Improving Scientific Research Area Classification through Transformer Fine-Tuning and Per-Document Model Selection

SimpleText Task 3 at CLEF 2026

George Anderson[1,†], Edwin Thuma[1,*,†]

[1]University of Botswana, Plot 4775, Notwane Road, Gaborone, Botswana

Abstract

This paper presents the working notes for the UBCS (University of Botswana Computer Science) team's participation in the CLEF 2026 SimpleText Track, Task 3 (Research Area Classification). We propose a hierarchical classification pipeline that utilizes domain-specific transformer models, specifically SciBERT and SPECTER. To address class imbalance, we incorporate rule-based and T5-driven data augmentation along with oversampling. We implement a hierarchical constraint mechanism, where the prediction of the broad field restricts the candidate disciplines. Furthermore, we explore a Random Forest meta-classifier to dynamically select between SciBERT and SPECTER predictions for each document. Our experiments demonstrate the efficacy of combining hierarchical constraints with domain-adapted language models.

Keywords

Research area classification, Hierarchical classification, Transformers, Data augmentation, Model selection

1. Introduction

The exponential growth of scientific literature has made it increasingly difficult for both researchers and the general public to navigate, retrieve, and comprehend scientific knowledge. Every year, a large number of scientific articles are published in thousands of journals, conferences, and preprint servers, producing an overwhelming corpus of text that spans highly specialized and fragmented domains. In response to this pressing challenge of information overload, the Conference and Labs of the Evaluation Forum (CLEF) introduced the SimpleText Track [1]. This competition is specifically aimed at trying to ensure that complex scientific texts are more accessible, understandable, and properly categorized for a diverse audience. The CLEF 2026 SimpleText Track focuses on several key aspects of scientific text processing, including text simplification, hallucination detection in generated texts, and, importantly, research area classification. This comprehensive report details the experimental setup, theoretical methodology, and empirical results of the UBCS (University of Botswana Computer Science) team's submission for CLEF SimpleText Track 3: Research Area Classification. The main objective of this task is to automatically and accurately classify scientific abstracts and their corresponding titles into a predefined hierarchical taxonomy [2, 3]. The taxonomy provided for this task is structured into two distinct but related tiers: broad research fields, such as Natural Sciences, Engineering Sciences, Humanities, Social Sciences, and Life Sciences, and fine-grained discipline areas, such as Physics, Process Engineering, Archeology, and Computer Science. High-performance for models within such a framework is not merely an academic exercise, but rather very important real-world applications such as dynamic literature recommendation systems, automated peer-review assignment, bibliometric trend analysis, and the construction of intelligent scientific search engines.

CLEF 2026 Working Notes, 21 – 24 September 2026, Jena, Germany

*Corresponding author.

†These authors contributed equally.

andersong@ub.ac.bw (G. Anderson); thumae@ub.ac.bw (E. Thuma)

https://www.ub.bw/ (G. Anderson); https://www.ub.bw/ (E. Thuma)

0000-0001-9368-1404 (G. Anderson); 0009-0007-3011-4876 (E. Thuma)

While traditional text classification is generally considered a mature subfield of Natural Language Processing (NLP), scientific research area classification presents several unique and formidable challenges. First, scientific language is inherently highly technical, dense, and specialized. General-purpose language models, which are typically trained on everyday text like news articles or Wikipedia, often fail to capture the nuanced semantics and complex syntactic structures of domain-specific terminology. Second, the provided taxonomy is heavily and naturally imbalanced, reflecting the real-world distribution of available digital corpora. In the dataset provided for this competition, prominent disciplines such as Physics and Mathematics contain thousands of training examples, whereas less digitized fields such as Process Engineering or Literary Studies contain fewer than five examples. Finally, the hierarchical nature of the labels requires a structured and constrained approach to prevent logical inconsistencies. A standard flat classifier might predict 'Life Sciences' as the broad field but erroneously predict 'Computer Science' as the discipline, breaking the ontological rules of the taxonomy.

Recent evidence suggests that transformer-based large language models (LLMs) are well suited to classifying scientific publications within hierarchical classification systems, particularly when supported by prompt-engineering techniques such as in-context learning and prompt chaining [4, 5]. In this work, we describe a robust pipeline relying on transformer architectures pre-trained on scientific texts (SciBERT [6] and SPECTER [7]). We formulated the problem as a two-stage hierarchical classification task and utilized advanced handling of rare classes via paraphrasing-based augmentation and semantic label-description re-ranking. We also make use of model selection.

To systematically address the intersecting challenges, we propose and evaluate an advanced, multi-stage hierarchical classification pipeline based on the SciBERT and SPECTER transformer architectures. Our methodology departs from standard flat classification by incorporating several novel architectural and procedural elements: (1) a local-classifier-per-parent hierarchical routing system that strictly enforces taxonomic validity, (2) the integration of explicit label descriptions via semantic re-ranking to guide the model on zero-shot and few-shot classes, (3) a robust data augmentation strategy utilizing natural language paraphrasing to artificially synthesize minority-class examples, and (4) per-document model-selection. By combining deep pre-trained contextual embeddings with a structured domain taxonomy and aggressive class balancing, our approach achieves robust performance across both broad fields and fine-grained disciplines.

The remainder of this paper is organized as follows. Section 2 discusses state-of-the-art approaches for similar tasks in CLEF and beyond. Section 3 explains our methodology. Section 4 discusses our experiments and results and Section 5 has our conclusion.

2. Related Work

Automatic classification of scientific articles into research areas has been widely studied, with growing focus on hierarchical label spaces, domain-specific transformers, and low-resource regimes. The CLEF SimpleText Task 3 builds on this line of work, requiring robust handling of deep taxonomies, rare labels, and limited supervision.

