Enhancing automated indexing of publication types and study designs in biomedical literature using full-text features

Joe D. Menke[a], Shufan Ming[a], Shruthan Radhakrishna[b], Halil Kilicoglu[a], Neil R. Smalheiser[a,c]

[a] *School of Information Sciences, University of Illinois Urbana-Champaign, 501 E Daniel Street, Champaign, 61820, IL, USA*
[b] *Department of Computer Science, University of Illinois Urbana-Champaign, 201 North Goodwin Avenue, Urbana, 61801, IL, USA*
[c] *Department of Psychiatry, University of Illinois Chicago, 1601 W Taylor Street, Chicago, 60612, IL, USA*

Abstract

Objective: Searching for biomedical articles by publication type or study design is essential for tasks like evidence synthesis. Prior work has relied solely on PubMed information or a limited set of types (e.g., randomized controlled trials). This study builds on our previous work by leveraging full-text features, alternative text representations, and advanced optimization techniques.
Methods: Using a dataset of PubMed articles published between 1987 and 2023 with human-curated indexing terms, we fine-tuned BERT-based encoders (PubMedBERT, BioLinkBERT, SPECTER, SPECTER2, SPECTER2-Clf) to investigate whether text representations based on different pre-training objectives could benefit the task. We incorporated textual and verbalized metadata features, full-text extraction (rule-based, extractive, and abstractive summarization), and additional topical information about the articles. To improve calibration and mitigate label noise, we used asymmetric loss and label smoothing. We also explored contrastive learning approaches (SimCSE, ADNCE, HeroCon, WeighCon). Models were evaluated using precision, recall, F1 score (both micro- and macro-), and area under ROC curve (AUC).
Results: Fine-tuning SPECTER2-base with adding the MeSH term "Animals", asymmetric loss with label smoothing, and WeighCon contrastive loss improved performance significantly over the previous best architecture

(micro-F1: 0.664 → 0.679 [+2.2%]; macro-F1: 0.663 → 0.690 [+4.1%]; $p < 0.0001$). Asymmetric loss and using SPECTER2-base instead of PubMedBERT contributed most to this gain. Full-text features boosted performance by 2.4% (micro-F1) and 1.8% (macro-F1) over the baseline (micro-F1: 0.616 → 0.631; macro-F1: 0.556 → 0.566; $p < 0.0001$). Topical label splitting and contrastive learning provided minor, non-significant improvements.
Conclusion: Full-text features, enhanced document representations, and fine-tuning optimizations improve publication type and study design indexing. Future work should refine label accuracy, better distill relevant article information, and expand label sets to meet needs of the research community. Data, code, and models are available at `https://github.com/ScienceNLP-Lab/MultiTagger-v2`.

Keywords: natural language processing, literature mining, publication type indexing, study design indexing, information retrieval, evidence synthesis

1. Introduction

The ability to search and filter research publications efficiently is crucial for all researchers. This is especially true in areas, such as evidence-based medicine, which uses the best available clinical evidence to inform treatment [1]. The size and rapid expansion of biomedical literature poses a significant challenge, making the identification of relevant publications a lengthy and labor-intensive task [2, 3, 4]. Improvements to literature screening processes are poised to significantly benefit downstream tasks, including evidence synthesis. The U.S. National Library of Medicine (NLM) indexes publications in MEDLINE, their bibliographic database, by Medical Subject Headings (MeSH) and publication types (PT) to help facilitate article retrieval. Historically, the task of indexing MEDLINE articles has been carried out manually by skilled medical indexing experts [5]. Over the past two decades, NLM's Medical Text Indexer (MTI) program has increasingly automated article indexing [6, 7, 8, 9]. The most recent MTI model uses a combination of convolutional neural networks (CNN) and PubMedBERT to generate and rank MeSH candidates [9]. Since 2022, automatic MeSH indexing has been applied to all journals indexed for MEDLINE with human indexers performing subsequent review and curation to ensure accuracy and completeness, which reduced the average indexing time from 145 days down to just one day [10].

