Single-Label Multi-Modal Field of Research Classification

Florian Ruosch[1][*] , Rosni Vasu[1][*] , Ruijie Wang[1,2] , Luca Rossetto[1] , and Abraham Bernstein[1]

[1] Department of Informatics, University of Zurich, Switzerland
[2] University Research Priority Program "Dynamics of Healthy Aging", University of Zurich, Switzerland
{ruosch, rosni, ruijie, rossetto, bernstein}@ifi.uzh.ch

Abstract. The automated field of research classification for scientific papers is still challenging, even with modern tools such as large language models. As part of a shared task tackling this problem, this paper presents our contribution **SLAMFORC**, an approach to single-label classification using multi-modal data. We combined the metadata of papers with their full text and, where available, images into a pipeline to predict their field of research with an ensemble voting on traditional classifiers and large language models. We evaluated our approach on the shared task dataset and scored the highest values for two of the four metrics used in the evaluation of the competition, with the other two being the second highest.

Keywords: Natural Scientific Language Processing · Field of Research Classification · Multi-Modality.

1 Introduction

Keywords and other classifications may help when searching or organizing scholarly publications [20]. They can be annotated by the authors or the publishers, with a corresponding manual effort, or may be machine-generated. The latter has been an application of natural language processing which, with the advent of pre-trained large language models such as BERT [14], has recently gained momentum. Still, the automated classification of research papers remains challenging [27].

This paper describes our submission to the shared task *Field of Research Classification of Scholarly Publications*[1] of the *1st Workshop on Natural Scientific Language Processing and Research Knowledge Graphs (NSLP 2024)*. Its *Subtask I*, which our contribution addresses, is to develop a single-label classifier for general scholarly publications. We trained and tested it on a dataset of

---

[*] Equal contribution
[1] https://nfdi4ds.github.io/nslp2024/docs/forc_shared_task

around $60,000$ English scientific papers [1,2], each from one of 123 hierarchical classes of a subset of the Open Research Knowledge Graph taxonomy.[2] Our approach, dubbed **SLAMFORC** (short for *Single-Label Multi-modal Field of Research Classification*), is multi-modal in that we incorporated data from three different sources: the dataset provided by the organizers of the challenge containing metadata of the articles (e.g., title, abstract), the semantic information provided by Crossref[3], and the contents of the papers (i.e., full text and images). Using this data as features, we engineered a classifier that produces single-label predictions for a given input document. For this endeavor, we computed the embeddings with two different flavors of a pre-trained BERT [14] model and, subsequently, fed these vectors to a handful of traditional classifiers. Then, we applied a voting ensemble [21] to their output to combine them into a final classifier, incorporating all of them as well as the entirety of the available features.

The shared task was very competitive, with 13 system submissions. The margin among the top five submissions was very narrow ($\pm0.75\%$), illustrating that the boundaries were pushed of what was possible with the provided data and task. In the end, our approach came in among the top results, scoring the highest values for two out of four evaluated aspects and the second-best for the others: accuracy (75.6%), precision (75.7%), recall (75.6%), and F1 (75.4%).

The remainder of this paper is structured as follows. Section 2 presents the related work, and Section 3 introduces our methodology. In the ensuing Section 4, we describe our experiments. Finally, we draw conclusions in Section 5.

2   Related Work

The classification of scholarly papers into research fields has found ample applications: for example, to ease organizing or searching the flood of new publications.

One such system [8] groups biomedical papers by applying non-negative matrix factorization [17] to the term relevance vectors of the documents. It uses bisecting k-means clustering [6], and, at the same time, assigns semantic meaning to each document and cluster inferred from the matrix decompositions.

The work by Taheriyan [27] describes an approach to classifying papers by using relationships such as common authors and references as well as citations in a graph. This information allows new papers to be assigned topics automatically instead of requiring manual annotations.

Nguyen and Shirai [20] focus on various text features such as the segmentation of the paper and apply three different classifiers: multi-label kNN [30], binary approach [28], and their newly proposed back-off model. While the latter performs the best, another interesting insight from their results is that only using the title, abstract, and the sections *Introduction* and *Conclusions* of papers improves over using the full text as a feature.

