An Extended Evaluation of Single-Label Multi-Modal Field of Research Classification Using a Taxonomy-Based Metric

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. This paper presents a multi-modal approach to the single-label field of research classification shared task. Our method, **SLAMFORC**, incorporates metadata, full text, and image data from scholarly articles to generate comprehensive document embeddings. We built a voting ensemble of pre-trained BERT models (SciBERT and SciNCL) and traditional classifiers and achieved competitive performance in the *Field of Research Classification of Scholarly Publications* shared task. SLAMFORC scored highest in F1 score and precision and second best in recall and accuracy. We extend our original analysis by examining misclassified samples to improve future iterations. Additionally, we apply a taxonomy-based evaluation metric to better assess our results.

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

1 Introduction

Keywords and other classifications may help when searching or organizing scholarly publications [22]. 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 [15], has recently gained momentum. Still, the automated classification of research papers remains challenging [31].

This paper describes an extension to our original submission [29] for the shared task *Field of Research Classification of Scholarly Publications*[1] [3] of the *1st Workshop on Natural Scientific Language Processing and Research Knowledge Graphs (NSLP 2024)* [28]. 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 around $60,000$ English scientific papers [1,2], each

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

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 [15] model and, subsequently, fed these vectors to a handful of traditional classifiers. Then, we applied a voting ensemble [23] 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 (±0.75%), 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%).

In addition to the original paper, we perform an extended analysis of our results to better understand the misclassified samples in the hopes of enabling future systems to avoid these fallacies in the field of research classification. We identify problematic cases and point out possible improvements for future iterations. Also, we add a taxonomy-based metric to the evaluation. Wwe use it to show that SLAMFORC classifies the majority of the samples correctly but struggles in some cases.

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. Then, Section 5 analyzes the results more in-depth. Finally, we draw conclusions in Section 6.

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 [9] groups biomedical papers by applying non-negative matrix factorization [19] to the term relevance vectors of the documents. It uses bisecting k-means clustering [7], and, at the same time, assigns semantic meaning to each document and cluster inferred from the matrix decompositions.

The work by Taheriyan [31] describes an approach to classifying papers by using relationships such as common authors and references as well as citations in

[2] https://orkg.org/fields
[3] https://www.crossref.org/

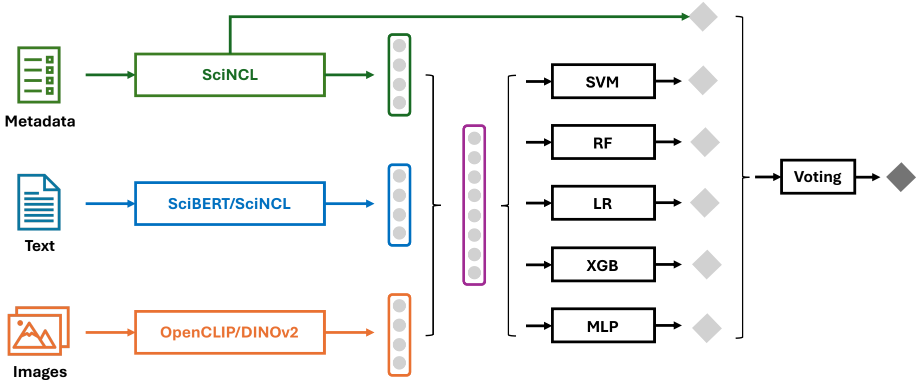

Fig. 1. Overview of the system architecture.

a graph. This information allows new papers to be assigned topics automatically instead of requiring manual annotations.

Nguyen and Shirai [22] focus on various text features such as the segmentation of the paper and apply three different classifiers: multi-label kNN [34], binary approach [32], 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.

Another approach is presented by Kim and Gil [18]: They describe a classification system based on latent Dirichlet allocation [8] and term frequency-inverse document frequency [27]. The former is employed to extract relevant keywords from the abstracts, the latter for k-means clustering [5] papers with similar topics.