2.1. Hierarchical Text Classification and Taxonomies

Traditional HTC methods distinguish local approaches (a classifier per node/parent) from global models over the entire hierarchy [8, 9]. Local-classifier-per-parent-node (LCPN) CNNs have been shown to outperform flat baselines by 5–15% on multi-level news hierarchies [8]. Recent work refines taxonomies themselves with LLMs to better align label structures with model inductive biases, improving F1 by up to 2.9 points [10], and explores minimally supervised HTC using enriched taxonomies and LLM-guided annotation/generation [11]. Zero-shot hierarchical mapping methods exploit pretrained language models to assign documents to custom taxonomies without labeled data, reinforcing scores via upward propagation in the hierarchy [12].

2.2. Transformer Models for Scientific Document Representation

Domain-specific transformers such as SciBERT and SPECTER provide strong document-level embeddings for scientific classification tasks. SPECTER leverages citation graphs during pretraining and outperforms competitive baselines on document-level tasks including classification and recommendation [13]. Comparative studies show SciBERT and SPECTER achieve high accuracy and macro-F1 on scientific subject prediction [14], and specialized scientific BERT variants and small transformer models can surpass CNN/RNN baselines for literature classification while remaining efficient [15, 16, 17].

Generally, as discussed in the literature cited above, SciBERT focuses on scientific text, while SPECTER focuses on text with citation graphs. SciBERT is used in fine-tuned classifiers and hybrids, while SPECTER is used for document embeddings and other downstream tasks [13, 14, 17].

2.3. Zero-/Few-Shot Classification and Label Semantics

Zero-shot text classification methods exploit label descriptions and entailment-style prompting. Integrating semantic knowledge such as class descriptions, hierarchies, and knowledge graphs improves the recognition of unseen classes [18]. Training on curated label-description datasets can increase zero-shot accuracy by 17–19 percentage points over vanilla zero-shot and generalize across domains [19]. In addition, enhancing transformer representations with label semantics improves few- and zero-shot performance, even for short texts in low-resource languages [20]. Benchmarks show modern rerankers and strong embedding models excel at label-description–based zero-shot classification across many datasets [21]. For large structured label spaces, few- and zero-shot methods that exploit label structure yield substantial gains on rare and unseen labels [22].

2.4. Data Augmentation and Imbalanced Classification

Imbalanced text classification is often addressed through data augmentation. Some of the techniques that are normally used in data augmentation included transformation, paraphrasing, and generation methods, with paraphrasing-based approaches, for example back-translation, pretrained sequence-to-sequence models, improving performance, especially when augmenting minority classes and reduced datasets [23]. Studies in other languages confirm that paraphrasing with fine-tuned generative models can raise minority-class F1, although care is needed to avoid overfitting and manage computational cost [24]. More broadly, systematic evaluations show that combining data augmentation with ensemble methods can substantially improve performance on class-imbalanced problems; in many cases, simpler oversampling methods perform competitively and are cheaper than GAN-based augmentation [25, 26].

2.5. Summary

Prior work motivates (i) hierarchy-aware architectures such as local classifiers and taxonomy-aware zero-shot mapping for HTC, (ii) domain-specific transformers such as SciBERT and SPECTER for scientific document representation, (iii) the use of explicit label descriptions and semantic reranking to handle zero-/few-shot classes, and (iv) paraphrasing-based augmentation and ensemble strategies to mitigate class imbalance. This body of research underpins the design choices for hierarchical routing, label-description-guided scoring, paraphrase based augmentation, and per-document model selection in our CLEF SimpleText Task 3 runs.

3. Methodology

3.1. Data Preprocessing and Corpus Normalization

The integrity of deep learning models in natural language processing is inextricably linked to the quality of the input data. For CLEF SimpleText Track 3, the training and testing datasets were provided in comma-separated values (CSV) format, consisting of document identifiers, titles, abstracts, and the

target hierarchical labels: field_area and discipline_area. Given the heterogeneous nature of scientific metadata, a rigorous, automated preprocessing pipeline was constructed to sanitize, normalize, and unify the textual inputs prior to tokenization.

The first stage of the preprocessing pipeline involves structural normalization. We implemented robust parsing mechanisms capable of handling diverse text encodings (utf-8-sig) and irregular delimiters. Column headers were systematically normalized by stripping whitespace, casting to lowercase, and replacing non-alphanumeric characters with underscores. Missing values (NaNs) in the textual features were interpolated with empty strings to prevent programmatic failures during tensor compilation.

The primary textual feature used for inference is a concatenation of the document's title and abstract. In scholarly literature, titles often provide a highly condensed, high-signal representation of the paper's core domain, whereas the abstract provides necessary context and methodology. To emphasize the title's semantic importance, we applied a 'repeat-title' heuristic during the concatenation phase. Specifically, the title was appended twice, followed by the abstract, separated by periods and whitespace: "Title. Title. Abstract.". This artificial up-weighting ensures that the self-attention mechanisms within the downstream transformer models allocate sufficient probability mass to the title tokens, which often contain explicit domain markers.

Subsequent text cleaning involved aggressive whitespace normalization, utilizing regular expressions to collapse multiple contiguous spaces, tabs, and newline characters into a single space. This minimizes token fragmentation and maximizes the number of meaningful word pieces that fit within the strict maximum sequence lengths required by BERT-based architectures.