---

[2] https://orkg.org/fields
[3] https://www.crossref.org/
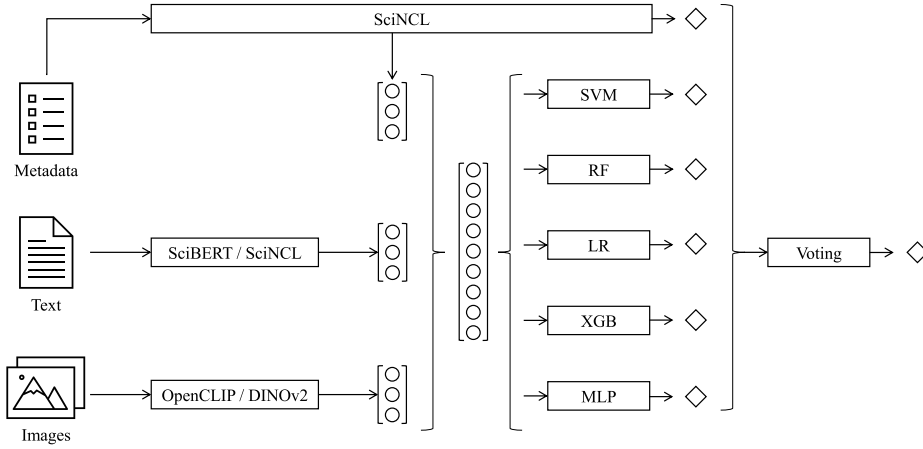
Fig. 1. Overview of the system architecture.

Another approach is presented by Kim and Gil [16]: They describe a classification system based on latent Dirichlet allocation [7] and term frequency-inverse document frequency [25]. The former is employed to extract relevant keywords from the abstracts, the latter for k-means clustering [4] papers with similar topics.

More recently, SPECTER [12] uses pre-trained language models (e.g., SciBERT [5]) to generate document-level embeddings from the titles and abstracts. These can be used for downstream tasks, such as predicting the class of a document, which is demonstrated by applying SPECTER to a new dataset with papers in 19 classes. In this work, incorporating the entire text of papers remains an open issue due to limitations on memory and the availability of the paper contents.

3   The SLAMFORC System

This section describes our approach to solving the shared task. We first explain the multi-modal data of our system. Then, we detail the classifiers we used with this data.

Figure 1 shows an overview of the system. Its code is publicly available.[4]

3.1   Multi-Modal Data

The dataset for the shared task [1,2] consisted of approximately $60,000$ scholarly articles, compiled from various sources such as the *Open Research Knowledge Graph* [3], *arXiv*[5], *Crossref*[6], and the *Semantic Scholar Academic Graph* [29]. It

---

[4] https://gitlab.ifi.uzh.ch/DDIS-Public/forc24
[5] https://arxiv.org
[6] https://www.crossref.org/

spans 123 fields of research (FoR) across five major domains and four hierarchical levels, with mapping to the *ORKG* taxonomy.[7] The challenge of imbalanced data is evident in the dataset, where the distribution of fields is uneven, varying from as low as eight articles (for *Molecular, Cellular, and Tissue Engineering*) to over $6,000$ (for *Physics*).

We utilized Crossref[3] to further enhance the text data of papers. Specifically, for each paper, we used its Digital Object Identifier (DOI) and the Crossref API client[8] to retrieve its annotated subjects and references from the Crossref Unified Resource API.[9] For the paper with the DOI "*10.1007/JHEP06(2012)126*," for example, we retrieved the subject "*Nuclear and High Energy Physics*" and the metadata of 37 reference papers. Despite Crossref adopting a different taxonomy, this retrieved subject remains highly useful for predicting the target label of this paper (i.e., "*Physics*"). Also, the reference papers are mostly in the Physics domain, and this information can be very useful.

We used the title, abstract, and publisher information from the provided dataset, along with the subject data, to generate the metadata embeddings for each paper. We appended all this data as input text to SciNCL [23], a pre-trained BERT model, for computing an embedding as a comprehensive representation of each paper.

