


Advancing Automatic Subject Indexing: Combining Weak Supervision with Extreme Multi-label Classification

Lakshmi Rajendram Bashyam¹[0009–0009–9455–2829] 
and Ralf Krestel^{1,2}[0000–0002–5036–8589]

¹ ZBW – Leibniz Information Centre for Economics, Germany

l.rajendram-bashyam@zbw-online.eu

² Kiel University, Germany

rkr@informatik.uni-kiel.de

Abstract. The multi-label automatic classification of scientific publications based on a pre-defined taxonomy, also called automatic subject indexing is a continuing research endeavor with significant cross-domain applicability. In this paper, we assess the performance of X-transformer and its variants with other extreme multi-label classification models for the above task. Our model *Weak X-transformer* achieves a micro F1-score of 0.65 and 64% accuracy on the task outperforming all other methods. We also investigate the impact of incorporating additional unlabelled data and hierarchical structure into the models. Our findings demonstrate that the transformer-based model with weak supervision outperforms other approaches, providing insights into effective strategies for extreme multi-label classification in scholarly publications.

Keywords: Extreme multi-label classification · Automatic subject indexing · Digital libraries · Semi-supervised learning.

1 Introduction

Automatic subject indexing in libraries involves the use of computational techniques to assign relevant subject headings or descriptors to library resources such as books, articles, and other materials. This process utilizes algorithms, machine learning, and natural language processing to analyze the content of the resources and determine their main topics or subjects. By automating this task, libraries can efficiently organize their collections. Automatic subject indexing not only saves time and resources for librarians but also enhances the accuracy and consistency of indexing across the library catalog. In addition, automatic subject indexing also provides a keyword-based summary of the publication to the user. Additionally, it allows libraries to keep pace with the growing volume of digital materials and ensures that their collections remain organized and easily navigable in the digital age.

The subjects/labels assigned to the publication usually are obtained from a pre-defined thesaurus maintained by the respective subject authority. Typically

each publication is assigned multiple subjects by the indexers, thus making it a multi-label classification task. However, the nature of the topics observed in the publication makes the assigned subjects highly imbalanced. Depending on the type of thesaurus used, it can also be classified as an XMLC task.

In this work, we compare XMLC models on a shared task as a part of the challenge "FoRC: Field of Research Classification of Scholarly Publications" organized by the "Natural Scientific Language Processing and Research Knowledge Graphs (NSLP 2024)" workshop. For the second task "Fine-grained multi-label classification of Computational Linguistics scholarly papers", we evaluate multi-label and extreme multi-label classification models. Additionally tweaking it to support the hierarchical nature of the task.

2 Related Work

The task of extreme multi-label classification (XMLC) is characterized by an imbalanced distribution of labels, posing significant challenges, particularly in improving the performance of less frequent labels. BONSAI [5] and PARABEL [9] represent tree-based approaches widely adopted for addressing the XMLC problem. Building upon this foundation, the X-transformer [16] introduces innovations such as X-linear, recursive linear models, and XR-transformers, a transformer-based framework that recursively fine-tunes pre-trained transformers. The pecos [14] library offers a robust implementation of these models, along with several other recent XMLC solutions including PINA [2] and FINGER [1]. More recently, the XLGEN model [4] has explored leveraging a text-to-text transformer model to tackle the challenges posed by XMLC.

Assigning labels or subjects to scientific publications constitutes a fundamental aspect of library organization. The size of the thesaurus employed for this purpose varies significantly among different organizations. The annotators responsible for assigning labels typically adhere to a predefined methodology for conducting this task. However, due to the slow annotation process, a considerable number of publications remain unlabeled. Integrating semi-supervised techniques to augment the training data becomes imperative in such scenarios. Addressing this need, recent research [17] conducts a comprehensive analysis utilizing unlabeled data through weak supervision techniques. The authors compare the efficacy of well-known weak supervision methods, including COSINE and WRENCH [15], under real-world conditions characterized by limited availability of clean labels. Widely adopted libraries such as setfit [13] and skweak [6] have significantly contributed to the adoption of semi-supervised and weak supervision techniques in resource-constrained settings. A method for generating weak labels through a noisy labeling scheme and subsequent refinement via a two-level approach [3] offers a computationally efficient solution that can prove invaluable in low-resource scenarios.