More recently, SPECTER [13] uses pre-trained language models (e.g., SciBERT [6]) 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]

[4] https://gitlab.ifi.uzh.ch/DDIS-Public/forc24

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* [4], *arXiv*,[5] *Crossref*,[6] and the *Semantic Scholar Academic Graph* [33]. It 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 [25], 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 [21]. 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 [6] and SciNCL [25]. Because of BERT's limitation to processing 512 tokens at a time [15] 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.

[5] https://arxiv.org

[6] https://www.crossref.org/

[7] https://huggingface.co/spaces/rabuahmad/forcI-taxonomy/blob/main/taxonomy.json

[8] https://github.com/fabiobatalha/crossrefapi

[9] https://api.staging.crossref.org/swagger-ui/index.html

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 [12] model pre-trained on the LAION-5B dataset [30] as well as a pre-trained DINOv2 [24] 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 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 [23]. 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 [14], 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) [17] since it reduces 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) [11], we used another popular method that can achieve good performance while sacrificing interpretability. Next, we also employed a fully-connected Multilayer Perceptron (MLP) to deal with the nonlinearity of the data. Furthermore, we fine-tuned SciNCL [25] as an end-to-end solution on the metadata.

Finally, we combined the output of the classifiers described above into an ensemble method [23] with hard voting [20]. 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 [26] 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 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.

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

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

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 decently independently. 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 shows 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

[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 underlined). Our submission is the first line.

ID	Accuracy	Precision	Recall	F1 ↓
683689	0.756	**0.757**	0.756	**0.754**
688995	0.754	0.755	0.754	0.752
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

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.

5 Analysis

In this section, we perform an extended analysis of our results from the shared task *Field of Research Classification of Scholarly Publications* [3]. First, we explore which categories of misclassified samples contributed negatively to our F1 score. Then, we investigate how collapsing the field of research categories to their respective super-classes affects our results. Finally, we perform an additional evaluation using a metric with regard to taxonomy-based assignments that also takes the distances among the categories into account. Data and code of the extended analysis are available in our online repository.[4]

5.1 Impact of Misclassified Samples

To analyze the misclassification errors of SLAMFORC, we look into the categories that contributed negatively to our F1 score. That is fields of research classes for which the F1 is below our threshold of 75.4%. Additionally, to avoid overfitting, we also set the requirements that these categories need to make up at least 1% of the test data set; i.e., they need to have a support of at least 89. Table 4 shows the 12 categories we identified in this manner, along with their top-level class, precision, recall, F1, and number of samples (support) in the test data set.

Table 4. Results for categories making up at least 1% of the test data set ($n \geq 89$) that influenced our F1 results negatively.

Category	**Top-Level**	**Precision**	**Recall**	**F1 ↑**	**Support**
Computer Sciences	Physical Sciences & Mathematics	0.39	0.37	0.38	112
Applied Mathematics	Physical Sciences & Mathematics	0.56	0.53	0.54	142
Mathematics	Physical Sciences & Mathematics	0.52	0.64	0.57	145
Condensed Matter Physics	Physical Sciences & Mathematics	0.58	0.59	0.59	200
Analysis	Physical Sciences & Mathematics	0.62	0.68	0.65	131
Artificial Intelligence	Physical Sciences & Mathematics	0.65	0.67	0.66	137
Cosmology	Physical Sciences & Mathematics	0.73	0.61	0.66	137
Algebra	Physical Sciences & Mathematics	0.68	0.65	0.67	92
External Galaxies	Physical Sciences & Mathematics	0.69	0.66	0.68	162
Atomic, Molecular and Optical Physics	Physical Sciences & Mathematics	0.65	0.72	0.69	197
Theory/Algorithms	Physical Sciences & Mathematics	0.78	0.62	0.69	99
Materials Science and Engineering	Engineering	0.71	0.70	0.71	239

The *Computer Sciences* category exhibits the lowest F1 score by far at 0.38. This indicates significant challenges in correctly identifying true positives, leading to a substantial number of false positives and false negatives, as evident by precision (0.39) and recall (0.37). It is important to note that this only concerns samples with the target field of research *Computer Sciences* and none of its various sub-fields. Therefore, it is understandable that misclassifications occur as they might also be rooted in the ground truth annotations, which may not be specific enough. Accordingly, we cannot necessarily fault SLAMFORC for this shortcoming but rather a closer look at this part of the data set is required.