In order to make use of the full text for the papers in the dataset, we first had to obtain the respective documents. This was straightforward for items that already had a download link annotated. For all other papers, we used the DOI field, where available, to find the PDFs. There were some cases where neither was available. For those, we relied on Crossref's API to resolve the paper title to its DOI, which allowed us to download the full text document, if it was available.

To extract the text from the PDFs, we employed PaperMage [19]. For each PDF, it produces a JSON file with information about its content and structure. We only relied on the extracted symbols, which we used to reconstruct the full text of the respective papers. Using this data, we computed the document-level embeddings with two pre-trained BERT models: SciBERT [5] and SciNCL [23]. Because of BERT's limitation to processing 512 tokens at a time [14] and papers exceeding this, we batched the input data accordingly. We employed a sliding window of size 512 tokens with an overlap of 128 to conserve semantics near the window borders. After computing the embeddings for each such chunk, we averaged them to obtain the final document-level embedding.

To incorporate the visual information contained in the PDFs, we extracted all their images and converted them to raster graphics. For each image, we used an OpenCLIP [11] model pre-trained on the LAION-5B dataset [26] as well as a pre-trained DINOv2 [22] model to extract image features. When PDFs contained multiple images, we used mean-pooling to aggregate the multiple feature vectors per model, resulting in two vectors per PDF; one for each applied model. For

---

Table 1. The first results on the *validation set* of the individual classifiers with all features (embeddings of metadata, full text, and images) as measured by accuracy, weighted precision, recall, and F1.

Classifier	Accuracy	Precision	Recall	F1
Support Vector Machine	0.755	0.754	0.755	0.752
Random Forest	0.755	0.754	0.755	0.753
Logistic Regression	0.750	0.755	0.750	0.750
XGBoost	0.731	0.735	0.731	0.731
Multilayer Perceptron	0.743	0.748	0.743	0.743

papers where the PDF did not contain any images or the PDF was not available, these vectors were set to zero.

3.2   Classifier

For the final system, we used a mixture of traditional classifiers and neural methods that we combined with an ensemble voting method [21]. Figure 1 shows an overview of the system. After computing the embeddings for the various data sources, we trained several classifiers that could handle vectors as input and predict the single-label class for each item in the dataset.

An obvious choice are Support Vector Machines [13], or SVM for short. Due to the nature of the input data, they can naturally classify them in a high-dimensional space and predict the field of research label. We employed a Random Forest (RF) [15] since they avoid overfitting to the training data, which was an overt problem to be expected because of the skew in classes in the dataset. Logistic Regression (LR) is another widely used traditional classifier to predict single labels on linearly separable data. With eXtreme Gradient Boosting (XGB) [10], we used another popular method that can achieve good performance while sacrificing interpretability. Next, we also employed a fully connected neural net that is a Multilayer Perceptron (MLP), able to deal with not linearly separable data. Furthermore, we also trained SciNCL [23] as an end-to-end solution on the metadata.

Finally, we combined the output of the classifiers described above into an ensemble method [21] with hard voting [18]. This enabled the use of all techniques and all available data at the same time while still producing a single predicted label for each item in the dataset.

4   Experiments

Table 1 shows the results of the initial experiments. We used a set of traditional classifiers as implemented by scikit-learn [24] with all of the available data for each paper consisting of the stacked embedding vectors. Since no method significantly outperformed the others, we combined all of them post-hoc using a

Table 2. Ablation study on the *validation set* by feature combination.

Meta	Text	CLIP	DINO	Accuracy	Precision	Recall	F1-Score
✓				0.774	0.776	0.774	0.773
	✓			0.437	0.709	0.437	0.463
		✓		0.254	0.552	0.254	0.241
			✓	0.259	0.505	0.259	0.250
✓	✓			0.776	0.778	0.776	0.775
✓		✓		0.772	0.775	0.772	0.772
✓			✓	0.776	0.777	0.776	0.774
	✓	✓		0.438	0.700	0.438	0.463
	✓		✓	0.437	0.700	0.437	0.461
		✓	✓	0.271	0.529	0.271	0.265
✓	✓	✓		0.776	0.777	0.776	0.775
✓	✓		✓	0.779	0.781	0.779	0.778
✓		✓	✓	0.776	0.778	0.776	0.774
	✓	✓	✓	0.433	0.701	0.433	0.456
✓	✓	✓	✓	0.777	0.779	0.777	0.775

voting ensemble method, giving us our final classifier for the results of which we submitted to the shared task.