Access to full-texts represents a significant asset within library systems. Annotators are furnished with the title, abstract, and full-text of a publication to ensure accurate labeling. Without all these resources, particularly full-texts,

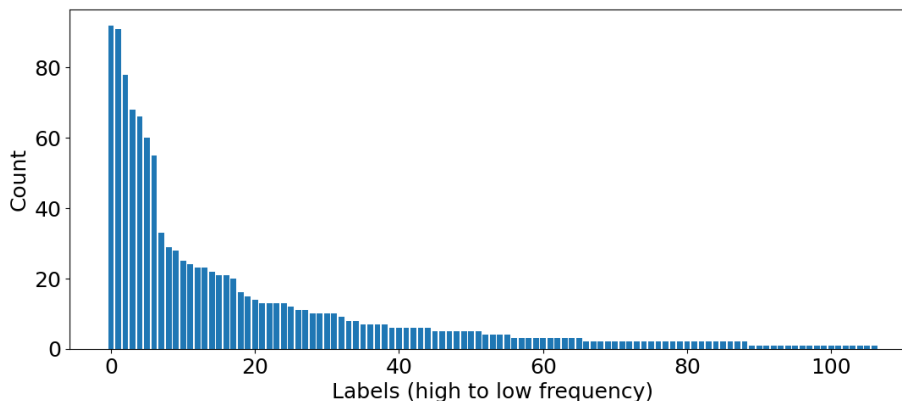


Fig. 1. Histogram of label distribution in the train set of FORC corpus

achieving human-level performance would pose a formidable challenge for any system. Previous research [7] has examined the performance of machine learning models when provided with an abundance of abstracts versus a comparatively limited number of full-texts.

3 Data Analysis

The training set for task 2 has 1051 publications. Each publishing record contains the title, abstract, and acl-id. Other information fields, such as date, venue, and publisher, are also available for the majority of records. Each publication receives three levels of coarse-to-fine-grained labels: Level 1, Level 2, and Level 3. The annotated labels are based on a predefined taxonomy created by extracting subjects from the publication and correlating them with existing topics from multiple paper sources. As specified by the workshop guidelines, the quality of annotations is evaluated based on inter-annotator agreement scores (IAA) using Krippendorff’s Alpha for multi-label annotations on each one of the three taxonomy levels. The average IAA scores for each level is given as part of the results in table 3.

Figure 1 shows the distribution of labels across all levels. The figure clearly shows that the annotated labels are severely unbalanced. It was also discovered that several of the labels used in the validation and test splits did not appear once in the training split. This makes it more difficult to detect these labels.

In addition to the existing data, we were able to collect the full text of every paper using the acl-id [11]. The average length of full-text was approximately 16,000 words. The same source also gave abstracts for over 70,000 other publications from ACL articles and posters. The publications collected ranged from 2001 to 2021 and were published in a variety of venues. The additional publication dataset also includes the title, abstract, full-text, and all other metadata elements found in the FORC dataset. However, no labels are associated with the

new dataset collected. Thus, we have approximately 70,000 unlabeled publication records from the same source as the FORC dataset.

4 Experiments

The text for training and testing the model is created by merging the fields *title*, *abstract*, *venue*, *publisher*, and *book title*, each separated by a specific token. By integrating the label sets from the three levels, we can define the Fine-grained multi-label classification task as a general multi-label classification. This allows us to use models intended for multi-label and XMLC tasks with minimal changes to the model.