The other two categories with striking results are *Theory/Algorithmns* and *Cosmology*. Both show a distinctly higher precision (0.73 and 0.78, respectively) than recall (0.61 and 0.62, respectively). This indicates the tendency of SLAMFORC to miss many true instances but also to have few false positives. Thus, the capabilities of future iterations of SLAMFORC to capture these two specific categories should be addressed in particular.

The results for the other categories in Table 4 do not reveal any useful new insights. The balance between precision and recall numbers suggests that SLAMFORC misclassifies these categories by including incorrect and missing actual instances. This shows the need for improved model discrimination capabilities in these fields.

Finally, Figure 2 shows the normalized confusion matrix of the top-level classes for the categories in Table 4. Notably, *Arts and Humanities* is absent from the plot since no sample was misclassified from or to it.

We can observe that SLAMFORC largely classifies the samples in (the subcategories of) *Physical Sciences & Mathematics* correctly (0.94). Only a small fraction of 0.04 is misclassified as belonging to *Engineering* or one of its subcategories. Vice versa, there is a significant amount of samples from *Engineering* that are wrongly predicted to be from *Physical Sciences & Mathematics* or one of its sub-categories (0.29). This is understandable since there is some connection between the two fields since it can be considered an application of *Physical Sciences & Mathematics*. Still, it highlights a shortcoming of SLAMFORC and

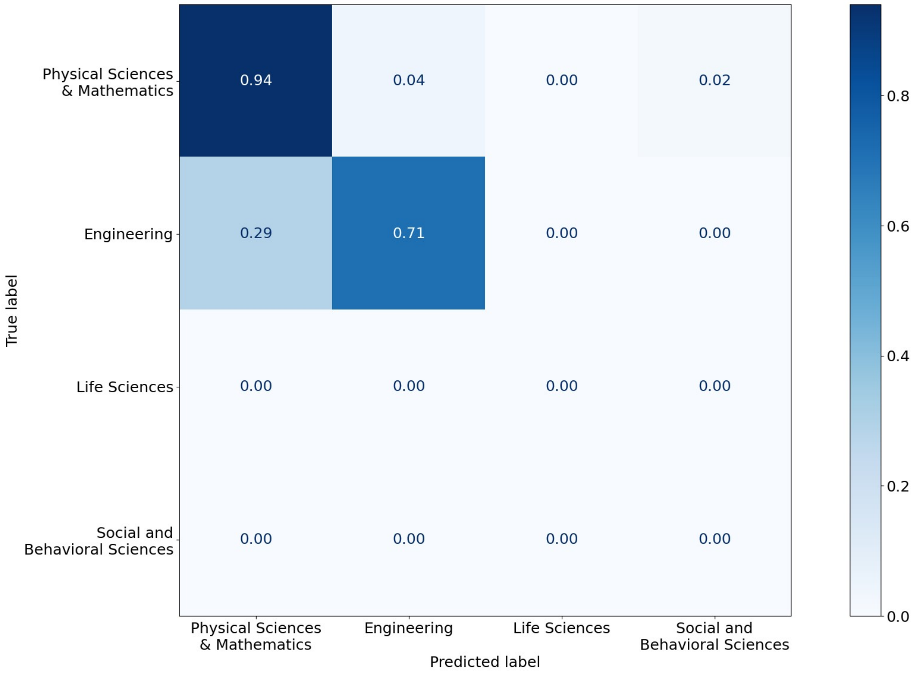

Fig. 2. Normalized confusion matrix of the top-level categories for the results negatively contributing to our F1 (see Table 4. Notice the absence of the top-level category *Arts and Humanities*.

its underlying classifiers. All other cells in the confusion matrix show negligible percentages of misclassification (i.e., ≤ 0.01).