To illustrate the impact of each data source and dissect our multi-modal approach, we performed a feature ablation study, the results of which are shown in Table 2. We used our final system architecture with all classifiers combined with voting on the powerset of possible feature combinations. It is evident that the (embeddings of the) metadata have the most positive influence on the results. Still, adding extra information to the classifier is not detrimental but rather contributes to a higher score. This holds more for the (embeddings of the) full texts of the papers which perform decent on their own. Using the embeddings of the images in the papers alone, where applicable, achieves clearly worse results than the other two data sources. Nevertheless, the combination of all features is among the highest scoring for all four employed metrics, and there was no reason not to rely on everything available.

Finally, Table 3 shows the results of the shared task evaluation.[10] Our submission (ID 683689, top row) scored the highest for precision (75.7%) and F1 (75.4%) while achieving the second-best values for accuracy (75.6%) and recall (75.6%). This goes to show that our multi-modal approach worked and performed well in this competition. Without further knowledge of the other systems, no comparisons can be made or insights gained, and are, thus, left for future work. In conclusion, the automated field of research classification of scientific papers is still challenging, but the submissions for this shared task seemed to have pushed the boundaries of what was possible with the given tools and information, seeing how close the top results were.

---

[10] https://codalab.lisn.upsaclay.fr/competitions/16684#results

Table 3. The final evaluation results on the *test set* as measured by accuracy, weighted precision, recall, and F1 (best for each in **bold**, runner-up <u>underlined</u>). Our submission is the first line.

ID	Accuracy	Precision	Recall	F1 ↓
683689	<u>0.756</u>	**0.757**	<u>0.756</u>	**0.754**
688995	0.754	<u>0.755</u>	0.754	<u>0.752</u>
687946	**0.757**	0.754	**0.757**	0.750
689747	0.748	0.744	0.748	0.743
688510	0.743	0.742	0.743	0.739
651741	0.730	0.725	0.730	0.724
649383	0.726	0.719	0.726	0.720
686435	0.706	0.703	0.706	0.693
686384	0.702	0.695	0.702	0.692
689251	0.682	0.679	0.682	0.678
689454	0.658	0.659	0.658	0.653
647796	0.058	0.061	0.058	0.057
678150	0.004	0.002	0.004	0.002

5   Conclusions

In this paper, we presented **SLAMFORC**, a system for the *Single-Label Multi-modal Field of Research Classification*. We used it to produce the results for our submission to the shared task *Field of Research Classification of Scholarly Publications*. Pursuing a multi-modal approach by incorporating not only the given dataset containing metadata of the papers but also the full text of publications as well as images in these documents, we built an ensemble classifier by combining a set of traditional classifiers using a voting ensemble. We computed the embeddings with pre-trained large language models, stacked these vectors, and trained the individual classifiers. Then, we used them jointly to obtain a single-label prediction for each item in the dataset.

As one of the conclusions of this work, we would like to raise some issues with the evaluation method. Possibly, some metric that also considers the semantics in the taxonomy might have enabled a more effective evaluation and allowed for insights into the inner workings of the systems, especially in connection with the misclassified items. One such metric was proposed by Chen et al. [9], which evaluates the performance of taxonomic assignments based on said given taxonomy.

Our system achieved the highest precision and F1 and the second-best accuracy and recall values of all submissions, demonstrating its effectiveness. While the ceiling seems to have been reached of what was possible in the shared task, judging by the range of the top submissions We hope to have contributed to the still challenging classification of research fields for scientific publications.