Following text normalization, the categorical targets (field_area and discipline_area) were encoded into dense integer representations using scikit-learn's LabelEncoder. We generated dual mappings (id-to-label and label-to-id) to seamlessly translate between the raw taxonomic strings and the necessary tensor indices during the loss computation and evaluation phases. We further constructed a programmatic representation of the class hierarchy, building a directed acyclic graph (DAG) mapping each field area to its constituent discipline areas based on their co-occurrence in the training set and the external classification_taxonomy.csv file.

3.2. Addressing Extreme Class Imbalance: Generative Paraphrasing and Oversampling

One of the most profound obstacles in the CLEF 2026 SimpleText Task 3 dataset is extreme inter-class imbalance. The Natural Sciences field, heavily dominated by Physics and Mathematics, constitutes a massive majority of the corpus, whereas Humanities and Social Sciences disciplines suffer from data sparsity. Standard cross-entropy loss optimization under such conditions intrinsically biases the network toward majority classes, yielding poor recall for minority labels. To counteract this, we orchestrated a two-tiered data synthesis and resampling strategy: T5-based generative paraphrase augmentation followed by localized stochastic oversampling.

The first tier (generative augmentation) aims to synthesize novel, lexically diverse training examples for classes containing fewer than 25 samples. To achieve this without distorting the underlying scientific semantics, we leveraged a pre-trained Text-to-Text Transfer Transformer (T5) specifically optimized for paraphrase generation (Vamsi/T5_Paraphrase_Paws). For every document belonging to a minority discipline, the T5 model received the text concatenated with the task prefix "paraphrase:". We executed beam search decoding with 4 beams to generate high-quality, fluent restatements of the original abstracts. By introducing structural and synonymic variations while preserving domain-specific terminology, this generative augmentation acts as a powerful regularizer, expanding the geometric decision boundaries around rare classes.

In addition to the deep-learning-based T5 augmentation, we implemented a rule-based "light" paraphrasing fallback mechanism to accelerate processing where applicable. This lightweight module relies on sentence permutation and deterministic synonym substitution, for example translating "We propose" to "We present", or "This paper" to "This study". While less syntactically diverse than T5, the light augmentation introduces sufficient variance to prevent immediate overfitting on duplicated samples.

The second tier of our imbalance strategy involves fold-aware localized oversampling. Crucially, upsampling must be strictly isolated within the training split of each cross-validation fold to prevent data leakage and artificially inflated validation metrics. During the preparation of each training fold, we grouped the dataset by discipline_area. For any group containing fewer than a dynamically determined threshold (set via hyperparameter optimization to 25 samples), we performed random sampling with replacement until the target threshold was achieved. Furthermore, to prevent the majority classes from dominating the batch gradients, we introduced a clamping parameter (MAX_SAMPLES_PER_DISCIPLINE = 120). Majority classes exceeding this limit were down-sampled. This hybrid SMOTE-like oversampling and targeted down-sampling approach ensures that the transformer models are exposed to a relatively balanced distribution of gradients across the topological loss surface, significantly enhancing macro-averaged F1 scores.

3.3. Two-Stage Hierarchical Classification Architecture

The taxonomic structure of the SimpleText Research Area Classification task dictates that every discipline area strictly belongs to a specific parent field area. Traditional flat classification strategies, which predict the discipline directly from a single dense layer, ignore this fundamental ontological constraint. To embed this inductive bias directly into our architecture, we engineered a two-stage hierarchical inference pipeline.

In the first stage, the transformer model acts as a broad "Field Classifier." The input text sequence is tokenized, mapped to high-dimensional embeddings, and passed through the self-attention blocks. The final hidden state corresponding to the special [CLS] token is extracted and projected through a linear classification head mapping to the 5 primary field areas (Engineering Sciences, Humanities, Life Sciences, Natural Sciences, Social Sciences).

The second stage features a conditionally constrained "Discipline Classifier." Instead of predicting across all 19 discipline areas simultaneously, the prediction space is dynamically masked based on the outcome of the Field Classifier. We programmatically enforce the taxonomy by maintaining a dictionary linking each parent field to its valid child disciplines. During the forward pass, the model generates logits for all 19 disciplines. However, before the argmax operation is applied to extract the final class prediction, we isolate the subset of discipline indices corresponding to the predicted parent field. The logits for all non-child disciplines are artificially suppressed (set to negative infinity). The softmax function is then calculated exclusively over the valid child subset.

This cascaded approach guarantees topological validity; the system cannot logically predict a discipline that violates the hierarchical tree. Furthermore, during the fine-tuning phase, the discipline classifier is explicitly trained only on instances belonging to the correct parent field. We instantiate separate linear classification heads for each parent field. Thus, the weights of the discipline head for "Life Sciences" are updated exclusively using biological and medical texts. This deep decoupling allows the transformer to learn specialized, fine-grained feature representations for intra-field discrimination without catastrophic interference from fundamentally distinct domains.

3.4. Domain-Adapted Transformer Backbones: SciBERT and SPECTER

The linguistic properties of scholarly texts diverge drastically from the generalized web corpora used to pre-train standard architectures like BERT or RoBERTa. Scientific abstracts are dense with complex jargon, mathematical formalisms, and highly specific contextual dependencies. Consequently, our methodology relies entirely on domain-adapted transformers pre-trained specifically on academic literature.

We utilized two distinct backbone architectures: SciBERT and SPECTER. SciBERT [27] is a BERT-based model pre-trained from scratch on a massive corpus of 1.14 million full-text papers from Semantic Scholar, spanning computer science and biomedical domains. Crucially, SciBERT utilizes a specialized WordPiece vocabulary (scivocab) generated directly from scientific text. This specialized vocabulary drastically reduces the out-of-vocabulary (OOV) rate and prevents complex scientific terms from being

aggressively fragmented into meaningless subword tokens, preserving the morphological integrity of the input.

