritmos *transformer* para automatizar a classificação de tarefas em MQTs. O objetivo é mapear os requisitos do MQT do dono de obra para uma estrutura normalizada reconhecida pelo empreiteiro, apoiando a tomada de decisão no processo de contratação.
O estudo avalia empiricamente três abordagens: um modelo *few-shot* baseado no ChatGPT, uma rede neural baseada em BERT e um modelo LlaMA refinado. A comparação incide sobre desempenho, adaptabilidade e aplicabilidade prática no contexto da construção. O artigo está organizado da seguinte forma: a Secção 2 apresenta o estado da arte, a Secção 3 descreve a metodologia, a Secção 4 apresenta os modelos desenvolvidos, a Secção 5 analisa os resultados experimentais e a Secção 6 apresenta as conclusões e contributos principais.

2. Estado da Arte

O deep learning tem vindo a apoiar a gestão de projetos de construção, melhorando a previsão, a tomada de decisão e a mitigação de incertezas [8]. Os avanços recentes em NLP permitiram o desenvolvimento de modelos de classificação de texto mais precisos e escaláveis [31]. Estas técnicas têm sido aplicadas em tarefas como verificação automática de regras, classificação documental e identificação de riscos contratuais [9].
Os modelos baseados em *transformers* apresentam desempenho de referência em tarefas de classificação de texto, utilizando mecanismos de autoatenção para captar relações contextuais [10]. Arquiteturas como BERT [5], GPT [11] e LLaMA [12] são amplamente utilizadas em diferentes tarefas de NLP. O BERT destaca-se pela sua capacidade de compreensão bidirecional

do contexto, enquanto o GPT e o LLaMA suportam cenários de aprendizagem *zero-sho*t e *few-shot*. Estes modelos têm demonstrado elevado desempenho em classificação multi-rótulo em domínios como o financeiro e o jurídico [13, 14]. No setor da construção, os *transformers* têm sido aplicados, por exemplo, na correção de erros em documentos técnicos e no reconhecimento de entidades [15], bem como na geração de dados sintéticos para mitigar a escassez de dados anotados [16]. Embora as aplicações diretas à classificação de MQTs sejam ainda limitadas, os estudos existentes evidenciam o potencial destes modelos para enfrentar desafios específicos do setor da construção.

A automatização da classificação de MQTs oferece vantagens significativas nos processos de concurso e contratação, onde os empreiteiros operam sob prazos apertados e elevado risco financeiro. A análise manual de MQTs é morosa, repetitiva e propensa a erros, representando uma carga elevada para os medidores orçamentistas. Enquanto na América do Norte existem sistemas de classificação normalizados, muitos países europeus carecem de um enquadramento unificado, resultando em documentos heterogéneos. Sistemas de classificação automática baseados em IA permitem mapear diferentes formatos de MQTs para bases de dados internas de custos, aumentando a eficiência, a interoperabilidade e a precisão. A automatização acelera a preparação de propostas, reduz erros de classificação e permite que os profissionais se concentrem em tarefas de maior valor acrescentado, como a análise de risco e o planeamento estratégico de custos [8].

Apesar dos avanços na classificação de texto com modelos baseados em *transformers*, a sua aplicação à classificação de tarefas em MQTs permanece pouco explorada. A maioria dos estudos existentes foca-se em classificação de rótulo único, não considerando a natureza hierárquica e multi-rótulo dos MQTs. Os MQTs estão organizados em capítulos e subcapítulos com relações hierárquicas do tipo pai-filho, podendo uma única tarefa pertencer a múltiplas subcategorias. A ausência de modelos capazes de preservar esta estrutura resulta frequentemente em classificações genéricas e pouco informativas. Adicionalmente, os textos dos MQTs são curtos, técnicos e ambíguos, originando cenários de classificação multi-rótulo difíceis de tratar com abordagens tradicionais. A escassez de conjuntos de dados anotados e a inexistência de *benchmarks* limitam ainda o treino, a avaliação e a reprodutibilidade dos modelos. A anotação manual exige conhecimento especializado, sendo dispendiosa e inconsistente.