The baseline model metrics were provided by the workshop committee. The metrics was reproduced by fine-tuning the scincl model [8] on FORC subtask 2 training dataset. The training text consists of combining only the title and abstract of each publication. Level wise labels for each publication were combined together as a list to form a multi-class multi-label classification task. The SciNCL is the state of the art pre-trained BERT language model to generate document-level embeddings of research papers. Since the training data for the FORC task consists of similar scientific documents, fine-tuning scincl model as baseline gives a good starting point.

We train a basic model like tf-idf on the train set in addition to the baseline that the organizers provided. Unlike the baseline, the tf-idf and all subsequent models was training on the train text created by merging *title*, *abstract*, *venue*, *publisher*, and *book title*, each separated by a specific token. The tf-idf model returns most similar subjects based on similarity in sparse tf-idf normalized bag-of-words vector space. Our XMLC models including tf-idf are trained using the Annif [12] toolset. Annif tf-idf implementation is based on the topic modelling library Gensim [10].

For this work, the label set size is 170. We model the problem as an extreme multi-label classification because this is on the higher end for a multi-label classification task. Regarding label occurrences, the label distribution in the picture likewise complies with Zipf’s law. The high imbalance in the label distribution can affect the performance of the model. The model performs best on labels that appear frequently in the training set and poorly on labels that are rare. This is a typical occurrence in contexts with extreme multi-label classification. For the reasons listed above, this task could still benefit from being modeled as an XMLC problem even though most XMLC projects have label sets larger than 500 labels.

We use tree-based models like parabel [9] in the XMLC space. The label space is divided recursively by these models. Because there are an equal number of labels in each cluster, it is balanced. We tokenize the dataset using a nltk tokenizer and choose trigram tokens for training. We tune the parameters of the model such as number of clusters and maximum tree depth on the validation split of the dataset.

Furthermore, we utilize the implementation of transformer based XMC framework, the X-transformer framework [16] provided by the Pecos library [14]. This XR-transformer framework allows fine-tuning pre-trained transformers recursively on multi-resolution objectives. There are three steps involved in the fine-tuning. For starters, the label space is clustered. In the second stage, a matcher is trained to classify the publication to one of the clusters and finally, a ranker is trained to rank the labels inside each cluster. Based on this method, we train the model on the training set for 10 epochs with early stopping.

Further, we were able to obtain the full text for each publication in train, val and test split [11]. Due to hardware restrictions and because BERT-based transformer models only support up to 512 tokens. However, we train the parabel model on the full texts called parabel-ft and evaluate on the full-text corpus of the test split.

With the new dataset gathered, we now have access to more fields and around 70000 more ACL abstracts. However, they are not tagged with any labels. First, we train the X-transformer model on this unlabeled dataset to further enhance its performance. We produce weak labels from each unlabelled publication to enable supervised training. The previously trained tf-idf model is used to generate the weak labels. Since the model’s performance determines the quality of these labels, they are noisy. We then combine the annotated clean labels and generate weak labels into the training dataset for the x-transformer model [17] [3]. Finally, the model is then fine-tuned on the dataset containing combination of weak and clean labels. We name this model the weak X-transformer owing to the training data containing weak/noisy labels.

There are three levels of subjects, ranging from coarse to fine-grained subjects. The levels designated as Level1 > Level2 > Level3 are arranged in a rigid hierarchy. The hierarchy of the labels is not taken into account when training the model because we utilize a general XMLC model. As an alternative, we add hierarchy to the output of the model. This is accomplished by comparing the label set of each publication to the taxonomy hierarchy. The labels in Levels 2 and 3 that do not adhere to the preceding level’s hierarchical structure are pruned from the collection of results. Furthermore, we only keep labels above a confidence score of 0.2 to increase the system’s accuracy.

5 Results and Discussion

Table 1 provides the performance metrics for each model in our experiment. Overall, the baseline model performance is subpar. The poor macro average scores for recall, precision, and F1 indicate that the model does not work well for multi-label task. This suggests that optimizing a transformer-based model directly for our goal is not a good idea.