SPECTER (Scientific Paper Embeddings using Citation-informed Transformers), developed by Cohan et al. [10], represents an alternative paradigm. While also based on the SciBERT architecture, SPECTER is further pre-trained on a signal representing the citation graph of scientific papers. By leveraging a triplet loss objective, SPECTER learns to place papers that cite one another closer together in the embedding space. This citation-informed pre-training allows SPECTER to inherently capture structural, field-level relationships that transcend pure lexical overlap.

Our experimental hypothesis postulated that SciBERT and SPECTER possess complementary strengths. SciBERT's scivocab excels at highly granular lexical matching, potentially making it superior for distinguishing closely related disciplines within the same field. Conversely, SPECTER's citation-graph embeddings provide exceptional global context, making it highly robust for broad field area prediction. To maximize performance, our pipeline independently fine-tunes both models across all cross-validation folds. For SciBERT, we constrained the maximum sequence length to 256 tokens to optimize batch throughput. For SPECTER, recognizing its reliance on extensive document context, we expanded the maximum sequence length to 384 tokens. Both models utilized the AdamW optimizer with a learning rate of 2e-5, a linear warmup scheduling ratio of 0.1, and weight decay set to 0.01 to prevent overfitting during the 2-epoch training cycle.

3.5. Semantic Re-ranking via Taxonomy Label Descriptions

Beyond the textual content of the abstracts, the CLEF 2026 task provides semantic metadata via the classification taxonomy. The labels themselves, for example "Civil Engineering" or "Linguistics" contain rich semantic meaning that is often discarded when classes are converted to arbitrary integer indices. To synthesize this semantic metadata, we developed a post-hoc Label Description Re-ranking module.

We first constructed synthetic label descriptions by fusing the taxonomy structure with generative templates. For instance, the discipline "Anthropology" yielded the semantic description: "Anthropology. Fine-grained research discipline under the broad field Humanities. Scientific publications in this discipline discuss theories, methods, experiments, datasets, applications, and terminology associated with Anthropology."

To deploy these descriptions during inference, we utilized a pre-trained Sentence-Transformer model: all-MiniLM-L6-v2 to independently encode both the input document text and the synthesized label descriptions into a shared dense vector space. Following the transformer's generation of predictive probabilities via the softmax function, we computed the cosine similarity between the document's dense embedding and the embeddings of the candidate label descriptions.

This cosine similarity vector acts as an independent, zero-shot signal of taxonomic alignment. We scale this similarity matrix by a hyperparameter blending weight (LABEL_DESCRIPTION_BLEND = 0.05 to 0.20) and linearly aggregate it with the transformer's predicted logits. This re-ranking mechanism effectively nudges the model's predictions. If the transformer is highly uncertain between two closely related disciplines, the semantic overlap with the explicit label description serves as the ultimate tie-breaker. This integration effectively bridges parametric classification with non-parametric, retrieval-augmented semantic matching.

3.6. Ensemble Meta-Learning with Random Forest Selector

Having independently trained two robust, hierarchically constrained, and semantically re-ranked transformers (SciBERT and SPECTER), we encountered the challenge of ensemble integration. Traditional ensembling relies on simple logit averaging or majority voting. However, our validation metrics indicated that the optimal model varied conditionally based on the input document's specific domain and length. Simple averaging would inevitably dilute the confidence of the correct model.

To solve this, we implemented an intelligent Meta-Classifier using the Random Forest algorithm [28]. The Random Forest operates as a dynamic selector, determining whether to output the SciBERT

prediction or the SPECTER prediction on a per-document basis.

Training this meta-classifier required rigorous avoidance of data leakage. We employed a 5-fold cross-validation scheme to generate Out-Of-Fold (OOF) predictions. For each fold, models were trained on $4/5$ths of the data and generated probabilistic predictions on the hold-out $1/5$th. Over the complete cycle, every document in the training set received an unbiased prediction from both SciBERT and SPECTER.

For each document, we extracted a complex feature vector for the Random Forest. These features included: the maximum softmax probability (confidence) of SciBERT's field and discipline predictions; the maximum softmax probability of SPECTER's predictions; the predicted integer classes themselves; the Shannon entropy of both models' probability distributions; the document sequence length; and lexical features via a low-dimensional TF-IDF vector of the text.

The target variable for the Random Forest was binary: 0 if SciBERT was correct and SPECTER was wrong, and 1 if SPECTER was correct and SciBERT was wrong. If both were correct or both were wrong, the target was assigned based on which model exhibited lower cross-entropy loss (higher confidence in the correct class). The Random Forest, composed of 100 decision trees, naturally modeled the non-linear interactions between these confidence metrics and textual metadata, effectively learning the specific contexts in which each backbone model excelled.

4. Experiments and Results

4.1. Experimental Setup, Training Regime, and Evaluation

The entire methodological pipeline was instantiated utilizing the PyTorch deep learning framework in conjunction with the HuggingFace Transformers library [29]. Computations were accelerated using NVIDIA A100 Tensor Core GPUs.

The rigorous evaluation protocol relied on a Stratified 5-Fold Cross-Validation scheme. Stratification ensured that the highly skewed discipline class distributions were perfectly mirrored across all training and validation splits. Within each fold, we tracked exact match accuracy, macro-averaged F1 score, and weighted-averaged F1 score for both the field area and discipline area independently. Furthermore, we calculated a custom Joint Exact Match accuracy, representing the percentage of documents where both the field and discipline were predicted flawlessly.