Em síntese, embora o *deep learning* apresente resultados promissores na classificação de tarefas na construção, persistem desafios significativos. A classificação de MQTs requer modelos capazes de lidar com estruturas hierárquicas e múltiplos rótulos, sendo particularmente escassa a investigação aplicada a documentos em língua portuguesa. Este trabalho procura colmatar estas lacunas através do desenvolvimento de modelos baseados em *transformer*s para classificação hierárquica multi-rótulo, captando de forma mais eficaz a complexidade da documentação de custos na construção.

3. Métodos

Este artigo propõe uma abordagem inovadora baseada em algoritmos *transformer* para a classificação automática de tarefas em MQTs, tendo em conta os desafios identificados na secção anterior.

3.1. Framework

Conforme ilustrado na Figura 1, o modelo utiliza algoritmos de *deep learning* para classificar tarefas segundo uma estrutura hierárquica predefinida, com base num conjunto de dados de MQTs anotados. No início de um novo processo de contratação, o MQT do dono de obra é introduzido no modelo, que classifica cada tarefa de acordo com um formato hierárquico reconhecido. Este processo automatiza a tradução do MQT do cliente para o enquadramento interno do empreiteiro, convertendo uma tarefa tradicionalmente manual num processo de validação por parte dos técnicos.

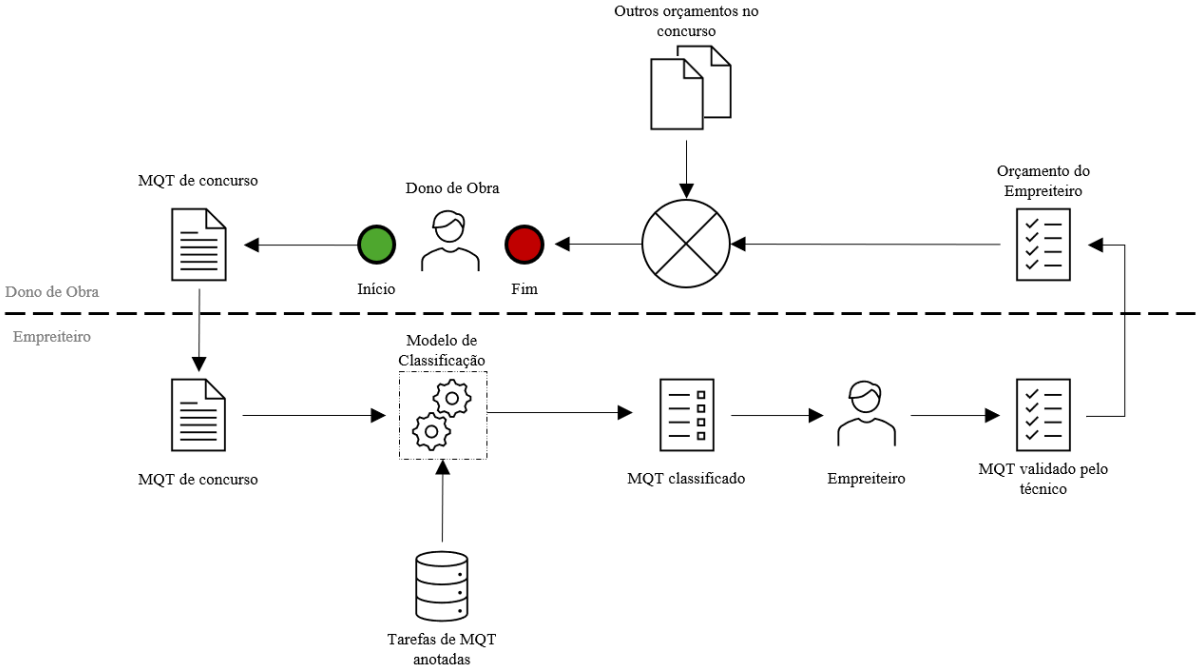

Figura 1: *Framework* para classificação automática de MQTs inserido no processo habitual de contratação.

