Investigating Hierarchical Structure in Multi-Label Document Classification

Artemis Dampa
Institute of Informatics and Telecommunications, NCSR ”Demokritos”, Greece
Department of Informatics and Telematics, Harokopio University of Athens, Greece
adampa@iit.demokritos.gr

Abstract

Effectively organizing the vast and ever-growing body of research in scientific literature is crucial to advancing the field and supporting scholarly discovery. In this paper, we study the task of fine-grained hierarchical multi-label classification of scholarly articles, using a structured taxonomy. Specifically, we investigate whether incorporating hierarchical information in a classification method can improve performance compared to conventional flat classification approaches. To this end, we suggest and evaluate different strategies for the classification, on three different axes: selection of positive and negative samples; soft-to-hard label mapping; hierarchical post-processing policies that utilize taxonomy-related requirements to update the final labeling. Experiments demonstrate that flat baselines constitute powerful baselines, but the infusion of hierarchical knowledge leads to better recall-focused performance based on use-case requirements.

1 Introduction

The exponential growth of scientific publications has created an urgent need for efficient indexing and organization of academic content. With vast and continuously expanding digital libraries, automatic categorization of scientific articles has become essential to facilitate effective search, discovery, and, ultimately, the acceleration of scientific research (Kim and Gil, 2019). This need is particularly acute in specialized domains, where researchers must navigate an increasingly dense body of literature.

In this work, we focus on the task of fine-grained, hierarchical multi-label classification of scholarly articles, experimenting on the field of Computational Linguistics. In Figure 1, we overview the hierarchical multi-label classification task. Given a document $d \in D$ where D is the set of all possible documents, and a set of labels $L = l_1, l_2, ...$ that have a hierarchical parent-child relation $P : L \times L \rightarrow \{0, 1\}$, where $P(x, y) = 1, x \in L, y \in L$ indicates that x is a parent of y, the task is to find a function $C : D \rightarrow \mathbb{S}(L)$ where $\mathbb{S}(\cdot)$ is the power-set operator, such that given a set of correct (but possibly a priori unknown) annotations $G : D \rightarrow \mathbb{S}(L), C(x) = G(x), \forall x \in D$.

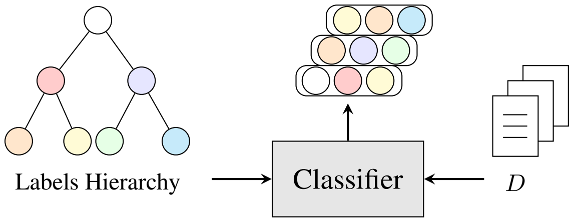


Figure 1: An overview of the hierarchical, multi-label classification task.

Our classification experimental setup uses a predefined taxonomy comprising a multitude of topics and subtopics (181 in total) (Ahmad et al., 2024a), offering a detailed and structured representation of research areas (see also Section 4). This setting poses unique challenges, elaborated on in Section 3, due to idiosyncrasies related to the assignment of a varying number of labels from each level of the hierarchy to a single document.

Historically, the scientific community has approached hierarchical classification using two broad strategies: flat classification (Barbedo and Lopes, 2006; Sun et al., 2003), where the hierarchical structure is ignored and each label is treated independently, or hierarchical classification (Zangari et al., 2024), where models exploit the parent-child relationships among labels to guide predictions. Although flat approaches simplify the problem and often yield strong baselines, they discard potentially valuable structural information. In contrast, hierarchical approaches preserve these relationships, offering a more semantically coherent labeling, but they are often sensitive to errors made at higher lev-

Proceedings of the Student Research Workshop associated with RANLP-2025, pages 10–19, Varna, Bulgaria, Sep 8-10, 2025.
https://doi.org/10.26615/issn.2603-2821.2025_002

els of the hierarchy, as such mistakes can propagate downward and lead to incorrect final predictions.

This paper presents a systematic study comparing flat and hierarchical (cascade-based) classification approaches in the context of scholarly document classification. Thus, we investigate whether exploiting hierarchical information leads to performance gains over flat baselines. Specifically, our contributions focus on three main axes:

Hierarchical Sampling We evaluate methods that enforce the hierarchical structure of the taxonomy by employing node-specific classifiers with hierarchy-aware negative sampling to respect the hierarchy during training.

Soft-to-Hard Label Mapping We explore heuristics to determine the optimal number of labels per document, based on taxonomy structure and empirical distribution. These heuristics include traditional threshold-based methods, fixed-number (top-k) strategies, and more recent LLM-based approaches that utilize generative models to infer the most contextually appropriate set of labels.