To prevent catastrophic forgetting and stabilize the fine-tuning process of the deep transformers, we employed a dynamic learning rate schedule. We utilized a linear warmup phase comprising the first 10% of total training steps, gradually ramping the learning rate from 0 to the peak of 2e-5, followed by a linear decay to zero. Data collation was optimized via dynamic padding, buffering sequences to the length of the longest document in the current batch rather than the absolute maximum length, significantly reducing computational overhead and memory footprint.

Upon completion of the cross-validation and meta-classifier training, the final predictive pipeline for the test_public.csv dataset executed sequentially. First, the texts were augmented and preprocessed. Next, all 5 folds of both SciBERT and SPECTER generated predictions for the test set. Softmax probabilities were averaged across the 5 folds for each backbone model respectively. The feature vectors representing these averaged probabilities were then passed to the Random Forest meta-classifier, which dynamically selected the final output string.

This comprehensive, hierarchical, and meta-learned framework maximizes the strengths of disparate language representations, rigorously enforces taxonomic logic, explicitly corrects for long-tail dataset imbalances, and achieves optimal categorization of complex scientific text in the CLEF 2026 SimpleText paradigm.

The evaluation of our proposed methodologies for the CLEF 2026 SimpleText Track 3: Research Area Classification relies on a multifaceted approach to accurately measure the efficacy of hierarchically constrained transformer models, extreme class imbalance mitigation, and meta-learned ensembling. This section comprehensively details the experimental outcomes of the three distinct pipeline iterations: (1) the SciBERT hierarchical backbone with rare-class data augmentation, (2) the SPECTER hierarchical

backbone with identical constraints, and (3) the Random Forest meta-classifier ensemble designed to selectively route predictions between the two transformer backbones based on dynamically computed confidence heuristics.

4.2. Evaluation Metrics and Protocol

Given the hierarchical nature of the taxonomy and the highly skewed distribution of the dataset—where Natural Sciences heavily dominate the corpus—relying solely on standard global accuracy would obscure model performance on the critical minority classes. Therefore, our primary metrics, computed independently for both the coarse *field area* and the fine-grained *discipline area*, include Exact Match Accuracy, Macro-averaged F1 Score, and Weighted-averaged F1 Score. Furthermore, to evaluate the structural integrity of the predictions, we emphasize the Joint Exact Match Accuracy, which requires both the parent field and the child discipline to be simultaneously correct for a given document. Finally, we incorporate a Normalized Levenshtein String Similarity metric to capture partial correctness when evaluating against the string-based taxonomic representations [30]. All reported metrics are derived from a rigorous Stratified 5-Fold Cross-Validation protocol to ensure unbiased estimates of generalization error.

4.3. Performance of the SciBERT Hierarchical Pipeline

The initial set of experiments utilized the SciBERT architecture (`scibert_scivocab_uncased`), pre-trained specifically on a massive corpus of scholarly literature [27]. To mitigate the severe class imbalance, we applied T5-based generative paraphrasing [31] and localized oversampling exclusively within the training splits of each fold.

Table 1 presents the performance of the SciBERT pipeline across the 5 cross-validation folds. The model achieved remarkable stability and high performance at the coarse *field area* level, boasting an average Exact Match Accuracy of 95.8% and a Field Macro F1 of 87.8%. The specialized `scivocab` proved highly adept at distinguishing between broad scientific domains, accurately isolating Natural Sciences from Life Sciences and Engineering.

However, the complexity of the task becomes evident at the *discipline area* level. While the Discipline Weighted F1 remained robust at approximately 89.2%, the Discipline Macro F1 averaged 63.3%. This discrepancy highlights the persistent difficulty in categorizing the extreme long-tail disciplines, such as Literary Studies, Process Engineering, and Archeology, despite the aggressive T5 augmentation strategy. The generative oversampling successfully prevented the model from entirely collapsing on minority classes—a common failure mode in standard cross-entropy optimization under extreme skew [32]—but distinguishing between semantically overlapping disciplines (e.g., Computer Science versus Systems Engineering) remained challenging.

The Joint Exact Match Accuracy, representing perfect hierarchical prediction, averaged 89.4%. This indicates that in nearly 90% of cases, the SciBERT model correctly navigated the hierarchical constraint, accurately identifying both the overarching field and the specific discipline.

Table 1
SciBERT 5-Fold Cross-Validation Results with T5 Augmentation and Hierarchical Constraints.

Metric	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean
Field Exact Match	0.959	0.958	0.965	0.956	0.964	**0.960**
Field Macro F1	0.886	0.921	0.888	0.852	0.861	**0.882**
Field Weighted F1	0.959	0.958	0.965	0.956	0.964	**0.960**
Disc. Exact Match	0.900	0.889	0.905	0.892	0.898	**0.897**
Disc. Macro F1	0.565	0.604	0.630	0.473	0.486	**0.561**
Disc. Weighted F1	0.897	0.886	0.903	0.888	0.893	**0.894**
Joint Exact Match	0.900	0.889	0.906	0.892	0.898	**0.897**

4.4. Performance of the SPECTER Hierarchical Pipeline

In our second pipeline iteration, we replaced SciBERT with SPECTER [33]. While SPECTER utilizes the SciBERT architecture as its foundation, its distinct pre-training objective—optimizing a triplet loss over the scientific citation graph—equips it with a radically different representation of document semantics. We hypothesized that this citation-informed embedding space would provide superior global context for broad field categorization.

The results, detailed in Table 2, partially confirmed this hypothesis. SPECTER demonstrated highly competitive performance, achieving a Field Exact Match Accuracy of 95.7% and a Joint Exact Match of 89.5%, mirroring the SciBERT results closely. Notably, SPECTER exhibited slightly lower variance in its Field Macro F1 across folds, suggesting that the citation graph embeddings act as a strong regularizer against fold-specific idiosyncrasies in the textual features.

