Ma 412 - Mathematical Foundations for Statistical Learning

-

Multi-Label Classification of Scientific Literature Using the NASA SciX Corpus

Atilla Kaan Alkan
atilla1.alkan@ipsa.fr

May 2025

1 Introduction

The NASA Astrophysics Data System (ADS) (Kurtz et al., 2000) is a research database supporting the astronomy and astrophysics community, covering over 20 million records spanning astronomy, astrophysics, and general physics publications. ADS has significantly increased research efficiency for astronomers over the last decades (Kurtz et al., 2005). Today, ADS is broadening to new domains such as heliophysics, earth sciences, and biophysics, evolving into a multidisciplinary platform that benefits a broader scientific audience. The Science Explorer (SciX) portal[1] was developed to support this expansion, enabling researchers across diverse disciplines—such as earth science, physics, heliophysics, and planetary science—to access a rich, interconnected library. SciX integrates ADS's original resources with new capabilities. As SciX grows, a crucial enhancement involves implementing *automated keyword labelling* through multi-label text classification with machine learning techniques. Multi-label classification enables the prediction of multiple relevant keywords for each document, allowing for accurate and efficient categorization across scientific domains (Sadat and Caragea, 2022). This task would significantly reduce the time and resources required for manual indexing and improve the user experience on SciX, where researchers rely on precise and relevant search results to navigate the expanding interdisciplinary content (Toney and Dunham, 2022).

[1] https://scixplorer.org/

2 Task Description

This project aims to develop a system that predicts relevant keywords (e.g., "solar wind", "lunar composition", etc.) by analyzing scientific papers' titles and abstracts. You will train a machine learning model to identify associations between document content and specific keywords, enabling it to recognize and label topics based on linguistic patterns. This task involves multi-label text classification, meaning each document may have multiple labels to capture the diverse themes or topics within scientific texts, unlike single-label classification, where only one label is assigned. The annotated dataset for this task is provided by the NASA ADS/SciX team[2] and is available through the HuggingFace repository[3]. The SciX corpus, comprising titles and abstracts of published papers, is split into training and test sets with 18,677 and 3,025 documents, respectively[4],[5].

3 Instructions

You must provide a system description report (.pdf format) presenting the work. This involves explaining the problem, the possible solutions, and those you have chosen by presenting the strengths and weaknesses. Particular attention should be paid to the explanation of the considered algorithms. The report should consist of the following sections:

Data Exploration and Preprocessing Please familiarize yourself with the dataset, i.e., examine the SciX dataset and understand the class distribution. Perform necessary preprocessing steps to prepare the texts for training a machine learning model.

Model Selection and Experimental Settings Identify an appropriate architecture for this multi-label classification task. It would help if you justified your choice by providing reasoning for your choice of architecture (you can read papers in section 4).

Evaluation and Metrics Check the common metrics used for this task by reading the papers of section 4 and test your model's performance on the validation set[6] using these metrics. Compare the results of your different models and analyze areas for improvement.

[2] https://ui.adsabs.harvard.edu/about/team/

[3] https://huggingface.co/datasets/adsabs/SciX_UAT_keywords

[4] https://huggingface.co/datasets/adsabs/SciX_UAT_keywords/resolve/main/data/train-00000-of-00001-b21313e511aa601a.parquet

[5] https://huggingface.co/datasets/adsabs/SciX_UAT_keywords/resolve/main/data/val-00000-of-00001-66ce8665444026dc.parquet

[6] https://huggingface.co/datasets/adsabs/SciX_UAT_keywords/resolve/main/data/val-00000-of-00001-66ce8665444026dc.parquet

Results Analysis Interpret the results, identifying strengths and potential weaknesses in the model's prediction of different keywords. You should perform an error analysis by examining cases where the model makes incorrect predictions. This will help you understand if specific keywords are more challenging to classify.

3.1 GitHub Repository

Please provide a GitHub repository containing your code for this project. The repository should be organised and include a README.md explaining the project, setup instructions, and how to run the code. Ensure the code is executable, the documentation is clear, and experiments are reproducible (use fixed random seeds to save any essential model checkpoints or outputs).

3.2 Timeline

The project will end with the system report submission on Friday, June 20th (11:59 pm). Make sure to finalise and upload your GitHub repository by this date, ensuring all code is executable, well-documented, and reproducible.

4 External Resources

- An introduction to multi-label text classification: https://medium.com/analytics-vidhya/an-introduction-to-multi-label-text-classification-b1bcb7c7364c;
- Hierarchical Multi-Label Classification of Scientific Documents, by Sadat and Caragea (2022);
- Extreme Multi-Label Legal Text Classification: A Case Study in EU Legislation, by Chalkidis et al. (2019);
- Evaluating Extreme Hierarchical Multi-label Classification, by Amigo and Delgado (2022).
- A Survey on Recent Advances in Hierarchical Multi-label Text Classification, by Liu et al. (2023)

References

Amigo, E. and Delgado, A. (2022). Evaluating extreme hierarchical multi-label classification. In Muresan, S., Nakov, P., and Villavicencio, A., editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5809–5819, Dublin, Ireland. Association for Computational Linguistics.

Chalkidis, I., Fergadiotis, E., Malakasiotis, P., Aletras, N., and Androutsopoulos, I. (2019). Extreme multi-label legal text classification: A case study in EU legislation. In Aletras, N., Ash, E., Barrett, L., Chen, D., Meyers, A., Preotiuc-Pietro, D., Rosenberg, D., and Stent, A., editors, *Proceedings of the Natural Legal Language Processing Workshop 2019*, pages 78–87, Minneapolis, Minnesota. Association for Computational Linguistics.

Kurtz, M. J., Eichhorn, G., Accomazzi, A., Grant, C. S., Demleitner, M., and Murray, S. S. (2005). Worldwide Use and Impact of the NASA Astrophysics Data System Digital Library. *Journal of the American Society for Information Science and Technology*, 56:36.

Kurtz, M. J., Eichhorn, G., Accomazzi, A., Grant, C. S., Murray, S. S., and Watson, J. M. (2000). The NASA Astrophysics Data System: Overview. , 143:41–59.

Liu, R., Liang, W., Luo, W., Song, Y., Zhang, H., Xu, R., Li, Y., and Liu, M. (2023). Recent advances in hierarchical multi-label text classification: A survey.

Sadat, M. and Caragea, C. (2022). Hierarchical multi-label classification of scientific documents. In Goldberg, Y., Kozareva, Z., and Zhang, Y., editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8923–8937, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Toney, A. and Dunham, J. (2022). Multi-label classification of scientific research documents across domains and languages. In Cohan, A., Feigenblat, G., Freitag, D., Ghosal, T., Herrmannova, D., Knoth, P., Lo, K., Mayr, P., Shmueli-Scheuer, M., de Waard, A., and Wang, L. L., editors, *Proceedings of the Third Workshop on Scholarly Document Processing*, pages 105–114, Gyeongju, Republic of Korea. Association for Computational Linguistics.