Field of Science and Technology Classification of Academic Documents in Portuguese

Ivo Simões and Hugo Gonçalo Oliveira and João Correia
CISUC, University of Coimbra
ivosimoes@student.dei.uc.pt, hroliv@dei.uc.pt, jncor@dei.uc.pt

Abstract

Towards improving metadata in academic repositories, this study evaluates the efficacy of different transformer-based models in the automatic classification of the Field of Science and Technology (FOS) of academic theses written in Portuguese. We compare the performance of four different encoder models, two multilingual and two Portuguese-specific, against five larger decoder-based LLMs, on a dataset of 9,696 theses characterized by their title, keywords, and abstract. Fine-tuned encoder-based models achieved the best scores ($F_1 = 88\%$), outperforming general-purpose decoder models prompted for the task. These results suggest that, for localized academic domains, task-specific fine-tuning remains more effective than general-purpose LLM prompting.

1 Introduction

Digital scientific repositories are essential pillars of modern research, by facilitating access to scientific production. Between commercially-owned and non-profit repositories, it is common practice for research and higher-education institutions to have their own repositories of scientific publications authored by their members. However, despite the adoption of common standards, such repositories tend to have inconsistent, incorrect, or incomplete metadata, which may result from human error. Documents can be submitted by their authors, who are frequently unaware of the standards, or by staff, who are unfamiliar with the covered topics. Therefore, tools for improving the quality of document metadata are necessary, especially for non-English languages, such as Portuguese. This includes the automatic classification of the scientific area, which will benefit document organization, discovery, and retrieval, among others.

We tackle the task of document classification for theses in Estudo Geral (EG) (Miguéis and Neves, 2021), the institutional repository of the University of Coimbra, in Portugal. When working with EG's metadata, we found that a significant portion of the theses did not have an assigned Field of Science and Technology (FOS), a compulsory classification by the OECD[1] for statistics of branches of scholarly and technical fields, which has six top-level categories for R&D works. For some theses, the FOS could not even be inferred from their collection, which had a generic name (*UC - Dissertações e Teses*) instead of mentioning the unit where the work had been developed.

For classifying the FOS of the theses from EG, considering their title, keywords, abstract, or their combination, we explored a range of transformer-based models. Evaluation relied on 20% of the 9,696 theses which had an assigned FOS, leaving the remaining available for fine-tuning the encoder-based models used in experimentation.

To the best of our knowledge, this is the first study to address FOS classification of academic theses in Portuguese. Our main contribution is a comparative analysis of transformer models in this task, covering encoder-based, which we fine-tune, as well as open and proprietary decoder-based models, prompted in zero-shot mode.

The remainder of the paper is organized as follows: Section 2 reviews work on the automatic classification of scientific publications; Section 3 describes our experimental setup, including data and models used; and, before concluding, Section 4 presents and discusses the obtained results.

2 Related Work

Natural Language Processing for Science has been attracting increasing interest, with large domain-specific corpora (Saier and Färber, 2019; Lo et al., 2020) and language models (Beltagy et al., 2019; Taylor et al., 2022; Asai et al., 2024) available.

[1] https://web-archive.oecd.org/2012-06-15/138575-38235147.pdf

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 1, pages 1021–1026

On the classification of scientific documents, the best performing methods are generally encoder-based transformers fine-tuned for the task. We highlight SciBERT (Beltagy et al., 2019), a BERT model pretrained on a corpus of 1.1M publications within computer science and biomedicine. After fine-tuning, SciBERT consistently outperforms: $\text{BERT}_{\text{BASE}}$ (Devlin et al., 2019) in a range of tasks in the scientific domain, including the classification of the field of study according to the Microsoft Academic Graph (MAG) (Sinha et al., 2015); or other BERT models pretrained on the biomedical domain (Lee et al., 2020; Peng et al., 2019) in the classification into 3 and 7 categories, using either abstracts or keywords (Rostam and Kertész, 2024), with top F1 scores (87%, 97%) achieved with the abstracts. Besides abstracts and keywords, the title, references, citations, recommendations, and argumentative zones were used for classification in one of 36 categories by an ensemble of transformers (Mendoza et al., 2022). Citations and references were also used for classification in predefined FOS taxonomies, based on a multilayer graph network (Gialitsis et al., 2022).