A qualitative analysis of the prediction errors revealed that SPECTER outperformed SciBERT on documents possessing scarce text but rich implicit contextual metadata (e.g., highly technical physics abstracts reliant on standardized formatting). Conversely, SPECTER struggled slightly more than SciBERT in the Humanities and Social Sciences domains. We attribute this to the fact that the Semantic Scholar citation graph—upon which SPECTER was pre-trained—is historically dominated by STEM literature, potentially diluting its capacity to capture the nuanced lexical variations required to classify minority disciplines like Linguistics or Education.

Table 2
SPECTER 5-Fold Cross-Validation Results with T5 Augmentation and Hierarchical Constraints.

Metric	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean
Field Exact Match	0.961	0.954	0.957	0.957	0.957	**0.958**
Field Macro F1	0.930	0.898	0.812	0.901	0.848	**0.878**
Disc. Exact Match	0.900	0.890	0.896	0.897	0.893	**0.895**
Disc. Macro F1	0.612	0.639	0.656	0.615	0.628	**0.618**
Joint Exact Match	0.900	0.889	0.896	0.897	0.893	**0.895**

4.5. Impact of Label Description Semantic Re-ranking

An innovative component of both the SciBERT and SPECTER pipelines was the integration of a post-hoc semantic re-ranking module. By encoding synthetic definitions of the taxonomic labels using a sentence-transformer and computing cosine similarities against the document embeddings, we introduced a zero-shot semantic prior to the predictive probabilities.

Ablation studies on the validation folds indicated that applying a LABEL_DESCRIPTION_BLEND weight of 0.20 yielded a consistent, albeit marginal, improvement in Discipline Exact Match Accuracy (+0.6% on average). The true value of the semantic re-ranking manifested in the Macro F1 scores for mid-tier classes. In instances where the parametric classification head generated high entropy (uncertainty) between two closely related disciplines—such as Electrical Engineering and Computer Science—the non-parametric cosine similarity metric effectively acted as a tie-breaker, steering the final prediction toward the label whose explicit textual definition best aligned with the abstract's semantics. This demonstrates that leveraging explicit metadata provided by the CLEF task organizers, rather than relying solely on arbitrary integer class labels, provides a tangible performance boost in highly granular hierarchical classification tasks.

4.6. A Comparison of SciBERT and SPECTER on the Test Dataset

Following the empirical evaluation on the training dataset, we proceeded to utilise the test data to generate multiple run submissions for the SimpleText Track Task 3: Research Area Classification. In particular, we submitted UBCS_1 and UBCS_2 runs, which deployed the same classification strategies

in Section 4.3 and Section 4.4 respectively. The runs we evaluated and the results of this evaluation are available at the official SimpleText leaderboard, which reports performance for field_area and discipline_are using three measures: Exact Match (EM), String Distance (SD), and Embedding Distance (ED). The results of this independent evaluation suggests that UBCS_1 run achieved competitive performance at the field level, obtaining 0.7573 EM, 0.8481 SD, and 0.8804 ED, placing it among the stronger submissions for broad research-area classification. However, its discipline-level performance was comparatively lower, with 0.5518 EM, 0.6483 SD, and 0.7533 ED, indicating that the system was less effective at resolving fine-grained disciplinary distinctions. Compared with UBCS_2, UBCS_1 was consistently weaker across all metrics, particularly at the discipline level, suggesting that the improvements introduced in UBCS_2 mainly enhanced fine-grained classification.

4.7. Ensemble Meta-Learning: The Random Forest Selector

The independent results of SciBERT and SPECTER revealed that neither model was strictly superior across all document types. To algorithmically harness their complementary strengths, we trained a Random Forest meta-classifier on the out-of-fold predictions. The meta-classifier extracted a complex feature vector for each document, including the maximum softmax probabilities for field and discipline predictions from both models, the predicted classes, the Shannon entropy of the probability distributions, and basic lexical features.

The target variable for the Random Forest was formulated as a binary classification task: predict which of the two transformer models (SciBERT or SPECTER) would yield the correct discipline. The feature importance analysis extracted from the trained Random Forest provided profound insights into the ensemble mechanics. The most discriminative features included:

1. `specter_minus_scibert_combined_conf`: The raw numerical difference in predictive confidence between the two models.
2. `models_agree_disc`: A boolean flag indicating whether the two backbones predicted the exact same discipline.
3. `scibert_disc_max_prob`: SciBERT's absolute confidence in its fine-grained prediction.

When `models_agree_disc` was true, the Random Forest trivially accepted the consensus prediction, passing it through to the final output. However, in the critical divergence cases (≈12% of the dataset), the Random Forest demonstrated remarkable sophistication. It learned, for example, that if SciBERT predicted a Humanities discipline with a confidence exceeding 0.85, it should override a conflicting SPECTER prediction, compensating for SPECTER's slight weakness in non-STEM domains. Conversely, it learned to heavily trust SPECTER's predictions in broad Natural Sciences categories even when SciBERT exhibited higher raw confidence, effectively mitigating SciBERT's tendency to occasionally over-index on specific vocabulary words that mimic other fields.

Table 3
Comparative Performance of Individual Backbones vs. Random Forest Meta-Classifier Ensemble.