A integração deste sistema no fluxo de contratação procura obter benefícios claros, nomeadamente:

- Eficiência temporal, reduzindo o tempo necessário para a classificação;
- Apoio à decisão baseado em dados, aumentando a fiabilidade do orçamento e planeamento;
- Redução de erros, minimizando falhas humanas na classificação de tarefas.

3.2. Obtenção e Estruturação dos dados

O conjunto de dados utilizado foi construído a partir de MQTs públicos portugueses relativos a projetos de edifícios (ver [17]) e inclui tarefas classificadas em três níveis hierárquicos:

- 12 capítulos, correspondentes a categorias gerais de construção;
- 58 subcapítulos, representando subdivisões funcionais;
- 72 itens, descrevendo tarefas específicas.

A estrutura hierárquica adotada reflete o sistema interno de classificação de uma empresa de construção colaboradora e não foi alterada, uma vez que corresponde à sua prática operacional

e documentação histórica. O objetivo não foi redefinir a taxonomia, mas desenvolver modelos capazes de aprender e operar dentro dessa estrutura pré-existente.
Na prática, nem todos os capítulos incluem subcapítulos, nem todos os subcapítulos contêm itens. Para garantir uma hierarquia consistente de três níveis, foram introduzidas classes auxiliares nos níveis em falta. Estas classes funcionam apenas como marcadores técnicos, preservando a semântica original do MQT e permitindo um processamento uniforme pelos modelos de classificação hierárquica.
Cada tarefa segue uma classificação estruturada, correspondendo aos níveis de capítulo, subcapítulo e item, como por exemplo:

"Estruturas de Betão Armado → Lajes → Lajes Maciças"

O conjunto de dados inclui 7 096 tarefas reais anotadas, extraídas de repositórios públicos. Para mitigar a escassez e o desequilíbrio dos dados, foi aplicada uma estratégia de aumento sintético de dados, triplicando o conjunto inicial e originando três subconjuntos com mais de 20 000 exemplos cada [17]. Este processo permitiu equilibrar as classes ao nível dos subcapítulos, garantindo o mesmo número de exemplos por classe.
Apesar deste equilíbrio parcial, a natureza multi-rótulo e hierárquica dos MQTs dificulta a obtenção de uma distribuição perfeitamente uniforme. Ao nível dos itens, o conjunto de dados permanece fortemente desequilibrado, com predominância da classe Item Auxiliar, utilizada quando não existem itens associados. Este desequilíbrio representa um desafio significativo, exigindo cuidados adicionais no desenho e avaliação dos modelos de classificação, de forma a evitar previsões enviesadas e garantir capacidade de generalização.

4. Modelos

Esta secção descreve as arquiteturas de classificação avaliadas neste estudo. Foram testadas três abordagens principais: um modelo *few-shot* baseado no GPT-4o Mini, um modelo *few-shot* refinado com LLaMA e um classificador hierárquico baseado em BERT. Adicionalmente, foram incluídos dois modelos de referência para comparação: uma Rede Neural Hierárquica (HNN) baseada em TF-IDF e uma HNN baseada em LSTM, permitindo avaliar o desempenho de abordagens clássicas e baseadas em *transformer*s.
A escolha destes modelos reflete a evolução progressiva do trabalho ao longo do tempo. Inicialmente, foram exploradas arquiteturas tradicionais, como TF-IDF e LSTM, devido à sua simplicidade. Contudo, a sua limitação na captura de contexto e dependências hierárquicas motivou a adoção de modelos *transformer*, em particular o BERT. Com o surgimento de LLMs mais acessíveis, foram integradas abordagens *few-shot* com GPT-4o Mini e LLaMA, representando a fase mais recente da investigação.