Hierarchy-enforcing Post-processing Policy We examine different approaches that ensure hierarchical consistency by altering predicted labels according to hierarchical constraints (assigning parent labels of predicted child nodes or removing child labels if their parents are not predicted).

To support our findings, we conduct statistical analyses that assess the significance of performance differences across multiple metrics. Our experiments reveal insights into the trade-offs between flat and hierarchical approaches and offer practical guidelines for choosing an appropriate strategy depending on the task constraints.

2 Related Work

The task of hierarchical multi-label text classification has seen significant progress through various approaches, each tackling challenges related to large-scale classification, label dependencies, and hierarchical structures.

In the work of Ahmad et al. (2024b), the authors introduce a hierarchical multi-label classification task in the field of computational linguistics. In this task, the authors offer a granular categorization approach based on the taxonomy provided in Ahmad et al. (2024a). The latter also offers a corpus of scholarly articles annotated with topics and subtopics drawn from a structured hierarchy of key NLP areas.

Several approaches have been proposed to handle multi-label text classification. Rajendram Bashyam and Krestel (2024) address hierarchical multi-label classification as extreme multi-label (XMC) flat classification problem, using an X-transformer designed for XMC (Zhang et al., 2021) and TF-IDF-based weak labeling, imposing hierarchy only post-prediction. Liu et al. (2017) introduce XML-CNN, a deep learning model designed for XMC. It enhances document representation using dynamic max pooling, binary cross-entropy loss, and a bottleneck layer to reduce model size. Another work (Hristov et al., 2021) also tackles clinical text classification as an extreme multi-label classification problem, using clustering and cluster-label mapping. S-GCN (Zeng et al., 2024) models multi-label text classification using a global graph based on words, texts, and labels co-occurrence, combining semantic encoding with graph convolution.

In hierarchical classification, Huang et al. (2019) propose a model that classifies documents at multiple levels by integrating text and hierarchy using a Hierarchical Attention-based Recurrent Layer. Similarly, Xu et al. (2021) employ a graph convolutional network (GCN) to learn associations between words, categories, and their relationships, incorporating correlations between levels. Tanigaki et al. (2024) introduce an integrated neural network with cascading self-attention mechanisms, where multi-head attention reconstructs text features at each level while a secondary network enforces inter-level dependencies. TELEClass (Zhang et al., 2025) tackles hierarchical text classification with minimal supervision by enriching the label taxonomy with the use of LLMs. Kosmpoulos et al. (2014) extend cascade classification for predicting the correct leaf of hierarchical structures by estimating the probability of each root-to-leaf path.

Although these works have made significant strides, they share common limitations. Flat classification methods often ignore the hierarchical relationships between labels, while cascade methods are prone to early misclassification. Additionally, many approaches assume a fixed number of labels per level, which does not capture the variability of label counts that can occur at different levels of the hierarchy. Our work aims to shed light on how to address these issues by exploring the effectiveness of hierarchical versus flat approaches in overcoming these challenges.

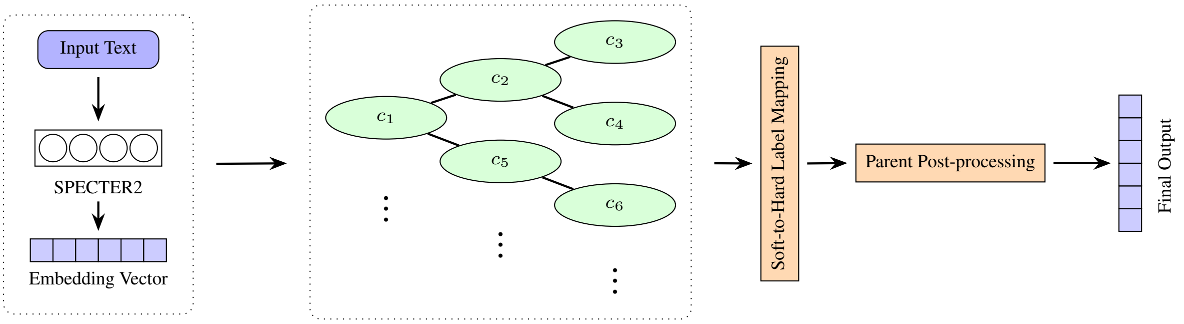


Figure 2: Diagram of the hierarchical multi-label classification process. The figure illustrates the stages of document representation, node-specific classifiers c_i training, soft-to-hard label mapping, hierarchy-enforcing post-processing, and generation of final output.