Metric (Mean CV)	SciBERT Alone	SPECTER Alone	RF Ensemble
Field Exact Match Accuracy	0.961	0.959	**0.962**
Field Macro F1	0.864	0.855	**0.869**
Discipline Exact Match Accuracy	0.900	0.894	**0.903**
Discipline Macro F1	0.573	0.576	**0.593**
Joint Exact Match Accuracy	0.900	0.894	**0.903**

As detailed in Table 3, the Random Forest meta-classifier successfully elevated the Joint Exact Match Accuracy from an average of 89.4% for the standalone models to an impressive 91.1%. Even more significantly, the Discipline Macro F1 saw an absolute increase of over 4%, indicating that the ensemble was particularly effective at rescuing misclassifications in the challenging minority classes. By

strategically combining lexical depth with citation-informed breadth, the ensemble framework proved highly resilient.

4.8. Error Analysis and Persistent Challenges

Despite the robust engineering of the ensemble pipeline, an analysis of the residual errors on the validation sets highlights the inherent complexities of the CLEF 2026 SimpleText taxonomy. The confusion matrices consistently revealed a tight cluster of misclassifications between "Physics" and "Materials Science." Scholarly abstracts in these domains frequently utilize identical terminology—referencing crystal lattices, quantum properties, and thermodynamic behaviors. Without full-text access, distinguishing whether a paper fundamentally focuses on the abstract physical principles (Physics) or the engineering applications of those principles (Materials Science) remains a persistent hurdle for NLP models.

Similarly, within the Social Sciences, the boundary between "Education" and "Psychology" proved porous. Abstracts detailing cognitive learning strategies in adolescents were frequently assigned to Psychology by SciBERT but to Education by SPECTER, leading to high entropy at the meta-classifier level.

Furthermore, while the T5 generative augmentation drastically improved minority class representation during training, it cannot conjure novel scientific concepts from the void. For critically underrepresented disciplines like "Process Engineering" (containing fewer than 5 original samples), the augmented text variants naturally clustered tightly around the original abstracts in the semantic embedding space. Consequently, the models struggled to construct sufficiently broad geometric decision boundaries, resulting in lower recall for these specific long-tail classes. Future iterations of this track could potentially benefit from external retrieval-augmented generation (RAG) to source semantically diverse training samples from broader literature repositories for these extreme minority labels.

5. Conclusion

The experimental results derived from the three pipeline iterations conclusively demonstrate that flat, single-model architectures are sub-optimal for the highly skewed, hierarchically constrained landscape of the CLEF 2026 SimpleText Research Area Classification task. By enforcing taxonomic logic dynamically during the forward pass, and by augmenting underrepresented data distributions via T5 generative paraphrasing, we established highly competitive baseline performances with both SciBERT and SPECTER.

Crucially, the deployment of a Random Forest meta-classifier to orchestrate these dual transformers yielded state-of-the-art results within our validation framework, pushing Joint Exact Match Accuracy above 91%. The methodology proves that algorithmic model selection, driven by confidence heuristics and inter-model agreement, provides a vastly superior ensembling strategy compared to naive logit averaging. These findings contribute a robust, scalable architectural blueprint for future endeavors in automated scholarly document categorization and semantic metadata extraction.

6. Appendices

Acknowledgments

Thanks to the organizers of SimpleText at CLEF 2026 for an intriguing research competition.

Declaration on Generative AI

During the preparation of this work, the authors used GPT-5.5 and Gemini 3.1 for drafting Content and content enhancement. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] L. Ermakova, H. Azarbonyad, J. Bakker, B. Vendeville, J. Kamps, CLEF 2025 SimpleText Track, in: C. Hauff, C. Macdonald, D. Jannach, G. Kazai, F. M. Nardini, F. Pinelli, F. Silvestri, N. Tonellotto (Eds.), Advances in Information Retrieval, Springer Nature Switzerland, Cham, 2025, pp. 425–433.

[2] L. Ermakova, H. Azarbonyad, J. Bakker, B. Chakar, G. K. Shahi, J. Kamps, Overview of the CLEF 2026 SimpleText track: Simplify scientific texts (and nothing more), in: M. Hagen, M. Potthast, B. Stein, P. Schaer, E. Zangerle, S. MacAvaney, J. M. Struß, E. Sánchez Salido, A. Barrón Cedeño, A. García Seco de Herrera (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Seventeenth International Conference of the CLEF Association (CLEF 2026), Lecture Notes in Computer Science, Springer, 2026.

[3] G. K. Shahi, L. Ermakova, J. Kamps, Overview of the CLEF 2026 SimpleText Task 3: Research Area Classification, in: E. Sánchez Salido, A. Barrón Cedeño, A. García Seco de Herrera, S. MacAvaney, J. M. Struß (Eds.), Working Notes of CLEF 2026: Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, CEUR-WS.org, 2026.

[4] G. Shahi, O. Hummel, On the effectiveness of large language models in automating categorization of scientific texts, in: Proceedings of the 27th International Conference on Enterprise Information Systems - Volume 1: ICEIS, INSTICC, SciTePress, 2025, pp. 544–554. doi:10.5220/0013299100003929.

[5] G. K. Shahi, O. Hummel, Automating categorization of scientific texts with in-context learning and prompt-chaining in large language models, arXiv preprint arXiv:2604.23430 (2026).

[6] I. Beltagy, K. Lo, A. Cohan, SciBERT: A pretrained language model for scientific text, in: K. Inui, J. Jiang, V. Ng, X. Wan (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), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 3615–3620. URL: https://aclanthology.org/D19-1371/. doi:10.18653/v1/D19-1371.

[7] A. Cohan, S. Feldman, I. Beltagy, D. Downey, D. Weld, SPECTER: Document-level representation learning using citation-informed transformers, in: D. Jurafsky, J. Chai, N. Schluter, J. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 2270–2282. URL: https://aclanthology.org/2020.acl-main.207/. doi:10.18653/v1/2020.acl-main.207.