Acknowledgments

References

1. Abu Ahmad, R., Borisova, E., Rehm, G.: FoRC-Subtask-I@NSLP2024 Testing Data (2024). https://doi.org/10.5281/zenodo.10469550, https://doi.org/10.5281/zenodo.10469550
2. Abu Ahmad, R., Borisova, E., Rehm, G.: FoRC-Subtask-I@NSLP2024 Training and Validation Data (2024). https://doi.org/10.5281/zenodo.10438530, https://doi.org/10.5281/zenodo.10438530
3. Auer, S., Oelen, A., Haris, M., Stocker, M., D'Souza, J., Farfar, K.E., Vogt, L., Prinz, M., Wiens, V., Jaradeh, M.Y.: Improving access to scientific literature with knowledge graphs. Bibliothek Forschung und Praxis **44**(3), 516–529 (2020)
4. Balabantaray, R.C., Sarma, C., Jha, M.: Document clustering using k-means and k-medoids. CoRR **abs/1502.07938** (2015), http://arxiv.org/abs/1502.07938
5. Beltagy, I., Lo, K., Cohan, A.: Scibert: A pretrained language model for scientific text. In: Inui, K., Jiang, J., Ng, V., Wan, X. (eds.) Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019. pp. 3613–3618. Association for Computational Linguistics (2019). https://doi.org/10.18653/V1/D19-1371, https://doi.org/10.18653/v1/D19-1371
6. Bishop, C.M.: Pattern recognition and machine learning, 5th Edition. Information science and statistics, Springer (2007), https://www.worldcat.org/oclc/71008143
7. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003), http://jmlr.org/papers/v3/blei03a.html
8. Bravo-Alcobendas, D., Sorzano, C.: Clustering of biomedical scientific papers. In: 2009 IEEE International Symposium on Intelligent Signal Processing. pp. 205–209. IEEE (2009)
9. Chen, C.Y., Tang, S.L., Chou, S.C.T.: Taxonomy based performance metrics for evaluating taxonomic assignment methods. BMC Bioinformatics **20**(1), 310 (Jun 2019). https://doi.org/10.1186/s12859-019-2896-0, https://doi.org/10.1186/s12859-019-2896-0
10. Chen, T., Guestrin, C.: Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. p. 785–794. KDD '16, Association for Computing Machinery, New York, NY, USA (2016). https://doi.org/10.1145/2939672.2939785, https://doi.org/10.1145/2939672.2939785
11. Cherti, M., Beaumont, R., Wightman, R., Wortsman, M., Ilharco, G., Gordon, C., Schuhmann, C., Schmidt, L., Jitsev, J.: Reproducible scaling laws for contrastive language-image learning. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023. pp. 2818–2829. IEEE (2023). https://doi.org/10.1109/CVPR52729.2023.00276

12. Cohan, A., Feldman, S., Beltagy, I., Downey, D., Weld, D.S.: SPECTER: document-level representation learning using citation-informed transformers. In: Jurafsky, D., Chai, J., Schluter, N., Tetreault, J.R. (eds.) Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020. pp. 2270–2282. Association for Computational Linguistics (2020). https://doi.org/10.18653/V1/2020.ACL-MAIN.207, https://doi.org/10.18653/v1/2020.acl-main.207

13. Cortes, C., Vapnik, V.: Support-vector networks. Mach. Learn. **20**(3), 273–297 (1995). https://doi.org/10.1007/BF00994018, https://doi.org/10.1007/BF00994018

14. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Burstein, J., Doran, C., Solorio, T. (eds.) Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019). https://doi.org/10.18653/v1/N19-1423, https://aclanthology.org/N19-1423

15. Ho, T.K.: Random decision forests. In: Third International Conference on Document Analysis and Recognition, ICDAR 1995, August 14 - 15, 1995, Montreal, Canada. Volume I. pp. 278–282. IEEE Computer Society (1995). https://doi.org/10.1109/ICDAR.1995.598994, https://doi.org/10.1109/ICDAR.1995.598994

16. Kim, S., Gil, J.: Research paper classification systems based on TF-IDF and LDA schemes. Hum. centric Comput. Inf. Sci. **9**, 30 (2019). https://doi.org/10.1186/S13673-019-0192-7, https://doi.org/10.1186/s13673-019-0192-7

17. Lee, D.D., Seung, H.S.: Learning the parts of objects by non-negative matrix factorization. Nature **401**(6755), 788–791 (1999)

18. Littlestone, N., Warmuth, M.K.: The weighted majority algorithm. Information and computation **108**(2), 212–261 (1994)