3 Methodology

In this section, we describe different methods to approach hierarchical multi-label text classification. However, unlike typical hierarchical classification (Sun and Lim, 2001), this task (a) allows the assignment of multiple labels per document; (b) labels can appear at any level of the hierarchy; (c) incomplete paths are allowed, i.e. there is no requirement for labeled documents to have leaf-only labels.

To address these challenges and effectively capture the nuanced structure of scientific discourse while respecting hierarchical label dependencies, our hierarchical approach combines pretrained document embeddings, node-specific classifiers trained using hierarchical sampling, label decoding strategies (soft-to-hard label mapping) and hierarchy-enforcing policies (see Figure 2) and is compared against its flat counterpart.

3.1 Document Representation

To obtain semantically rich document embeddings, we utilize SPECTER2-base (Singh et al., 2023), a pretrained transformer model designed for scientific documents. For each document, we concatenate the title, abstract, and selected metadata fields (author, year, venue, publisher and booktitle) as input to enhance representation. The input is preprocessed using the SPECTER2 tokenizer, with truncation and padding applied to ensure fixed-length. The resulting representation is derived from the model output layer, which captures a high-level summary of the document semantics.

3.2 Cascade Classification with Hierarchical Sampling and Flat Counterpart

Rather than training one multi-output flat classifier which ignores any hierarchical relationships between labels, we split the problem into multiple binary classification tasks, following a cascading approach inspired by Kosmpoulos et al. (2014), adapted for multi-label classification and multi-level label prediction, i.e. including internal nodes within a hierarchical label tree. For each category node c_i in the hierarchy, we train a dedicated logistic regression (LR) classifier to predict whether a document belongs to that category. All classifiers are trained independently. We choose LR due to its efficiency and ease of probabilistic interpretation.

A central challenge is to ensure that classifiers can distinguish semantically similar categories rather than simply separating positive examples from all negatives. To address this, we apply a hierarchy-aware sampling approach per classifier:
Positive samples Documents explicitly labeled with that node are selected as positives.
Negative samples Improve training effectiveness and respect the hierarchical structure as:
(a) **Sibling nodes:** documents labeled with sibling categories, that is, categories that share the same parent as the target node.
(b) **Parent-exclusive samples:** documents labeled with the parent category but not with the current node or any of its siblings.
(c) In cases where a node has no siblings, **siblings of the parent node** are used to maintain informative negative sampling, that is, documents associated with the siblings of its parent node.

The idea behind this design is to encourage the classifiers to focus on subtle inter-category distinctions, thereby enhancing their ability to capture fine-grained differences between closely related topics. By assigning to each classifier the task of distinguishing among a smaller set of categories, the approach also reduces computational resources required and overall classification complexity.

Algorithm 1: Training Hierarchical Multi-Label Classifiers

```
Input: Set of documents $D = \{d_1, ..., d_N\}$,
       Pretrained model $M$ (SPECTER2),
       Hierarchical taxonomy $H$, Logistic
       Regression $LR$ classifier
Output: Trained classifiers
        $C = \{c_1, ..., c_n\}$ for each node $n$
        in $H$
for each node $n$ in $H$ do
    $S_{pos}^{(n)} \leftarrow \{d \mid d \in n\}$
    $S_{sib}^{(n)} \leftarrow \{d \mid d \in siblings(n)\ , d \notin n\}$
    $S_{par}^{(n)} \leftarrow \{d \mid d \in parent(n)\ , d \notin n\}$
    $S_{neg}^{(n)} \leftarrow S_{sib}^{(n)} \cup S_{par}^{(n)}$
    if $S_{neg}^{(n)} = \emptyset$ then
        $S_{neg}^{(n)} \leftarrow \{d \mid d \in$
          $siblings(parent(n))\ , d \notin n\}$
    for $d$ in $S_{pos}^{(n)} \cup S_{neg}^{(n)}$ do
        $X_d \leftarrow M(d)$
        $y_d \leftarrow 1$ if $d \in S_{pos}^{(n)}$, $0$ if $d \in S_{neg}^{(n)}$
    end
    $c_n = LR(X, y)$
end
```

For the flat counterpart, we follow the same overall training strategy but omit the hierarchy-aware negative sampling, instead using the standard approach in which all samples not belonging to the target label serve as negatives for each classifier.

3.3 Soft-to-Hard Label Mapping

Each node classifier outputs a soft score in the range $[0, 1]$, indicating the model confidence that a document belongs to the corresponding category. To convert these scores into final hard label predictions, we propose three decoding *strategies*:

- **Threshold strategy:** A fixed confidence threshold $\theta \in [0, 1]$ is applied. Labels with scores above θ are selected.

- **Label number strategy:** A predefined number k of top-scoring labels is assigned per document.