Some authors target broader multi-level categories (Toney and Dunham, 2022; Rao et al., 2025; Gusenbauer et al., 2025), e.g., based on MAG or topics (Sadat and Caragea, 2022), and tackle the classification of scientific documents as a multi-label problem. A simple approach uses word embeddings and cosine similarity (Toney and Dunham, 2022), while others explore deep neural networks (Rao et al., 2025), including BERT (Sadat and Caragea, 2022; Gusenbauer et al., 2025).

Recent work (Arhiliuc et al., 2025) compares different methods in the classification of journal abstracts from 42 scientific disciplines (i.e., second-level FOS). It considered traditional SVMs, fine-tuned BERT models, and in-context learning with GPT-3.5-turbo. SciBERT achieved the best macro F1 (81%) and GPT-3.5 was the worst performer.

With few exceptions (Toney and Dunham, 2022), prior work is on English-written documents. This is natural, because English is by far the most frequent language in science, but we can easily find many scientific publications in other languages, including Portuguese, for which related work is limited.

3 Experimental Setup

In this section, we describe the dataset used for experimentation, followed by the tested models.

FOS	# of samples	
Natural Sciences	832	(8.6%)
Engineering and Technology	1,620	(16.7%)
Medical and Health Sciences	3,298	(34.0%)
Social Sciences	3,129	(32.3%)
Humanities	817	(8.4%)

Table 1: Number of documents in the final dataset for each Field of Science and Technology (FOS) label.

3.1 Dataset

We used theses from the institutional repository of the University of Coimbra (EG), available in open access, not embargoed, and written in Portuguese. For classification, we focus on the following metadata fields: title, keywords, abstract, and the FOS (category). Based on their values, we further excluded theses with any empty fields or more than one FOS, as well as those where the title was, in fact, a section header, after inspecting the most frequent values (e.g. *Introdução*, *Editorial*, *Apresentação*, *Prefácio*, *Relatório de Estágio*). The result was a dataset of 9,696 theses — bachelor (4.1%), master (80.2%), and doctoral (15.7%) — with their FOS distributed according to Table 1.

The FOS has six top-level categories: Natural Sciences, Engineering and Technology, Medical and Health Sciences, Agricultural Sciences, Social Sciences, and Humanities. However, since the university of Coimbra does not have a unit focused on Agricultural Sciences, our dataset is limited to the other five categories.

For the reported experimentation, we used a stratified train-test split of the data: 80% used for training and 20% for evaluation.

3.2 Models

We used the training split to fine-tune four different encoder-based models for FOS classification. These included two multilingual models, multilingual cased $\text{BERT}_{\text{BASE}}$ (Devlin et al., 2019) and $\text{mE5}_{\text{small}}$ (Wang et al., 2024), and two models pretrained for Portuguese, $\text{BERTimbau}_{\text{BASE}}$ (Souza et al., 2020) and Albertina-100M-PTPT (Santos et al., 2024a). Fine-tuning was performed for 10 epochs, using the AdamW optimizer with a learning rate of 3×10^{-5} and the remaining default parameters of the HuggingFace Trainer[2]. We conduct four independent fine-tuning runs for each combination of model and fields, and report the mean of the best macro-F1 scores obtained at each of

[2] https://huggingface.co/docs/transformers/en/main_classes/trainer

the four runs. Models were fine-tuned in a system equipped with an NVIDIA RTX A6000 GPU (48 GB VRAM).

We additionally test a set of significantly larger multilingual decoder-based models, known for performing tasks from natural language descriptions, without fine-tuning. The set of models of this kind included the proprietary GPT-5 Nano and the open model gpt-oss-20b (OpenAI et al., 2025), both by OpenAI, two open models by Meta, Llama3.1-8B Instruct and the larger Llama3.3-70B Instruct (Grattafiori et al., 2024), as well as Qwen3-8B (Yang et al., 2025). GPT-5 was accessed through OpenAI's API, Llama3.1 models were tested using the OpenRouter API[3], and Qwen3-8B and gpt-oss-20b were locally run using the same machine used for fine-tuning the encoder models.