5.2 Collapsing to Super-Classes

Table 5 shows accuracy, precision, recall, and F1 for the original results and their collapsing into the super-classes of the predicted research fields in three steps up to the top-level categories. For example, *Machine Learning* would be collapsed into *Artificial Intelligence* in the first step, then into *Computer Sciences* in the next, and, finally, into *Physical Sciences & Mathematics*. For taxa of lower than maximum rank in the taxonomy, they are collapsed into their parents up to the top level and then stay there for subsequent steps (e.g., *Computational Linguistics* → *Linguistics* → *Social and Behavioral Sciences* → *Social and Behavioral Sciences*). Details can be found in the ORKG taxonomy,[7] and the parent classes are also described in our online repository. Even though we used the same evaluation routines, the results slightly differ from the numbers reported in the shared task.[10] We rectified some errors in the annotations and taxonomy, all detailed in our online repository.

Table 5. Results when collapsing the categories to their super-classes in three steps to the top-level.

Collapse	**Accuracy**	**Precision**	**Recall**	**F1**
None	0.75	0.75	0.75	0.75
One Up	0.82	0.82	0.82	0.82
Two Up	0.93	0.93	0.93	0.93
Top-Level	0.94	0.94	0.94	0.94

Unsurprisingly, accuracy, precision, recall, and F1 improve for each collapse into the super-classes. After the first step, the metrics increase from 75% to 82%. The second collapse exhibits the largest jump for all four metrics to 93%. Finally, when all fields of research are collapsed into one of their five top-level categories (*Physical Sciences & Mathematics*, *Engineering*, *Life Sciences*, *Social and Behavioral Sciences*, *Arts and Humanities*), we achieve an accuracy, precision, recall, and F1 of 94%. This only being a small step up from the previous may be owed to the fact that few categories have maximum rank in the taxonomy. Still, these can be considered impressive results given the complexity of predicting the field of research for papers.

Figure 3 shows the normalized confusion matrix of the top-level classes for all fields of research in the taxonomy. As the majority of misclassifications by SLAMFORC is to assign them to *Physical Sciences & Mathematics* or one of its sub-categories, we see a tendency. The second-most wrongly predicted subcategories are from the top-level class of *Engineering*. As mentioned earlier, there is a connection between these two fields. Still, this reveals a possible bias in SLAMFORC and its underlying classifiers.

5.3 Taxonomy-Based Evaluation

Figure 4 shows the confusion matrix as a heatmap for all categories in the taxonomy. The full list for the axis labels can be found in the online repository.[4] It illustrates that most samples are predicted correctly, indicated by the strong diagonal line. Still, the following classes do not have any correctly classified instances in the test set: *Biomedical Engineering and Bioengineering*, *Cell Behavior*, *Civil and Environmental Engineering*, *Energy Systems*, *Hardware Systems*, *Molecular, Cellular, and Tissue engineering*, *OS/Networks*, *Other Quantitative Biology*, *Physical Sciences & Mathematics*, *Popular Physics*, *Quantitative Methods*, *Space Physics*, *Tissues and Organs*. Their support ranges from one to eight, making them only very small fractions considering the size of the test set ($n = 8903$).