One of the easiest models to train and assess is the tf-idf model. It is interesting to note that the tf-idf model has very good recall scores despite having lower precision scores than other models. One reason for this could be that the lexical matching component identifies many labels in the publication text in-

Table 1. Performance metrics of baseline and XMLC models

Model	Micro			Macro			Weighted		
	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1
weak x-transformer	0.64	0.65	0.65	0.56	0.52	0.52	0.64	0.65	0.63
x-transformer	0.70	0.60	0.64	0.50	0.37	0.40	0.66	0.60	0.61
parabel-ft	0.62	0.46	0.53	0.25	0.14	0.16	0.51	0.37	0.40
parabel	0.66	0.50	0.57	0.33	0.19	0.21	0.45	0.76	0.56
tfidf	0.41	0.75	0.53	0.44	0.54	0.45	0.47	0.75	0.55
baseline	0.35	0.32	0.34	0.01	0.04	0.02	0.16	0.33	0.17

creasing the recall scores. However, a large number of the labels in the result set have no semantic relevance to the publication. Thus, a simple model could not adequately represent this.

Table 1 further shows the outcome of our tree-based parabel approach studies. The performance gain over the tf-idf model is not evident in the findings. The macro average scores are in fact decreasing. This suggests that the model has extremely low performance on some classes. Upon examination of the data, we discovered that the label set contains a large quantity of false negatives. In the next experiment, we again train a parabel model using the full-text dataset for the publications to refine it. The results of the parabel-ft model also demonstrate a decline in macro average scores. This suggests that training using full-text does not always result in improved performance. The reason for the degradation in performance could be that the model cannot identify the relevant parts of the text correctly.

Out of the models previously discussed, the weak X-transformer model, which is a combination of weak supervision applied to an X-transformer model designed for extreme multi-label classification, exhibits the best performance. Table 2 provides a few examples of the labels generated by our final model. The predicted label set is sorted by the confidence score from the weak X-transformer model. In terms of micro average scores, the model maintains a relatively high recall score while improving on the precision score. Both the weighted average scores and the macro scores show similar results. It obtains weighted averages, macro, and micro F1 scores of 0.64, 0.65, and 0.63 respectively.

The table 3 provides the level-wise performance of the model in the label hierarchy. We compare the multi-label outputs of our model with the actual labels at every label hierarchy. With a level1 label set, the X-transformer model obtains a high score for precision and recall metric of 0.90 each. Performance slightly declines as one moves up the tiers of the label hierarchy. The gradual decrease through the label hierarchy is also observed in the inter-annotator agreement (IAA) scores.

In the realm of XMLC, a common hurdle lies in enhancing the performance of less frequent labels. As depicted in Figure 2, the precision scores of different models across each label class are illustrated. Labels are organized based on their occurrence frequency in descending order along the x-axis. Notably, the models

Table 2. Examples of labels obtained from the weak X-transformer model

ACL_ID:2021.sigdial-1.31		Title: Summarizing Behavioral Change Goals from SMS Exchanges to Support Health Coaches	
True Labels	Automatic Text Summarization	Predicted Labels	Model Architectures
	Discourse Analysis		Domain-specific NLP
	Domain-specific NLP		Data Management and Generation
	Data Management and Generation		Classification Applications
	Model Architectures		Discourse Analysis
	Transformer Models		Dialogue Systems
	Extractive Text Summarization		Medical and Clinical NLP
	Data Preparation		Data Preparation
	Medical and Clinical NLP		NLP for News and Media
ACL_ID:2021.unimplicit-1.1		Title: Let's be explicit about that: Distant supervision for implicit discourse relation classification via connective prediction	
True Labels	Learning Paradigms	Predicted Labels	Model Architectures
	Classification Applications		Learning Paradigms
	Discourse Analysis		Discourse Analysis
	Few-shot Learning		Classification Applications

generally exhibit superior performance for more commonly occurring labels compared to rare ones. Interestingly, the weak X-transformers model demonstrates higher precision scores for less frequent labels in comparison.