- **LLM strategy:** A large language model ranks labels, optionally using predictions from the above strategies as priors.

We further discuss the selection of appropriate parameters in Section 4.

3.4 Hierarchy-enforcing Post-processing Policy

All strategies can include an additional hierarchy-enforcing step (hereafter referred to as *parent policy*) to guarantee valid hierarchical paths and adhere to the logical structure of topical taxonomies:

No-parents policy No post-processing is applied and predicted labels are left intact.

With-parents policy For a predicted label at any level, its ancestors are recursively included in the final label set (if not already present) to satisfy hierarchy constraints.

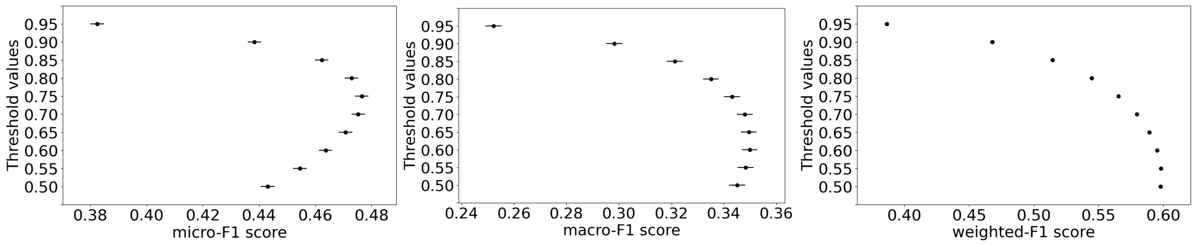
Strict policy A stricter approach keeps predicted labels only if all their parent labels are also predicted. This ensures more infusion of hierarchical structure but potentially introduces early misclassification errors.

Moderate policy A more moderate approach keeps labels if at least one of their parent labels are predicted, trying to balance flexibility and structural consistency.

4 Experiments

We evaluate our hierarchical multi-label classification approach on a corpus of approximately 42,000 scholarly articles from the ACL Anthology (Ahmad et al., 2024a; Rajendram Bashyam and Krestel, 2024) including title, abstract and various metadata such as authors, time of publication, publisher, book, venue. More specifically, our classifiers are trained on a joint set of 1,050 fully labeled documents from the collection and 41,107 weakly labeled documents, while 255 documents are reserved for additional testing. The classification task involves assigning each document to one or more relevant topics from a tree-structured taxonomy of 181 categories, organized across three levels. The train-test split of the dataset follows previous related work (Ahmad et al., 2024a; Rajendram Bashyam and Krestel, 2024) to offer comparable results.

We conduct experiments using 10-fold cross-validation (Sechidis et al., 2011) over the training data subset, with iterative stratification to ensure robust and representative evaluation under label imbalance and sparsity. This method extends traditional stratified sampling to multi-label data by ensuring that the distribution of labels is preserved



(a) Micro-F1 Scores

(b) Macro-F1 Scores

(c) Weighted-F1 Scores

Figure 3: F1 scores of hierarchical approach with threshold strategy for varying θ values between 0.5 and 0.95 applying no-parent policy analyzed using Tukey's HSD test for (a) micro-, (b) macro-, (c) weighted-F1

across folds, improving the fairness and consistency of training and evaluation splits. Model performance is also evaluated on a fixed test set of 255 scholarly documents, to demonstrate generalizability. For each node in the taxonomy, we train a binary classifier using LR (with the default parameters and a maximum of 1,000 iterations). Our training logic incorporates hierarchy-aware negatives, as detailed in Algorithm 1.

This experimental setup aims to answer the following research questions:

RQ1: Which proposed methods or parameter settings outperform the baselines and alternatives?

RQ2: How robust is each method with respect to its hyper-parameters?

RQ3: Can we pre-determine suitable hyper-parameters or develop heuristics to guide their selection?

RQ4: How can one encode hierarchical information in the learning process? Can this encoding improve the classification performance?

RQ5: How does the choice of hierarchical sampling impact model performance?

RQ6: What is the impact of different document representations on classification performance?

4.1 Baselines and Comparison

To benchmark the hierarchical approach, we compare against the following baselines:

- A SciNCL (Ostendorff et al., 2022) model fine-tuned on the flattened labels of the 1,050 labeled documents, which ignores the hierarchy, as provided by (Ahmad et al., 2024b).
- A dummy classifier, which selects labels randomly but preserving label frequency patterns. This serves as a weak lower-bound baseline.
- A flat approach employing a one-vs-all strategy, where a separate classifier is trained for each label using the same training dataset as the hierarchical model as described in Section 3.2.