The same zero-shot prompt was used for the decoder models and combined with structured outputs for extracting the predicted FOS. It included a system prompt with the high-level instructions, followed by a user prompt with the input fields, one per line, starting with their identification: `TITLE`, `KEYWORDS`, `ABSTRACT`. Figure 1 illustrates the full prompt, with the three input fields, as instantiated for a specific thesis.

```
Vais receber texto proveniente de diferentes
secções de uma dissertação ou de um relatório
(TÍTULO, PALAVRAS-CHAVE, RESUMO).
A tua tarefa consiste em classificar a área
científica (FOS - Field of Science and
Technology) do documento. Deves escolher a FOS
que melhor se adequa ao documento, entre as
seguintes: Ciências Sociais, Ciências Médicas
e da Saúde, Humanidades, Ciências da
Engenharia e Tecnologias, Ciências Exactas e
Naturais.
TÍTULO: Envelhecimento nas dificuldades
intelectuais
PALAVRAS-CHAVE: Deficiente mental, idoso
RESUMO: O presente estudo teve como principal
objetivo a caraterização das pessoas idosas
com dificuldades intelectuais, um grupo
recente da população. Procurámos conhecer as
suas principais competências ao nível do
comportamento adaptativo e das funções e
estruturas do corpo e as necessidades
percebidas pelos profissionais que com elas
trabalham. Aplicámos a Escala de Comportamento
Adaptativo, alguns itens da Checklist da CIF e
construímos uma Listagem de necessidades. Os
resultados obtidos apresentam as diferenças
encontradas em função do género, grupo etário
e grau de dificuldades intelectuais.
```

Figure 1: Example prompt for FOS classification.

[3] https://openrouter.ai/.

The temperature of the open models was set to 0.01. In GPT-5 Nano, however, this hyperparameter is unknown and, as a proprietary model, cannot be changed. For gpt-oss-20b, the reasoning level was set to 'high', which, according to its authors, leads to better results. However, we noted that, in some cases, the model got stuck in long reasoning loops. To circumvent this, we restrain the number of output tokens to 3,000 and add a condition such that, after four consecutive classification attempts exceeding this limit for the same prompt, the temperature is set to 1. This was triggered for less than 0.2% of the inputs.

4 Results

Table 2 reports the performance of each model with different combinations of fields. Overall, we observe that fine-tuned BERT models perform substantially better than zero-shot prompting general-purpose decoder models. With a macro F1 of 88%, the best BERT model, BERTimbau, outperforms the best decoder model by ≈17 points. We thus conclude that fine-tuning smaller encoder models is still preferable to prompting larger decoder models not trained in the target task. Nevertheless, considering that they were not trained for the task, the latter models still exhibit a positive performance. They should thus be seen as an option when data is not available for fine-tuning.

Despite achieving the best macro F1 in three combinations of fields, the performance of BERTimbau is not substantially better than that of the other BERT models, suggesting that fine-tuning was effective, even for the multilingual models. The best combination is to use all the available fields: title, keywords, and abstract (T+K+A), even if the contribution of the keywords is unclear. Nevertheless, with the fine-tuned models, F1 is higher than 80% for every combination except for title-only, where scores were 1 or 2 points away from this milestone, despite this field having, on average, only 25 tokens.