- **Modelo classificador hierárquico baseado em TF-IDF:** Implementação de uma HNN clássica baseada em vetores TF-IDF, seguindo uma estratégia de classificação *top-down*. As saídas dos níveis superiores são concatenadas com os vetores de entrada. Apesar de simples, o modelo serve essencialmente como *baseline* para comparação com arquiteturas mais avançadas.

- **Modelo classificador hierárquico baseado em LSTM:** Avaliação de uma HNN baseada em LSTM, destinada a capturar informação sequencial e contexto local nas descrições das tarefas. As representações geradas alimentam os classificadores de capítulo, subcapítulo e item, permitindo analisar o compromisso entre complexidade, custo computacional e desempenho.
- **Modelo classificador hierárquico baseado em BERT:** Desenvolvimento de um classificador hierárquico com uma *Hierarchical Attention Network* (HAN) e BERT (prajjwal1/bert-small) como extrator de características. O modelo realiza previsões sequenciais de capítulo, subcapítulo e item, preservando dependências hierárquicas.
- **Modelo few-shot com GPT-4o mini:** Classificação de tarefas MQT através de aprendizagem *few-shot*, recorrendo a prompting em vez de treino supervisionado. O prompt incluiu exemplos anotados, a estrutura hierárquica dos MQTs e o formato de saída esperado (capítulos, subcapítulos e itens).
- **Modelo few-shot com LlaMA:** Utilização de um modelo LLaMA 3.2 8B, quantizado a 4-bit e refinado com Unsloth e LoRA, explorando a capacidade generativa de LLMs em regime *few-shot*. O processo incluiu pré-processamento dos textos, engenharia de prompts e pós-processamento das saídas. O modelo foi treinado com mais de 18 000 exemplos anotados e avaliado de forma consistente com as restantes abordagens.

5. Resultados

Esta secção apresenta os resultados e está dividida em duas partes, a primeira acerca dos resultados de treino das redes neuronais classificadoras desenvolvidas e a segunda acerca dos resultados de validação de todos os modelos.

5.1. Desempenho de treino

A Tabela 1 apresenta os resultados de treino e teste dos modelos TF-IDF HNN, LSTM-HNN e BERT-HAN, avaliados nos níveis hierárquicos de capítulo, subcapítulo e item com base nas métricas de precisão, *recall*, *F1-score* e *Hamming Loss*.

A *Hamming Loss* foi utilizada como métrica complementar por ser particularmente adequada a cenários de classificação multi-rótulo. Esta métrica mede a proporção de rótulos incorretamente previstos e permite avaliar o erro global do modelo, mesmo em contextos com forte desequilíbrio de classes.

Os modelos TF-IDF e LSTM apresentaram bom desempenho ao nível do capítulo (F1 > 0,86), mas falharam na classificação de subcapítulos, com precisão e *recall* nulas, evidenciando limitações na captura de relações hierárquicas e semânticas mais finas. Ao nível do item, obtiveram *F1-scores* moderados (~0,70), fortemente influenciados pelo desequilíbrio dos dados e pela predominância da classe auxiliar.

Em contraste, o modelo BERT-HAN apresentou desempenho superior em todos os níveis. Alcançou um *F1-score* de 0,96 ao nível do capítulo, com *Hamming Loss* reduzida, e foi o único a obter resultados relevantes ao nível do subcapítulo (F1 = 0,66). Ao nível do item, manteve um desempenho robusto (F1 = 0,76), apesar da elevada granularidade e do forte desequilíbrio das classes.

De forma global, os resultados demonstram que modelos clássicos são insuficientes para capturar a complexidade dos MQTs, enquanto modelos *transformer* pré-treinados integrados em arquiteturas hierárquicas permitem melhorias substanciais em todos os níveis de classificação.