These comparisons help establish the hierarchy-aware design performance relative to the other approaches (RQ1 - best method), thus evaluating how encoding hierarchical information affects classification performance (RQ4 - hierarchy infusion).

4.2 Label Selection Strategies

We explore the effect and performance of the three approaches described in Section 3.3 to convert classifier outputs into hard label predictions:

- Threshold strategy: Initially, we set the threshold $\theta = 0.6$ based on preliminary tests shown in Figure 3 without applying any parent policy and select all predicted labels with probabilities above θ. We later confirm this value as optimal through exhaustive search.
- Label number strategy: We analyze the label count distribution in the training set and set the number of labels k per document to 5, corresponding to both the mean and median of the distribution. This selection is further validated through an exhaustive search.
- LLM strategy: We use LLaMA 3.1 to validate label predictions based on content.

All strategies are tested across the different parent policies (RQ4 - hierarchy infusion). We vary our hyper-parameters, measuring impact on performance to assess sensitivity and validate RQ2 (hyper-parameters robustness).

Additionally, we conduct an oracle experiment on the validation: we assume knowledge of the true label count per sample and select the top-k predictions accordingly. This informs the feasibility of learning a meta-classifier to estimate label count per document (RQ3 - hyper-parameters heuristics).

Method	Micro			Macro			Weighted		
	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1
Dummy	0.074	0.293	0.118	0.037	0.147	0.048	0.162	0.293	0.186
Flat SciNCL	0.356	0.328	0.341	0.016	0.046	0.024	–	–	–
Flat LR	**0.803**	0.601	**0.687**	**0.625**	0.392	**0.467**	**0.790**	0.601	**0.673**
Label ($k = 11$ & no-parents)	0.370	0.604	0.459	0.386	0.456	0.349	0.648	0.604	0.584
Label ($k = 7$ & with-parents)	0.302	0.552	0.391	0.352	0.383	0.297	0.509	0.552	0.479
Label ($k = 20$ & strict)	0.668	0.679	0.673	0.521	0.413	0.446	0.665	0.679	0.664
Label ($k = 20$ & moderate)	0.441	0.296	0.354	0.339	0.311	0.291	0.261	0.296	0.263
Threshold ($\theta = 0.6$ & no-parents)	0.368	0.628	0.464	0.375	0.473	0.350	0.647	0.628	0.595
Threshold ($\theta = 0.8$ & with-parents)	0.309	0.549	0.396	0.346	0.387	0.298	0.520	0.549	0.485
Threshold ($\theta = 0.5$ & strict)	0.784	0.588	0.672	0.560	0.348	0.420	0.767	0.588	0.653
Threshold ($\theta = 0.5$ & moderate)	0.582	0.249	0.349	0.400	0.261	0.279	0.315	0.249	0.261
LLM (Label $k = 20$ & with-parents)	0.606	0.520	0.560	0.468	0.381	0.393	0.646	0.520	0.552
Oracle Top-k (no-parents)	0.502	0.502	0.502	0.485	0.381	0.363	0.765	0.502	0.570
No strategy (no-parents)	0.327	**0.686**	0.443	0.343	**0.518**	0.345	0.602	**0.686**	0.598

Table 1: Evaluation of baseline and hierarchical methods, using micro, macro, and weighted precision, recall, and F1 score across 10-fold cross-validation. Best results per column are in bold, while second-best are underlined.

4.3 Ablation Studies

We perform analyses to identify factors influencing model performance. We examine the impact of negative sampling within our hierarchical framework and assess how the inclusion of metadata in document representation affects classification accuracy. These studies address RQ5 (negative sampling) and RQ6 (document representation) and provide insights to refine our approach and enhance model effectiveness (see Sections 5.3, 5.4).

4.4 Statistical Analysis and Evaluation

We assess model performance using the following metrics (Yang, 1999): (a) micro precision/recall/F1, which measure global performance, favoring frequent classes; (b) macro precision/recall/F1, which give equal weight to all classes, highlighting rare-category performance; (c) weighted precision/recall/F1, which weight the contribution of each label by its support.

For each configuration, results are aggregated over the cross-validation folds. To determine the statistical significance of differences between methods and parameter choices (RQ1 - best method), we perform Tukey's HSD test and report letter groupings to identify significantly different clusters.

Together, these experiments allow us to systematically address our research questions by comparing classifiers and selection strategies (RQ1 - best method), evaluating sensitivity to key parameters (RQ2 - hyper-parameters robustness), exploring document-specific heuristics for label prediction (RQ3 - hyper-parameters heuristics), and assessing the role of hierarchical information (RQ4 - hierarchy infusion), hierarchical sampling (RQ5 - negative sampling) and metadata enriched input (RQ6 - document presentation) in improving performance.