The application of traditional metrics, like done in the shared task, is, thus, problematic. They score everything not on the diagonal of the confusion matrix as zeroes by treating all the classes as independent. Still, the distribution in Figure 4 shows there are some connections between the labels. There is a defined hierarchy in the taxonomy and, inherently, some semantics. The labels on the

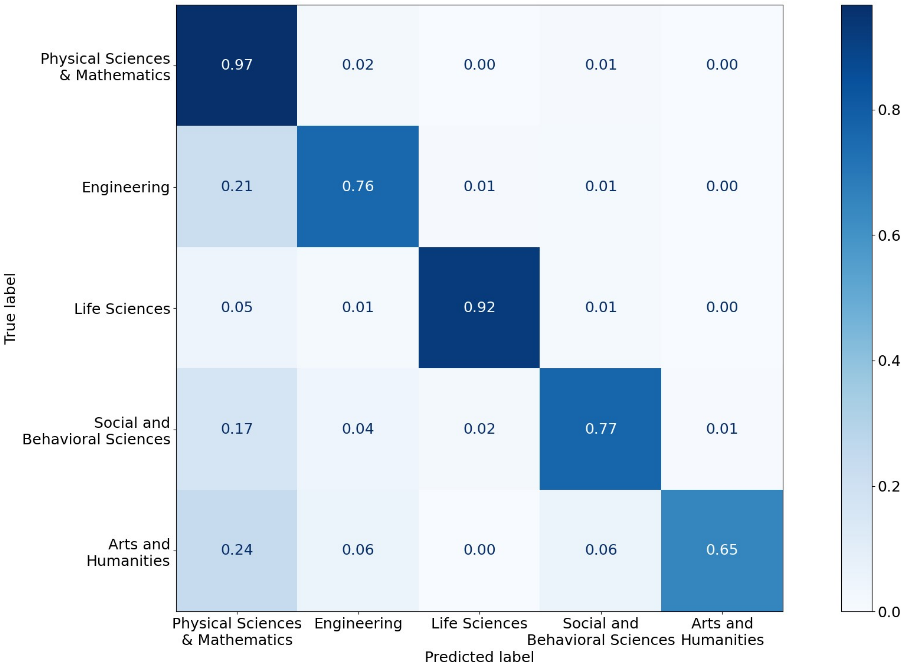

Fig. 3. Normalized confusion matrix of our results for the five top-level categories.

axis are grouped by that hierarchy. Therefore, not every distance is equal, but proximity in the confusion matrix also indicates "less of an error".

Table 6 shows the label pairs with a misclassification rate of more than 0.5. Looking at these sets, it becomes evident that the errors should not be punished by giving a score of zero instead of one (i.e., giving an error of one) since the classes in the pairs are somewhat related. Accordingly, we should evaluate considering these distances.

Hence, we applied a taxonomy-based metric [10,16] to assess the performance of SLAMFORC by considering the hierarchical structure of the labels. Inspired by the taxonomy similarity (TS) definition [16], we define the taxonomy distance $TD = 1 - TS$. For two labels a and b with the hierarchical sequences $a_1 \dots a_m$ and $b_1 \dots b_n$ respectively, the TS is the following:

$$TS = \frac{\sum_{l=1}^{\min(m,n)} 1[a_l = b_l]}{\max(m,n)}$$

Hence, we obtain $0 \leq TD = 1 - TS \leq 1$, where bigger signifies better. For the example above, the hierarchical sequence would look as follows: $a_1 =$ *Social and Behavioral Sciences*, $a_2 =$ *Linguistics*, $a_3 =$ *Computational Linguistics*.

The average taxonomy distance (ATD) is defined as the mean of the taxonomy distances between each sequence and its predicted label, calculated by