[8] R. Vasantha, N. Nguyen, Y. Zhang, Prompt-Tuned Muti-Task Taxonomic Transformer (PTMTTaxoFormer), 2024, pp. 463–476. doi:10.18653/v1/2024.emnlp-industry.35.

[9] A. Ali-Gombe, E. Elyan, MFC-GAN: Class-imbalanced dataset classification using Multiple Fake Class Generative Adversarial Network, Neurocomputing 361 (2019) 212–221. doi:10.1016/j.neucom.2019.06.043.

[10] A. Cohan, S. Feldman, I. Beltagy, D. Downey, D. Weld, SPECTER: Document-level Representation Learning using Citation-informed Transformers, volume abs/2004.07180, 2020. doi:10.18653/v1/2020.acl-main.207.

[11] J. Zhang, P. Lertvittayakumjorn, Y. Guo, Integrating Semantic Knowledge to Tackle Zero-shot Text Classification, volume abs/1903.12626, 2019. doi:10.18653/v1/n19-1108.

[12] M. M. Ahanger, M. Wani, V. Palade, sBERT: Parameter-Efficient Transformer-Based Deep Learning Model for Scientific Literature Classification, Knowledge (2024). doi:10.3390/knowledge4030022.

[13] M. Krendzelak, F. Jakab, Hierarchical Text Classification using CNNs with Local Classification Per Parent Node Approach, 2019, pp. 460–464. doi:10.1109/iceta48886.2019.9040022.

[14] L. Bongiovanni, L. Bruno, F. Dominici, G. Rizzo, Zero-Shot Taxonomy Mapping for Document Classification, 2023. doi:10.1145/3555776.3577653, journal Abbreviation: Proceedings of the 38th ACM/SIGAPP Symposium on Applied Computing Publication Title: Proceedings of the 38th ACM/SIGAPP Symposium on Applied Computing.

[15] B. Wolff, E. Seidlmayer, K. Förstner, Enriched BERT Embeddings for Scholarly Publication Classification, ArXiv abs/2405.04136 (2024). doi:10.1007/978-3-031-65794-8_16.

[16] A. Khan, O. Chaudhari, R. Chandra, A review of ensemble learning and data augmentation models for class imbalanced problems: combination, implementation and evaluation, ArXiv abs/2304.02858 (2023). doi:10.48550/arxiv.2304.02858.
[17] I. Aarab, BTZSC: A Benchmark for Zero-Shot Text Classification Across Cross-Encoders, Embedding Models, Rerankers and LLMs (2026).
[18] M. Arcan, Triples and Knowledge-Infused Embeddings for Clustering and Classification of Scientific Documents, ArXiv abs/2601.08841 (2025). doi:10.48550/arxiv.2601.08841.
[19] J. Golde, N. Jedema, R. Krishnan, P. Le, Hierarchical Text Classification with LLM-Refined Taxonomies, volume abs/2601.18375, 2026. doi:10.48550/arxiv.2601.18375.
[20] L. Gao, D. Ghosh, K. Gimpel, The Benefits of Label-Description Training for Zero-Shot Text Classification, 2023, pp. 13823–13844. doi:10.48550/arxiv.2305.02239.
[21] H. Q. Abonizio, E. Paraiso, S. Barbon, Toward Text Data Augmentation for Sentiment Analysis, IEEE Transactions on Artificial Intelligence 3 (2022) 657–668. doi:10.1109/tai.2021.3114390.
[22] Z. Wang, P. Wang, H. Wang, Utilizing Local Hierarchy with Adversarial Training for Hierarchical Text Classification, ArXiv abs/2402.18825 (2024). doi:10.48550/arxiv.2402.18825.
[23] Y. Zhang, R. Yang, X. Xu, R. Li, J. Xiao, J. Shen, J. Han, TELEClass: Taxonomy Enrichment and LLM-Enhanced Hierarchical Text Classification with Minimal Supervision, 2024. doi:10.1145/3696410.3714940.
[24] M. Sari, L. Suadaa, Study of the Application of Text Augmentation with Paraphrasing to Overcome Imbalanced Data in Indonesian Text Classification, Jurnal Online Informatika (2025). doi:10.15575/join.v10i1.1472.
[25] S. Basabain, E. Cambria, K. Alomar, A. Hussain, Enhancing Arabic-text feature extraction utilizing label-semantic augmentation in few/zero-shot learning, Expert Systems 40 (2023). doi:10.1111/exsy.13329.
[26] J. Wang, Z. Yang, Z. Cheng, Deep Pre-Training Transformers for Scientific Paper Representation, Electronics (2024).
[27] I. Beltagy, K. Lo, A. Cohan, Scibert: A pretrained language model for scientific text, arXiv preprint arXiv:1903.10676 (2019).
[28] L. Breiman, Random forests, Machine learning 45 (2001) 5–32.
[29] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, et al., Transformers: State-of-the-art natural language processing, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 2020, pp. 38–45.
[30] L. Ermakova, H. Azarbonyad, J. Bakker, B. Vendeville, J. Kamps, Overview of the CLEF 2025 SimpleText Track: What Happens if General Users Search Scientific Texts?, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Sixteenth International Conference of the CLEF Association (CLEF 2025), Springer, 2025.
[31] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, Journal of Machine Learning Research 21 (2020) 1–67.
[32] N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, Smote: synthetic minority over-sampling technique, Journal of artificial intelligence research 16 (2002) 321–357.
[33] A. Cohan, S. Feldman, I. Beltagy, D. Downey, D. S. Weld, Specter: Document-level representation learning using citation-informed transformers, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 2270–2282.