The best decoder model was Llama3.3 70B, with F1 scores around 70%. GPT-5 Nano, according to OpenAI, the "fastest, cheapest version of GPT-5" and "great for summarization and classification tasks", lies 1 point behind, with the open gpt-oss-20 performing very close to GPT-5. The worst performing model was clearly Llama3.1 8B, which was somewhat expected due to scaling laws, closely followed by Qwen3 8B, which managed to outper-

Model	T			T+K			T+A			T+K+A		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
$\text{mBERT}_{\text{BASE}}$	0.787	0.777	0.782	0.863	0.838	0.849	0.871	0.854	0.861	0.879	0.863	0.870
$\text{mE5}_{\text{small}}$	0.795	0.784	0.790	0.845	0.831	0.837	0.868	0.863	0.865	0.870	0.860	0.864
$\text{BERTimbau}_{\text{BASE}}$	0.798	0.785	**0.791**	0.874	0.843	**0.855**	0.882	0.875	**0.878**	0.887	0.877	**0.881**
Albertina-100M	0.790	0.776	0.781	0.862	0.842	0.851	0.865	0.853	0.858	0.864	0.845	0.853
GPT-5 Nano	0.714	0.677	0.686	0.732	0.683	0.695	0.742	0.684	0.699	0.747	0.682	0.699
gpt-oss-20b	0.700	0.667	0.677	0.715	0.667	0.683	0.730	0.674	0.693	0.729	0.679	0.696
Llama3.1 8B	0.640	0.626	0.629	0.658	0.631	0.640	0.641	0.635	0.636	0.648	0.639	0.640
Llama3.3 70B	0.704	0.698	**0.698**	0.713	0.701	**0.704**	0.726	0.697	**0.709**	0.721	0.701	**0.709**
Qwen3 8B	0.715	0.634	0.659	0.725	0.646	0.672	0.729	0.624	0.651	0.728	0.635	0.662

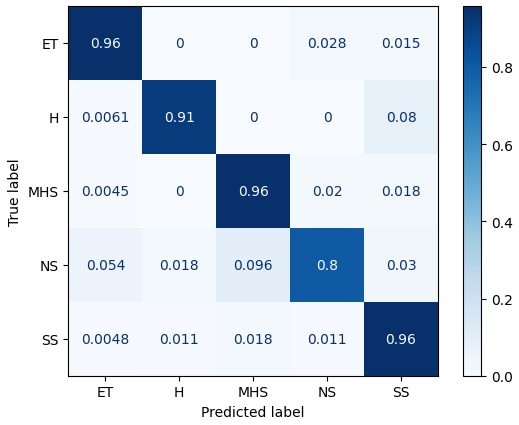
Table 2: Macro-averaged Precision (P), Recall (R), and F1 scores of each model for four different subsets of text fields (T = Title, T+K = Title and Keywords, T+A = Title and Abstract, T+K+A = Title, Keywords, and Abstract).

form the former across every field combination, despite having a similar number of parameters. This performance gap can be attributed to Qwen3's more robust pre-training on multilingual data, including Portuguese, whereas Llama3.1 employs a more English-focused training corpus. Llama3.1 8B and Qwen3 8B were the smallest decoder models tested, much smaller than Llama3.3 70B. Nevertheless, both still have ≈80 times more parameters than $\text{BERTimbau}_{\text{BASE}}$, further stressing that, whenever possible, one should opt for fine-tuning smaller models.

Since the decoder models were not trained on the target task, we would expect the number of provided fields to have a greater impact on their performance. However, the performance of every combination is even closer than with the BERT models.

For a more detailed analysis, Figure 2 depicts the confusion matrices of the best BERT and decoder-based models. It further reveals that Natural Sciences is the most difficult class, with F1 scores of 11 and 27 points below the second most difficult, respectively for BERTimbau and for Llama3.3. For both models, Engineering and Technology is the top-performing class, together with Medical and Health and Social Sciences for BERTimbau.

(a) $\text{BERTimbau}_{\text{BASE}}$ (T+K+A).



(b) Llama3.3-70B (T+K+A)

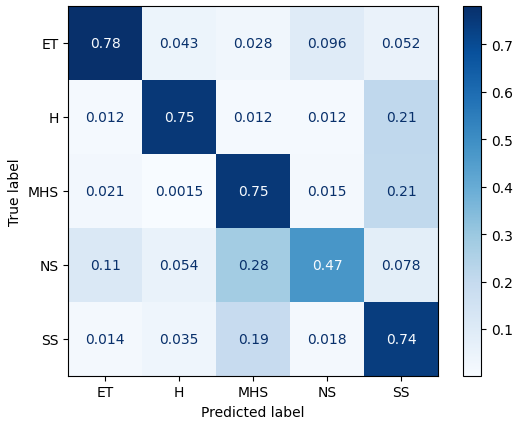


Figure 2: Confusion matrices of selected models (ET=Engineering and Technology, H=Humanities, MHS=Medical and Health Sciences, NS=Natural Sciences, SS=Social Sciences.)