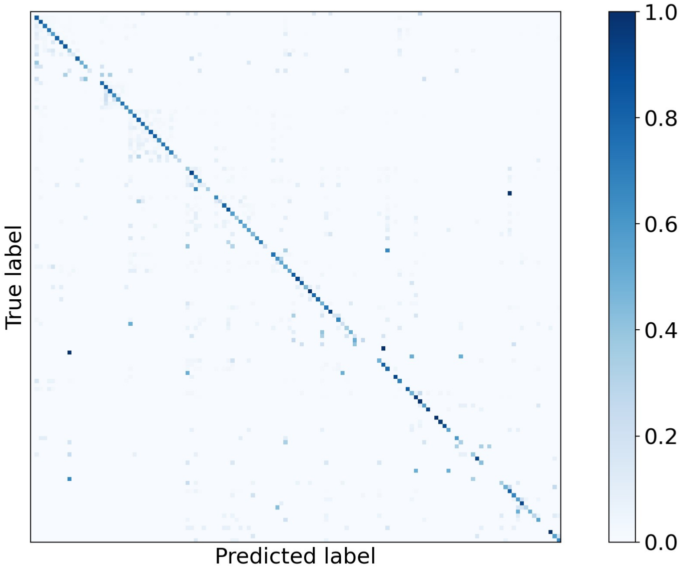

Fig. 4. Normalized confusion matrix of our results for the whole taxonomy. The full list can be found in the online repository.

the sum of these distances divided by the number of instances [10]. Figure 5 shows that the majority of the classifications are accurate, as indicated by the initial flat segment with zero taxonomy distance. However, there is a sharp increase towards the end, which reflects the misclassification in some instances. This suggests that SLAMFORC performs well overall, but it struggles with a few challenging instances that result in larger taxonomy distances.

6 Conclusions

In this paper, we presented an extended analysis of **SLAMFORC** [29], a system for the *Single-Label Multi-modal Field of Research Classification.* It was used to produce the results for our submission to the shared task *Field of Research Classification of Scholarly Publications* [3] of the *1st Workshop on Natural Scientific Language Processing and Research Knowledge Graphs (NSLP 2024)* [28]. 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.

Our system achieved the highest precision and F1 and the second-best accuracy and recall values of all submissions, demonstrating its effectiveness. Judging

Table 6. Label pairs where the misclassification rate exceeds 0.50.

True Label	Predicted Label	Miscl. Rate	Support
Audio and Speech Processing	Computational Linguistics	1.00	2
Biomedical Engineering and Bioengineering	Computer and Systems Architecture	1.00	1
Molecular, Cellular, and Tissue Engineering	Biological and Chemical Physics	1.00	1
Multiagent Systems	Artificial Intelligence	0.67	3
OS/Networks	Digital Communications and Networking	0.67	3
Cell Behavior	Biological and Chemical Physics	0.67	3
Controls and Control Theory	Mathematics	0.50	24
Energy Systems	Mechanical Engineering	0.50	4
Tissues and Organs	Life Sciences	0.50	2
Tissues and Organs	Quantitative Methods	0.50	2
Hardware Systems	Computer Sciences	0.50	2
Hardware Systems	Controls and Control Theory	0.50	2
Genomics	Bioinformatics	0.50	2
Genomics	Medicine	0.50	2
Economic Theory	Economics	0.50	2

by the range of top submissions, the ceiling of what was possible in the shared task seems to have been reached.

This paper adds to the original publication by performing an extended analysis of the results and also incorporates a taxonomy-based metric. We investigated the classes that contributed negatively to our F1 score. Furthermore, we explore how collapsing the categories to their parents affects the results, showing an increase in each step. Finally, we use the taxonomy distance, which is based on the taxonomy similaritys [16], to show that SLAMFORC classifies the majority of the samples correctly, struggling only with a few challenging instances.

Acknowledgments

This work was partially funded by the University Research Priority Program "Dynamics of Healthy Aging" at the University of Zurich and the Swiss National Science Foundation through Projects "CrowdAlytics" (Grant Number 184994) and "MediaGraph" (Grant Number 202125).

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. Ahmad, R.A., Borisova, E., Rehm, G.: FoRC@NSLP2024: Overview and Insights from the Field of Research Classification Shared Task. In: Rehm, G., Schimmler, S., Dietze, S., Krüger, F. (eds.) Proceedings of the Workshop on Natural Scientific Language Processing and Research Knowledge Graphs (NSLP 2024; co-located with ESWC 2024). Hersonissos, Greece (5 2024), 27 May 2024. Accepted for publication.