Tabela 1: Métricas de treino por modelo e nível de *output*

Modelo	**Nível**	**Treino**				**Teste**			
		Precisão	*Recall*	*F1-score*	*Hamming Loss*	Precisão	*Recall*	*F1-score*	*Hamming Loss*
TF-IDF HNN	Capítulo	0.9459	0.9058	0.9254	0.0123	0.9490	0.8992	0.9234	0.0126
	Subcapítulo	0.00	0.00	0.00	0.0183	0.00	0.00	0.00	0.0184
	Item	0.7613	0.6599	0.7077	0.0088	0.7598	0.6556	0.7031	0.0089
LSTM HNN	Capítulo	0.9273	0.8072	0.8624	0.0215	0.9408	0.8061	0.8681	0.0204
	Subcapítulo	0.00	0.00	0.00	0.0183	0.00	0.00	0.00	0.0184
	Item	0.7613	0.6599	0.7077	0.0088	0.7598	0.6556	0.7031	0.0089
BERT-HAN	Capítulo	0.9820	0.9789	0.9804	0.0033	0.9614	0.9553	0.9583	0.0070
	Subcapítulo	0.9498	0.5458	0.6927	0.0088	0.9079	0.5138	0.6552	0.0099
	Item	0.9825	0.6475	0.7785	0.0058	0.9540	0.6380	0.7645	0.0063

5.2. Desempenho de validação

A Tabela 2 apresenta a comparação do desempenho dos modelos no conjunto de validação.

Os modelos clássicos baseados em TF-IDF e LSTM demonstraram limitações significativas à medida que a granularidade da classificação aumenta. Embora o LSTM-HNN apresente bom desempenho ao nível do capítulo, ambos os modelos falham na distinção de subcapítulos. Estes resultados confirmam a dificuldade de modelos não contextuais ou sem *pretraining* em capturar dependências semânticas e hierárquicas complexas.

O modelo BERT-HAN apresentou desempenho robusto e consistente em todos os níveis hierárquicos, beneficiando da combinação de *embeddings* pré-treinados, atenção hierárquica, *threshold tuning* e *masking*. Destaca-se como o melhor modelo supervisionado tradicional, especialmente ao nível do subcapítulo, onde mantém um equilíbrio adequado entre precisão e *recall*, apesar do forte desequilíbrio de classes.

O modelo GPT-4o-mini, avaliado num contexto *few-shot* e sem afinação específica, obteve resultados competitivos, particularmente ao nível do capítulo, com degradação gradual à medida que aumenta a granularidade. Apesar de menor *recall* nos níveis inferiores, o modelo demonstra elevado potencial como solução rápida e de baixo custo em cenários com poucos dados anotados.

Tabela 2: Métricas de validação por modelo e nível de *output*

Modelo	**Nível**	**Precisão**	***Recall***	***F1-score (weighted)***	***Hamming Loss***
TF-IDF HNN	Capítulo	0.310	0.136	0.174	0.101
Total Params:	Subcapítulo	0.040	0.378	0.108	0.235
356,718	Item	0.654	0.439	0.347	0.016
	Geral	0.090	0.337	0.216	0.113
LSTM HNN	Capítulo	0.836	0.609	0.654	0.045
Total Params:	Subcapítulo	0.051	0.938	0.090	0.421
2,162,158	Item	0.654	0.439	0.525	0.164
	Geral	0.095	0.661	0.369	0.184
BERT-HAN	Capítulo	0.879	0.863	0.868	0.022
Total Params:	Subcapítulo	0.591	0.701	0.666	0.019
28,964,207	Item	0.672	0.635	0.580	0.014
	Geral	0.692	0.719	0.687	0.017
GPT-4o-mini	Capítulo	0.871	0.839	0.849	0.025
Total Params:	Subcapítulo	0.736	0.556	0.621	0.016
not revealed	Item	0.796	0.570	0.603	0.012
	Geral	0.800	0.636	0.675	0.015
LLaMA Fine-tuned	Capítulo	0.931	0.851	0.886	0.019
Total Params:	Subcapítulo	0.852	0.453	0.551	0.017
8,072,204,288	Item	0.822	0.599	0.645	0.012
Trainable Params: 41,943,040	Geral	0.867	0.611	0.673	0.014