19. Lo, K., Shen, Z., Newman, B., Chang, J., Authur, R., Bransom, E., Candra, S., Chandrasekhar, Y., Huff, R., Kuehl, B., Singh, A., Wilhelm, C., Zamarron, A., Hearst, M.A., Weld, D., Downey, D., Soldaini, L.: PaperMage: A unified toolkit for processing, representing, and manipulating visually-rich scientific documents. In: Feng, Y., Lefever, E. (eds.) Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. pp. 495–507. Association for Computational Linguistics, Singapore (Dec 2023). https://doi.org/10.18653/v1/2023.emnlp-demo.45, https://aclanthology.org/2023.emnlp-demo.45

20. Nguyen, T.H., Shirai, K.: Text classification of technical papers based on text segmentation. In: Métais, E., Meziane, F., Saraee, M., Sugumaran, V., Vadera, S. (eds.) Natural Language Processing and Information Systems - 18th International Conference on Applications of Natural Language to Information Systems, NLDB 2013, Salford, UK, June 19-21, 2013. Proceedings. Lecture Notes in Computer Science, vol. 7934, pp. 278–284. Springer (2013). https://doi.org/10.1007/978-3-642-38824-8_25, https://doi.org/10.1007/978-3-642-38824-8_25

21. Opitz, D., Maclin, R.: Popular ensemble methods: An empirical study. Journal of artificial intelligence research **11**, 169–198 (1999)

22. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P., Li, S., Misra, I., Rabbat, M.G., Sharma, V., Synnaeve, G., Xu, H., Jégou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: Di-

nov2: Learning robust visual features without supervision. CoRR **abs/2304.07193** (2023). https://doi.org/10.48550/ARXIV.2304.07193

23. Ostendorff, M., Rethmeier, N., Augenstein, I., Gipp, B., Rehm, G.: Neighborhood contrastive learning for scientific document representations with citation embeddings. In: Goldberg, Y., Kozareva, Z., Zhang, Y. (eds.) Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. pp. 11670–11688. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Dec 2022). https://doi.org/10.18653/v1/2022.emnlp-main.802, https://aclanthology.org/2022.emnlp-main.802

24. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al.: Scikit-learn: Machine learning in python. Journal of Machine Learning Research **12**, 2825–2830 (2011)

25. Ramos, J., et al.: Using tf-idf to determine word relevance in document queries. In: Proceedings of the first instructional conference on machine learning. vol. 242, pp. 29–48. Citeseer (2003)

26. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., Schramowski, P., Kundurthy, S., Crowson, K., Schmidt, L., Kaczmarczyk, R., Jitsev, J.: LAION-5B: an open large-scale dataset for training next generation image-text models. In: Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A. (eds.) Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022 (2022), http://papers.nips.cc/paper_files/paper/2022/hash/a1859debfb3b59d094f3504d5ebb6c25-Abstract-Datasets_and_Benchmarks.html

27. Taheriyan, M.: Subject classification of research papers based on interrelationships analysis. In: Proceedings of the 2011 workshop on Knowledge discovery, modeling and simulation. pp. 39–44 (2011)

28. Tsoumakas, G., Katakis, I., Vlahavas, I.P.: Mining multi-label data. In: Maimon, O., Rokach, L. (eds.) Data Mining and Knowledge Discovery Handbook, 2nd ed, pp. 667–685. Springer (2010). https://doi.org/10.1007/978-0-387-09823-4_34, https://doi.org/10.1007/978-0-387-09823-4_34

29. Wade, A.D.: The semantic scholar academic graph (s2ag). In: Companion Proceedings of the Web Conference 2022. p. 739. WWW '22, Association for Computing Machinery, New York, NY, USA (2022). https://doi.org/10.1145/3487553.3527147, https://doi.org/10.1145/3487553.3527147

30. Zhang, M., Zhou, Z.: ML-KNN: A lazy learning approach to multi-label learning. Pattern Recognit. **40**(7), 2038–2048 (2007). https://doi.org/10.1016/J.PATCOG.2006.12.019, https://doi.org/10.1016/j.patcog.2006.12.019