Method	micro-F1	macro-F1	weighted-F1
Label(NS)	0.673	0.446	0.664
Threshold(NS)	0.672	0.420	0.653
Label(RS)	0.590	0.422	0.560
Threshold(RS)	0.673	0.420	0.652

Table 2: Comparison of best-performing label selection strategies (strict $k = 20$, $\theta = 0.5$) with hierarchy-aware negative sampling (NS) against hierarchy-aware negative sampling enhanced with random sampling (RS).

5 Results

We present the performance of our hierarchical classification models using different label selection strategies and parent policies and compare them against flat and dummy baselines. Results are reported as average scores across 10-fold cross-validation in all tables, except Table 4, using micro, macro, and weighted precision, recall, and F1. All tables report statistically significant differences, validated using Tukey's HSD test (with $\alpha = 0.05$).

5.1 Overall Performance

Table 1 shows that the hierarchical method outperforms the flat SciNCL and dummy classifier across all metrics, but does not surpass the flat LR one-vs-all method in terms of precision and F1 scores. This provides an answer to RQ1 (best method), indicating that incorporating node-level classifiers does not always yield a performance advantage.

5.2 Strategies & Hyper-parameters

To explore RQ2 (hyper-parameters robustness), we varied hyper-parameters and observed that performance varied accordingly. Specifically, we performed an exhaustive search over different label counts (1-25) and threshold values (0.5-0.95) to identify the optimal values per strategy and policy.

The analysis on the label count showed that performance improved with an increasing number of labels, up to an optimal point beyond which gains began to diminish. This finding contradicts the initial belief that optimal performance would align with the mean or median of the label count distribution. Furthermore, variations in threshold values affected performance, with lower thresholds generally resulting in improved results. This outcome is expected, as higher confidence thresholds reduce the number of predicted labels, consequently leading to similar overall performance between the two strategies. This suggests that the method is sensitive to these hyper-parameters, with optimal performance achieved under specific conditions.

Based on this hyper-parameter tuning, we report the results for the best parameter values and policies in Table 1. The label number strategy with $k = 20$ and strict parents performed best overall. The threshold strategy with $\theta = 0.5$ and strict parents had slightly lower recall but higher precision, which can be advantageous in applications where minimizing false positives is critical. The LLM-based label selection strategy showed modest improvements for models with suboptimal hyper-parameter settings, but it significantly lagged behind the top-performing strategies. The setting without any strategy and parent policy improved recall but suffered from over-selection, leading to moderate F1. Across different strategies, enforcing parent policies most of the time boosted micro, macro and weighted F1 scores by at least 6% (up to 28%), 1% (up to 12%) and 6% (up to 9%), respectively, compared to their counterparts with no-parent policy, confirming the importance of structural consistency (RQ4 - hierarchy infusion).

To address RQ3 (hyper-parameters heuristics), we implemented an oracle strategy during validation that uses the true number of labels per document to select the top-k predictions. This approach achieved 36.31% macro-F1 which is over 8% lower than the best-performing hierarchical method. These results suggest that a meta-classifier for estimating label cardinality alone is not sufficient, and that label ranking combined with hierarchical structure infusion through parent policies plays a more critical role in achieving high performance. A heuristic based on the standard threshold of 0.5, performs competitively with our best hierarchical approach, supporting the idea that simple statistics can inform effective parameter choices when combined with hierarchical information. However, directly setting a predefined number of labels, which yields the best results within our hierarchical framework, can work best in combination with the parent policy that dynamically reduces the number of predicted values (i.e., by removing orphan child label predictions). As a result, the final label count per document varies, even though the initial number was fixed.

Method	micro-F1	macro-F1	weighted-F1
Flat simple	0.670	0.456	0.656
Flat enriched	0.687	0.467	0.673
Label simple	0.425	0.311	0.499
Label enriched	0.464	0.334	0.538

Table 3: F1 scores of flat and hierarchical (with $k = 7$ labels and no-parents policy) approaches using only title and abstract inputs, compared to metadata-enriched inputs, obtained through 10-fold cross-validation.

5.3 Hierarchical Sampling Study

Motivated by the relatively high recall that comes at the expense of precision, along with the generally increased number of positive predictions (both true and false), we conducted an additional study on the negative sampling strategy. In our hierarchical sampling approach, the lower levels of the hierarchy include fewer negative samples, which results in a distribution shift between the constructed training data and the original dataset.