Por fim, o modelo LLaMA *fine-tuned* apresentou o melhor desempenho global ao nível do capítulo e resultados sólidos nos restantes níveis, evidenciando forte capacidade discriminativa. No entanto, a sua natureza generativa limita o controlo direto sobre o compromisso precisão–*recall*, uma vez que não permite ajuste explícito de *thresholds*, tornando-o mais conservador em níveis hierárquicos intermédios.

Globalmente, os resultados demonstram que modelos *transformer*, sobretudo quando refinados ou integrados em arquiteturas hierárquicas, superam claramente abordagens clássicas na

classificação hierárquica multi-rótulo de MQTs, sendo mais adequados para lidar com desequilíbrio, ambiguidade semântica e dependências estruturais complexas.

6. Conclusão e estudos futuros

Este estudo abordou o problema da classificação automática de descrições de MQTs numa estrutura hierárquica composta por capítulos, subcapítulos e itens, um desafio central nos processos de orçamentação na indústria da construção. A motivação principal prendeu-se com a necessidade crescente de automatizar tarefas repetitivas e propensas a erro humano, num contexto marcado por prazos apertados, elevada complexidade técnica e grande volume de dados não estruturados.

Os resultados obtidos demonstram de forma consistente que modelos baseados em *transformer*s superam claramente abordagens tradicionais ao longo de todos os níveis hierárquicos. Em particular, o BERT-HAN revelou-se a solução supervisionada mais equilibrada e fiável. Em paralelo, os modelos baseados em LLMs, nomeadamente o LLaMA refinado e o GPT-4o-mini em regime *few-shot*, evidenciaram um desempenho competitivo, o último na ausência de treino supervisionado, reforçando o potencial destas abordagens para cenários de rápida implementação ou de escassez de dados anotados.

O estudo evidenciou também limitações estruturais importantes, sobretudo ao nível dos subcapítulos e itens, onde a ambiguidade semântica das descrições, a sobreposição entre classes e o desequilíbrio na distribuição dos rótulos impactam negativamente métricas como o *recall*. Estas dificuldades não refletem apenas limitações dos modelos, mas também inconsistências inerentes ao próprio processo de anotação e à natureza dos documentos de contratação pública.

Do ponto de vista prático, os resultados confirmam a viabilidade da integração de sistemas de classificação hierárquica baseados em IA em fluxos reais de contratação, permitindo acelerar a categorização de tarefas, melhorar a consistência dos orçamentos e facilitar a interoperabilidade entre ferramentas de planeamento, estimativa e gestão de projetos.

Como trabalho futuro, destaca-se a necessidade de expandir e refinar o conjunto de dados, privilegiando fontes internas de empresas com terminologia mais padronizada. A integração de abordagens de *Retrieval-Augmented Generation* (RAG) surge como uma linha particularmente promissora, permitindo enriquecer o contexto dos modelos com conhecimento externo relevante durante a inferência. Adicionalmente, a incorporação de mecanismos de explicabilidade e validação *human-in-the-loop* será fundamental para garantir confiança, transparência e adoção em ambientes industriais e de contratação pública.

Em síntese, este trabalho estabelece uma base sólida para a aplicação prática de modelos avançados de linguagem na classificação hierárquica de MQTs, demonstrando ganhos claros face a métodos tradicionais e abrindo múltiplas oportunidades de evolução técnica e operacional.

Agradecimentos

Este trabalho tem o cofinanciamento do PRR – Plano de Recuperação e Resiliência e União Europeia – www.recuperarportugal.gov.pt | PRR – Investimento RE-C05-i02: Missão Interface – CoLAB."