5 Conclusion

We presented a comparative study of different models for FOS classification of academic thesis in Portuguese, covering fine-tuned BERT-based models and larger decoder-models in zero-shot prompting. We conclude that fine-tuning smaller models is the best option, validating the conclusions of previous work in a similar task, for English (Arhiliuc et al., 2025), and in line with other studies on text classification in Portuguese (da Silva Oliveira et al., 2024; Pinto et al., 2025). Performance was surprisingly positive when considering just the title, but the best option is to consider a combination of title, abstract and keywords for classification. Moreover, the models consistently struggled in the classification of publications of Natural Sciences.

We stress that data availability remains a key determinant of the performance of smaller models, and, in cases where training data is scarce, prompting decoder models is a viable option. Therefore, the data used in our work can be found and ac-

cessed through a public collection on HuggingFace[4], together with a version of the best fine-tuned models for every field combination. These will establish baselines for the untouched problem of FOS classification of Portuguese academic work.

Since it is with the decoder models that there is more room for improvement, future work should include few-shot prompts and test other models of this kind, especially those pretrained for Portuguese, e.g., Gervásio (Santos et al., 2024b), Sabiá (Abonizio et al., 2024), or AMALIA[5], which can be further fine-tuned for the task. The exploration of other fields should also be considered, as well as multi-label and / or lower-level classifications.

Acknowledgements

This work was partially supported by the AMALIA project, funded by FCT/IP in the context of measure RE-C05-i08 of the Portuguese Recovery and Resilience Program; by the Portuguese Recovery and Resilience Plan (PRR) through project C645008882-00000055, Center for Responsible AI; and by national funds through FCT – Foundation for Science and Technology I.P., in the framework of the Project CISUC (UIDB/00326/2025 and UIDP/00326/2025).

References

Hugo Abonizio, Thales Sales Almeida, Thiago Laitz, Roseval Malaquias Junior, Giovana Kerche Bonás, Rodrigo Nogueira, and Ramon Pires. 2024. Sabi\'a-3 technical report. *arXiv preprint arXiv:2410.12049*.

Cristina Arhiliuc, Raf Guns, Walter Daelemans, and Tim CE Engels. 2025. Journal article classification using abstracts: a comparison of classical and transformer-based machine learning methods. *Scientometrics*, 130(1):313–342.

Akari Asai, Jacqueline He, Rulin Shao, Weijia Shi, Amanpreet Singh, Joseph Chee Chang, Kyle Lo, Luca Soldaini, Sergey Feldman, Mike D'arcy, and 1 others. 2024. OpenScholar: Synthesizing scientific literature with Retrieval-Augmented LLMs. *arXiv preprint arXiv:2411.14199*.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Amanda da Silva Oliveira, Thiago de Carvalho Cecote, João Paulo Reis Alvarenga, Vander Luis de Souza Freitas, and Eduardo José da Silva Luz. 2024. Toxic speech detection in Portuguese: A comparative study of large language models. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese-Vol. 1*, pages 108–116.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*, pages 4171–4186, Minneapolis, Minnesota. ACL.

Nikolaos Gialitsis, Sotiris Kotitsas, and Haris Papageorgiou. 2022. Scinobo: A hierarchical multi-label classifier of scientific publications. In *Companion Proceedings of the Web Conference 2022*, WWW '22, page 800–809, New York, NY, USA. Association for Computing Machinery.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Michael Gusenbauer, Jochen Endermann, Harald Huber, Simon Strasser, Andreas-Nizar Granitzer, and Thomas Ströhle. 2025. Fine-tuning scibert to enable asjc-based assessments of the disciplinary orientation of research collections. *Scientometrics*, pages 1–38.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. 2020. S2ORC: The semantic scholar open research corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4969–4983, Online. ACL.