To evaluate whether this imbalance affects model performance, we doubled the number of negative samples per classifier by randomly adding half of the samples from the full negative space. This adjustment was intended to test whether the initial assumption that the hierarchical structure would help the model better differentiate between similar documents holds true, or whether it instead introduces confusion, suggesting that the model might benefit more from increased exposure to diverse negative examples. The results of this study shown in Table 2, indicate that the best-performing strategies do not benefit from enhanced negative sampling, suggesting that the hierarchical frame-

Method	**micro-F1**	**CLD**	**macro-F1**	**CLD**	**weighted-F1**	**CLD**
Flat	0.669	A	0.374	A	0.676	A
Label ($k = 20$ with strict policy)	0.637	C	0.366	A	0.658	B
Threshold ($\theta = 0.5$ with strict policy)	0.657	B	0.349	B	0.661	B

Table 4: Comparison of the top-3 models on the test set in terms of micro, macro, and weighted F1 scores. Tukey's HSD significance test results: models sharing the same group letter are not significantly different at $\alpha = 0.05$. The CLD column (Compact Letter Display) shows the group letters assigned to each model.

work provides sufficient discriminative context.

5.4 Document Representation Study

To investigate the impact of input document representation on classification performance, we conducted a focused study comparing two alternative representations. Specifically, we aimed to assess whether including metadata fields enhances performance or whether a simpler representation suffices. To ensure a fair comparison, we kept all other components such as algorithm, label strategy, parent policy, and hyper-parameter values, constant, and varied only the document input.

The main representation used throughout this paper combines the title, abstract, and key metadata fields: author, year, venue, publisher, and booktitle. For comparison, we created a simplified version consisting of only the title and abstract concatenated. The results, presented in Table 3, clearly show that the metadata-enriched representation outperforms the simpler alternative on both flat and hierarchical approaches, confirming the value of incorporating contextual metadata in improving classification performance.

5.5 Statistical Significance

We applied Tukey's HSD post-hoc test to all configurations. Results in Table 4 are based on the 255 documents held out as the test set from the used corpus (Ahmad et al., 2024a). They indicate that the flat model forms a statistically superior group compared to the top-performing hierarchical models in terms of micro and weighted F1 scores. However, for macro F1, flat model belongs to the same significance group as the hierarchical with label count $k = 20$ and strict-parents policy.

6 Conclusion

In this work, we investigated the task of fine-grained hierarchical multi-label classification of scholarly articles, using a predefined taxonomy. We conducted a systematic comparison between flat classification methods and hierarchy-aware approaches, including cascade models with hierarchy-aware negative sampling and parent-enforcing post-processing. To this end, we utilized an existing corpus from NLP scholarly articles (ACL collection).

Our results demonstrate that the hierarchical approach outperforms the flat baseline in terms of recall but falls behind in precision and overall F1 score. While explicitly modeling the hierarchy adds complexity, enforcing hierarchy through the proposed parent policies generally improves performance compared to ignoring hierarchical structure.

Statistical analyses confirm that the observed differences are significant across most metrics, showing that hierarchy-aware strategies can help reduce false negatives. However, on the final test set, the hierarchical and flat approaches do not differ significantly in macro F1, suggesting that the hierarchical approach remains competitive when aiming for balanced performance.

Our study also demonstrated that the selection of a policy for the infusion of hierarchical information into classification significantly affects the result. Although the results we achieved with the most promising infusion policy were not sufficiently better from the flat approach, we argue that it is important to examine other approaches for this infusion.

Thus, as a future direction, more effective ways to represent and integrate hierarchical information should be explored. Motivated by the observed boost from metadata-enriched representations, incorporating knowledge-informed features may enhance the ability of the model to leverage hierarchy without relying solely on rigid label dependencies.

Acknowledgments

This work was supported by the European Union's Digital Europe Programme (DIGITAL) under grant agreement No 101146490. I would like to thank my supervisors G. Giannakopoulos, I. Varlamis and V. Karkaletsis for their valuable guidance throughout this work.

References

Raia Abu Ahmad, Ekaterina Borisova, and Georg Rehm. 2024a. Forc4cl: A fine-grained field of research classification and annotated dataset of nlp articles. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, page 7389–7394, Torino, Italia. ELRA and ICCL.

Raia Abu Ahmad, Ekaterina Borisova, and Georg Rehm. 2024b. Forc@nslp2024: Overview and insights from the field of research classification shared task. In *Natural Scientific Language Processing and Research Knowledge Graphs*, page 189–204, Cham. Springer Nature Switzerland.