Referências

[1] L. Jacques de Sousa, J. Martins, L. Sanhudo, and J. Baptista, "Automation of text document classification in the budgeting phase of the Construction process: a Systematic Literature Review," Construction Innovation, vol. 24, 2024, doi: 10.1108/CI-12-2022-0315.

[2] S. Durdyev, "Review of construction journals on causes of project cost overruns," Engineering, Construction and Architectural Management, vol. 28, no. 4, pp. 1241–1260, 2021, doi: 10.1108/ECAM-02-2020-0137.

[3] A. M. Abdelalim *et al.*, "An Analysis of Factors Contributing to Cost Overruns in the Global Construction Industry," Buildings, vol. 15, no. 1, doi: 10.3390/buildings15010018.

[4] S. Moon, S. Chi, and S.-B. Im, "Automated detection of contractual risk clauses from construction specifications using bidirectional encoder representations from transformers (BERT)," Automation in Construction, vol. 142, p. 104465, 2022, doi: 10.1016/j.autcon.2022.104465.

[5] J. Devlin *et al.*, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in North American Chapter of the Association for Computational Linguistics, 2019.

[6] D. Mahmud *et al.*, "Integrating LLMs With ITS: Recent Advances, Potentials, Challenges, and Future Directions," IEEE Transactions on Intelligent Transportation Systems, 2025.

[7] J. I. Deza, H. Ihshaish, and L. Mahdjoubi, "A Machine Learning Approach to Classifying Construction Cost Documents into the International Construction Measurement Standard," ArXiv, vol. abs/2211.07705, 2022.

[8] L. Siciliani *et al.*, "AI-based decision support system for public procurement," Information Systems, vol. 119, p. 102284, 2023, doi: 10.1016/j.is.2023.102284.

[9] M. Salem, A. Mohamed, and K. Shaalan, "Transformer Models in Natural Language Processing: A Comprehensive Review and Prospects for Future Development," in 11th International Conference on Advanced Intelligent Systems and Informatics (AISI 2025), Cham, A. E. Hassanien *et al.*, Eds., 2025: Springer Nature Switzerland, pp. 463–472.

[10] Z. Zheng *et al.*, "A Text Classification-Based Approach for Evaluating and Enhancing the Machine Interpretability of Building Codes," ArXiv, vol. abs/2309.14374, 2023.

[11] O. J. Achiam *et al.*, "GPT-4 Technical Report," 2023.

[12] A. Dubey *et al.*, "The Llama 3 Herd of Models," ArXiv, vol. abs/2407.21783, 2024.

[13] J. A. F. Costa, N. C. D. Dantas, and E. D. S. A. Silva, "Evaluating Text Classification in the Legal Domain Using BERT Embeddings," in Intelligent Data Engineering and Automated Learning (IDEAL 2023), Cham, P. Quaresma *et al.*, Eds., 2023: Springer Nature Switzerland, pp. 51–63.

[14] F. Zeidi, M. F. Amasyalı, and c. Erol, "LegalTurk Optimized BERT for Multi-Label Text Classification and NER," ArXiv, vol. abs/2407.00648, 2024.

[15] Z. Hou *et al.*, "Text Error Correction Method in the Construction Industry Based on Transfer Learning," in Communications and Networking, Cham, H. Gao *et al.*, Eds., 2022: Springer International Publishing, pp. 277–290.

[16] P. Ghimire, K. Kim, and M. Acharya, "Generative AI in the Construction Industry: Opportunities & Challenges," ArXiv, vol. abs/2310.04427, 2023.

[17] L. Jacques de Sousa, J. Poças Martins, L. Sanhudo, and J. M. Silva, "Portuguese Construction Dataset for AI: BOQ Text Extraction and Synthethic Data Augmentation Using LLMs," presented at the 2025 European Conference on Computing in Construction - CIB W78 Conference on IT in Construction, Porto, Portugal, 2025.