Óscar E. Mendoza, Wojciech Kusa, Alaa El-Ebshihy, Ronin Wu, David Pride, Petr Knoth, Drahomira Herrmannova, Florina Piroi, Gabriella Pasi, and Allan Hanbury. 2022. Benchmark for research theme classification of scholarly documents. In *Proceedings of the Third Workshop on Scholarly Document Processing*, pages 253–262, Gyeongju, Republic of Korea. ACL.

[4] https://huggingface.co/collections/ivosimoes/propor-fos-classification

[5] https://amaliallm.pt/

Ana Eva Miguéis and Bruno Neves. 2021. A visão dos gestores de repositórios. o caso da universidade de coimbra. *Sob a lente da Ciência Aberta: Olhares de Portugal, Espanha e Brasil*, pages 273–294.

OpenAI, :, Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, Boaz Barak, Ally Bennett, Tyler Bertao, Nivedita Brett, Eugene Brevdo, Greg Brockman, Sebastien Bubeck, and 108 others. 2025. gpt-oss-120b & gpt-oss-20b model card. *Preprint*, arXiv:2508.10925.

Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. Transfer learning in biomedical natural language processing: An evaluation of BERT and ELMo on ten benchmarking datasets. *BioNLP 2019*, page 58.

Tomás Pinto, Bruno Ferreira, Catarina Silva, and Hugo Gonçalo Oliveira. 2025. Prompting LLMs for relation classification in Portuguese: Is it worth it? In *Progress in Artificial Intelligence*, pages 249–261, Cham. Springer Nature Switzerland.

Susie Xi Rao, Peter H Egger, and Ce Zhang. 2025. Hierarchical classification of research fields in the mag science network using deep learning. *Quantitative Science Studies*, pages 1–60.

Zhyar Rzgar K Rostam and Gábor Kertész. 2024. Fine-tuning large language models for scientific text classification: A comparative study.

Mobashir Sadat and Cornelia Caragea. 2022. Hierarchical multi-label classification of scientific documents. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8923–8937, Abu Dhabi, United Arab Emirates. ACL.

Tarek Saier and Michael Färber. 2019. Bibliometric-enhanced arxiv: A data set for paper-based and citation-based tasks. In *Proceedings of the 8th International Workshop on Bibliometric-enhanced Information Retrieval (BIR) co-located with the 41st European Conference on Information Retrieval (ECIR 2019)*, pages 14–26. CEUR-WS.org.

Rodrigo Santos, João Rodrigues, Luís Gomes, João Silva, António Branco, Henrique Lopes Cardoso, Tomás Freitas Osório, and Bernardo Leite. 2024a. Fostering the ecosystem of open neural encoders for portuguese with Albertina PT-* family. In *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages@ LREC-COLING 2024*, pages 105–114.

Rodrigo Santos, João Ricardo Silva, Luís Gomes, João Rodrigues, and António Branco. 2024b. Advancing generative AI for Portuguese with open decoder Gervásio PT*. In *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @ LREC-COLING 2024*, pages 16–26, Torino, Italia. ELRA and ICCL.

Arnab Sinha, Zhihong Shen, Yang Song, Hao Ma, Darrin Eide, Bo-June Hsu, and Kuansan Wang. 2015. An overview of microsoft academic service (mas) and applications. In *Proceedings of the 24th international conference on world wide web*, pages 243–246.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. BERTimbau: Pretrained BERT models for Brazilian Portuguese. In *Brazilian Conference on Intelligent Systems (BRACIS)*, pages 403–417.

Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. 2022. Galactica: A large language model for science. *arXiv preprint arXiv:2211.09085*.

Autumn Toney and James Dunham. 2022. Multi-label classification of scientific research documents across domains and languages. In *Proceedings of the third workshop on scholarly document processing*, pages 105–114.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual e5 text embeddings: A technical report. *Preprint*, arXiv:2402.05672.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.