To improve on the previous model, we incorporate the unlabelled data into the training step. The result of the weak supervision/ weak x-transformer model is shown in the Table 1. While the enhancement in performance may not be substantial, the weak x-transformer still demonstrates superior performance across micro, macro, and weighted average scores in comparison. Both precision and recall scores are close to each other without any trade-offs of improvement of one score over the other. Hence, this model is comparatively better than the rest of the models.

Table 3. Performance metrics of weak X-transformer model by each level of hierarchy and the associated inter-annotator agreement

Hierarchy	Prec	Rec	F1	IAA
Level 1	0.80	0.92	0.90	0.67
Level 2	0.80	0.73	0.76	0.58
Level 3	0.75	0.70	0.72	0.54

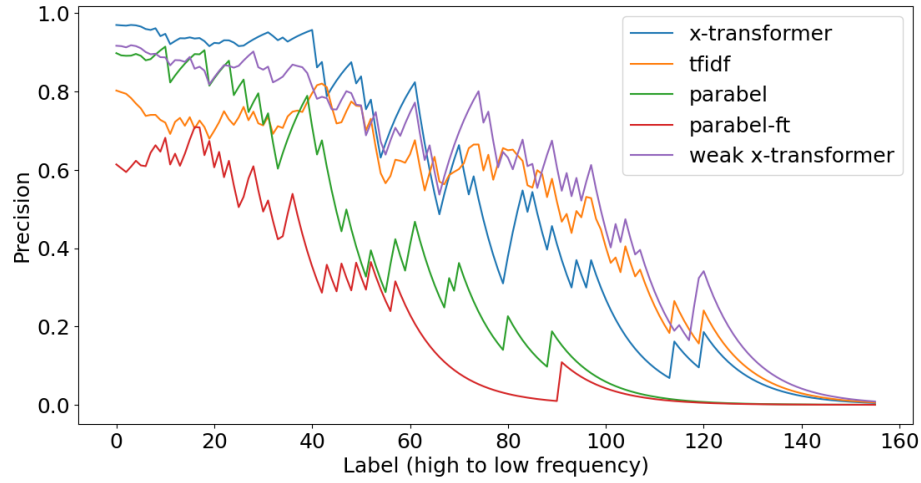


Fig. 2. Label-wise precision score for each model. The labels are arranged in descending order of count. For the presentation, the graph is smoothed.

Finally, in addition to the training of the model, label pruning is executed for each model. Pruning the labels that do not conform to hierarchical structure slightly improves the precision score by removing false labels without affecting the recall values negatively.

6 Conclusion and Future Work

This study presents an approach to modeling hierarchical fine-grain multi-label classification as an XMLC problem. Furthermore, we assess how well simple models such as tf-idf to complex transformer-based models perform for the given task. We also explore other approaches, such as training with full-text and utilizing unlabeled datasets in a weak supervision setting. Our weak X-transformer model, which is our best-performing model, attains an F1 score of 0.65 across all labels.

In future work, we would like to explore the possibility of incorporating hierarchy during training instead of at the output level. Furthermore, We would like to experiment with generative AI by grounding the model to a graph RAG for label space.

References

1. Chen, P., Chang, W.C., Jiang, J.Y., Yu, H.F., Dhillon, I., Hsieh, C.J.: Finger: Fast inference for graph-based approximate nearest neighbor search. In: Proceedings of the ACM Web Conference 2023. p. 3225–3235. WWW '23, Association for Computing Machinery, New York, NY, USA (2023). <https://doi.org/10.1145/3543507.3583318>