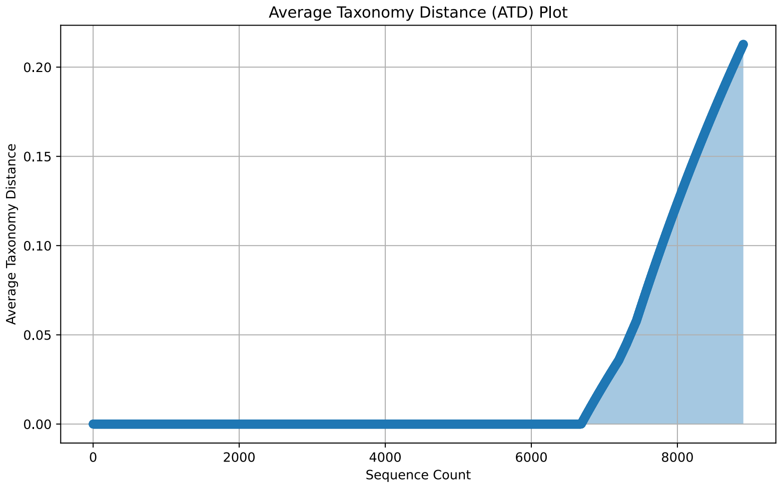

Fig. 5. Average Taxonomy Distance (ATD) plot.

4. 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)
5. 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
6. 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
7. Bishop, C.M.: Pattern recognition and machine learning, 5th Edition. Information science and statistics, Springer (2007), https://www.worldcat.org/oclc/71008143
8. 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
9. Bravo-Alcobendas, D., Sorzano, C.: Clustering of biomedical scientific papers. In: 2009 IEEE International Symposium on Intelligent Signal Processing. pp. 205–209. IEEE (2009)
10. 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
11. 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 Ma-

chinery, New York, NY, USA (2016). https://doi.org/10.1145/2939672.2939785, https://doi.org/10.1145/2939672.2939785

12. 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
13. 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
14. 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
15. 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
16. Hasson, I., Novgorodov, S., Fuchs, G., Acriche, Y.: Category recognition in e-commerce using sequence-to-sequence hierarchical classification. In: Proceedings of the 14th ACM International Conference on Web Search and Data Mining. pp. 902–905 (2021)
17. 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
18. 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
19. Lee, D.D., Seung, H.S.: Learning the parts of objects by non-negative matrix factorization. Nature **401**(6755), 788–791 (1999)
20. Littlestone, N., Warmuth, M.K.: The weighted majority algorithm. Information and computation **108**(2), 212–261 (1994)
21. 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
22. 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

23. Opitz, D., Maclin, R.: Popular ensemble methods: An empirical study. Journal of artificial intelligence research **11**, 169–198 (1999)
24. 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.: Dinov2: Learning robust visual features without supervision. CoRR **abs/2304.07193** (2023). https://doi.org/10.48550/ARXIV.2304.07193
25. 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
26. 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)
27. 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)
28. Rehm, G., Schimmler, S., Dietze, S., Krüger, F. (eds.): Proceedings of the Workshop on Natural Scientific Language Processing and Research Knowledge Graphs (NSLP 2024; co-located with ESWC 2024). No. 14770 in Lecture Notes in Artificial Intelligence (LNAI), Springer (2024), in print
29. Ruosch, F., Vasi, R., Wang, R., Rossetto, L., Bernstein, A.: Single-Label Multi-Modal Field of Research Classification. In: Rehm, G., Schimmler, S., Dietze, S., Krüger, F. (eds.) Proceedings of the Workshop on Natural Scientific Language Processing and Research Knowledge Graphs (NSLP 2024; co-located with ESWC 2024). Hersonissos, Greece (5 2024), 27 May 2024. Accepted for publication.
30. 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
31. 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)
32. 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

33. 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
34. 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