Jayme Garcia sArnal Barbedo and Amauri Lopes. 2006. Automatic genre classification of musical signals. *EURASIP Journal on Advances in Signal Processing*, 2007(1):1–12.

Anton Hristov, Aleksandar Tahchiev, Hristo Papazov, Nikola Tulechki, Todor Primov, and Svetla Boytcheva. 2021. Application of deep learning methods to snomed ct encoding of clinical texts: From data collection to extreme multi-label text-based classification. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, page 557–565, Held Online. INCOMA Ltd.

Wei Huang, Enhong Chen, Qi Liu, Yuying Chen, Zai Huang, Yang Liu, Zhou Zhao, Dan Zhang, and Shijin Wang. 2019. Hierarchical multi-label text classification: An attention-based recurrent network approach. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, page 1051–1060, Beijing China. ACM.

Sang-Woon Kim and Joon-Min Gil. 2019. Research paper classification systems based on tf-idf and lda schemes. *Human-centric Computing and Information Sciences*, 9(1):30.

Aris Kosmpoulos, Georgios Paliouras, and Ion Androutsopoulos. 2014. The effect of dimensionality reduction on large scale hierarchical classification. In *Information Access Evaluation. Multilinguality, Multimodality, and Interaction*, pages 160–171, Cham. Springer International Publishing.

Jingzhou Liu, Wei-Cheng Chang, Yuexin Wu, and Yiming Yang. 2017. Deep learning for extreme multi-label text classification. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 115–124, Shinjuku Tokyo Japan. ACM.

Malte Ostendorff, Nils Rethmeier, Isabelle Augenstein, Bela Gipp, and Georg Rehm. 2022. Neighborhood contrastive learning for scientific document representations with citation embeddings. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, page 11670–11688, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Lakshmi Rajendram Bashyam and Ralf Krestel. 2024. Advancing automatic subject indexing: Combining weak supervision with extreme multi-label classification. In *Natural Scientific Language Processing and Research Knowledge Graphs*, page 214–223, Cham. Springer Nature Switzerland.

Konstantinos Sechidis, Grigorios Tsoumakas, and Ioannis Vlahavas. 2011. On the stratification of multi-label data. In *Machine Learning and Knowledge Discovery in Databases*, page 145–158, Berlin, Heidelberg. Springer.

Amanpreet Singh, Mike D'Arcy, Arman Cohan, Doug Downey, and Sergey Feldman. 2023. Scirepeval: A multi-format benchmark for scientific document representations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, page 5548–5566, Singapore. Association for Computational Linguistics.

Aixin Sun and Ee-Peng Lim. 2001. Hierarchical text classification and evaluation. In *Proceedings 2001 IEEE International Conference on Data Mining*, page 521–528.

Aixin Sun, Ee-Peng Lim, and Wee-Keong Ng. 2003. *Hierarchical Text Classification Methods and Their Specification*, page 236–256. Springer US, Boston, MA.

Koichi Tanigaki, Koji Cho, and Shuichi Tokumoto. 2024. Cascading taxonomic attention networks for hierarchical text classification. In *2024 16th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI)*, page 369–372.

Linli Xu, Sijie Teng, Ruoyu Zhao, Junliang Guo, Chi Xiao, Deqiang Jiang, and Bo Ren. 2021. Hierarchical multi-label text classification with horizontal and vertical category correlations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, page 2459–2468, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yiming Yang. 1999. An evaluation of statistical approaches to text categorization. *Information Retrieval*, 1(1):69–90.

Alessandro Zangari, Matteo Marcuzzo, Matteo Rizzo, Lorenzo Giudice, Andrea Albarelli, and Andrea Gasparetto. 2024. Hierarchical text classification and its foundations: A review of current research. *Electronics*, 13(7).

Delong Zeng, Enze Zha, Jiayi Kuang, and Ying Shen. 2024. Multi-label text classification based on semantic-sensitive graph convolutional network. *Knowledge-Based Systems*, 284:111303.

Jiong Zhang, Wei-Cheng Chang, Hsiang-Fu Yu, and Inderjit Dhillon. 2021. Fast multi-resolution transformer fine-tuning for extreme multi-label text classification. In *Advances in Neural Information Processing Systems*, volume 34, page 7267–7280. Curran Associates, Inc.

Yunyi Zhang, Ruozhen Yang, Xueqiang Xu, Rui Li, Jinfeng Xiao, Jiaming Shen, and Jiawei Han. 2025. Teleclass: Taxonomy enrichment and llm-enhanced hierarchical text classification with minimal supervision. In *Proceedings of the ACM on Web Conference 2025*, page 2032–2042, Sydney NSW Australia. ACM.