2. Chien, E., Zhang, J., Hsieh, C.J., Jiang, J.Y., Chang, W.C., Milenkovic, O., Yu, H.F.: Pina: Leveraging side information in extreme multi-label classification via predicted instance neighborhood aggregation. In: International Conference on Machine Learning. pp. 5616–5630. PMLR (2023)
3. Goh, G.B., Siegel, C., Vishnu, A., Hodas, N.: Using rule-based labels for weak supervised learning: A chemnet for transferable chemical property prediction. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. p. 302–310. KDD '18, Association for Computing Machinery, New York, NY, USA (2018). <https://doi.org/10.1145/3219819.3219838>
4. Jung, T., Kim, J.k., Lee, S., Kang, D.: Cluster-guided label generation in extreme multi-label classification. In: Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics. pp. 1670–1685. Association for Computational Linguistics (2023). <https://doi.org/10.18653/v1/2023.eacl-main.122>
5. Khandagale, S., Xiao, H., Babbar, R.: Bonsai: diverse and shallow trees for extreme multi-label classification. Machine Learning **109**(11), 2099–2119 (2020). <https://doi.org/10.1007/s10994-020-05888-2>
6. Lison, P., Barnes, J., Hubin, A.: skweak: Weak supervision made easy for NLP. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations. pp. 337–346. Association for Computational Linguistics, Online (2021). <https://doi.org/10.18653/v1/2021.acl-demo.40>
7. Mai, F., Galke, L., Scherp, A.: Using deep learning for title-based semantic subject indexing to reach competitive performance to full-text. In: Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries. JCDL '18, ACM (2018). <https://doi.org/10.1145/3197026.3197039>
8. Ostendorff, M., Rethmeier, N., Augenstein, I., Gipp, B., Rehm, G.: Neighborhood Contrastive Learning for Scientific Document Representations with Citation Embeddings. In: The 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP 2022). Association for Computational Linguistics, Abu Dhabi (2022). <https://doi.org/10.48550/arXiv.2202.06671>, 7–11 December 2022. Accepted for publication.
9. Prabhu, Y., Kag, A., Harsola, S., Agrawal, R., Varma, M.: Parabel: Partitioned label trees for extreme classification with application to dynamic search advertising. In: Proceedings of the 2018 World Wide Web Conference. p. 993–1002. WWW '18, International World Wide Web Conferences Steering Committee (2018). <https://doi.org/10.1145/3178876.3185998>
10. Rehurek, R., Sojka, P.: Gensim–python framework for vector space modelling. NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic **3**(2) (2011)
11. Rohatgi, S.: Acl anthology corpus with full text. Github (2022), <https://github.com/shauryr/ACL-anthology-corpus>
12. Suominen, O.: Annif: Diy automated subject indexing using multiple algorithms. LIBER Quarterly: The Journal of the Association of European Research Libraries **29**(1), 1–25 (Jul 2019). <https://doi.org/10.18352/lq.10285>, <https://liberquarterly.eu/article/view/10732>
13. Tunstall, L., Reimers, N., Jo, U.E.S., Bates, L., Korat, D., Wasserblat, M., Pereg, O.: Efficient few-shot learning without prompts (2022). <https://doi.org/10.48550/ARXIV.2209.11055>

14. Yu, H.F., Zhong, K., Zhang, J., Chang, W.C., Dhillon, I.S.: Pecos: Prediction for enormous and correlated output spaces. *Journal of Machine Learning Research* **23**(98), 1–32 (2022)
15. Yu, Y., Zuo, S., Jiang, H., Ren, W., Zhao, T., Zhang, C.: Fine-tuning pre-trained language model with weak supervision: A contrastive-regularized self-training approach. In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. pp. 1063–1077 (2021). <https://doi.org/10.18653/v1/2021.naacl-main.84>
16. Zhang, J., Chang, W., Yu, H., Dhillon, I.S.: Fast multi-resolution transformer fine-tuning for extreme multi-label text classification. In: *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems (NeurIPS)*. pp. 7267–7280 (2021)
17. Zhu, D., Shen, X., Mosbach, M., Stephan, A., Klakow, D.: Weaker than you think: A critical look at weakly supervised learning. In: Rogers, A., Boyd-Graber, J., Okazaki, N. (eds.) *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. pp. 14229–14253. Association for Computational Linguistics, Toronto, Canada (2023). <https://doi.org/10.18653/v1/2023.acl-long.796>