Outside of the NLM, the challenge of automatically indexing the biomedical literature continues to garner substantial attention. For example, since 2013, the annual BioASQ shared tasks have focused on advancing MeSH indexing methods with MTI often serving as a baseline [11, 12, 13]. Notable approaches include ensembling models using a learning-to-rank framework (e.g., MeSHLabeler[14])), combining learning-to-rank with deep representations (e.g., DeepMESH [15]), using contextualized representations and fine-tuning (e.g., BERTMeSH [16]), incorporating full-text content (FullMeSH [17]), and models that also incorporate novel attention-based mechanisms (e.g., MeSHProbeNet-P [18], KenMeSH [19]). The focus of these methods is to assign MeSH headings and subheadings to publications. These headings are primarily based on an article's topic of study, e.g., “Brain” or “Asthma”. However, this is not the case for all MeSH terms. A relatively small subset of MeSH terms focus on methodological characteristics of research, indicating how a study was conducted rather than its subject matter, e.g., Cohort Studies or Double-Blind Method. This information is particularly important for evidence synthesis pipelines, serving as an initial automated filtering step before manual analysis and synthesis [20, 21, 22, 23]. A significant body of work has emerged focusing on these study design-related terms as well as publication types (collectively referred to here as PTs). For example, RCT Tagger classifies Randomized Controlled Trials (RCT) using a SVM model with n-gram based features and manually annotated features from MEDLINE (i.e., MeSH terms and publication type) to support systematic reviews [24]. Building on this work, Wallace et al. [25] used similar machine learning methods along with crowdsourcing, while Marshall et al. [26] built binary RCT classifiers using ensembles of SVMs, CNNs, and MEDLINE publication type tags. Extending beyond RCTs, MultiTagger, a suite of binary SVM classifiers, was designed [27] and developed [28] to generate probabilistic estimates of 50 different PTs. RCT Tagger and MultiTagger both showed high levels of recall in identifying RCTs, resulting in little information loss [29] at much faster speeds [30] demonstrating their value for systematic reviews. Focusing more on pre-clinical animal research, Neves et al. developed models to identify study designs that may serve as alternatives to animal experiments (e.g., *in vitro*) [31]. Most recently, we developed a new version of MultiTagger formulating the task as multi-label classification and fine-tuning PubMedBERT [32] using text (title, abstract) and metadata features from PubMed [33]. Using this model, we improved micro-F1 by 40% ($0.497 \rightarrow 0.697$) and macro-F1 by 52% ($0.416 \rightarrow 0.632$) compared to

MultiTagger. Performance was compared on averages across 49 labels using a test of 64,400 PubMed articles where title and abstracts were longer than 25 characters.

In this study, we extend our prior work in several directions:

- We investigate whether (and which) full-text features could improve PT classification. In prior work, full-text information was shown to benefit MeSH indexing [16, 17]. Due to the context size limitations of BERT-based models (512 tokens), which is much smaller than the average full-text article, and based on the hypothesis that most full-text content is irrelevant to PT classification, we conduct experiments with extractive and abstraction summarization methods.

- We assess whether enriched document representations could benefit PT classification task. Specifically, we hypothesize that publications may cite other studies that share methodological similarities and fine-tune several Transformer-based models that are pre-trained in part using citation-related information (BioLinkBERT [34], SPECTER [35]). Further, we hypothesize that, compared to PubMedBERT which uses standard masked language modeling and next sentence prediction as pre-training objectives, models pre-trained on document classification tasks could provide enriched representations that benefit the PT classification task and experiment with such models (SPECTER2 [36]). We also explore unsupervised and supervised contrastive learning in depth to in an effort to better align document representations with the labels. Specifically, we compare unsupervised contrastive loss approaches (SimCSE [37] and ADjusted InfoNCE (ADNCE) [38]) with supervised approaches (HeroCon [39] and WeighCon [40]).

- We attempt to address dataset-related issues using advanced optimization techniques in model fine-tuning and label splitting. Although we have limited our dataset to articles curated by NLM indexers only, there is some noise and inconsistencies in PT indexing [24, 29]. Given that relabeling articles at large scale would be infeasible, we opt for mitigating the effect of label noise through more advanced regularization techniques. These include asymmetric loss [41], an adaptation of cross-entropy loss that focuses on harder data and disregards possibly mislabeled data during training, and label smoothing [42, 43], which disperses small probabilities from correct labels among incorrect labels for regularization. Furthermore, given the hierarchical nature of some PT labels and their heterogeneity, we add labels for more fine-grained categories during training to evaluate whether more discriminative representations can be learned.

Statement of significance	
Problem	Manually indexing research articles by publication type and study design is difficult and time-consuming.
What is Already Known	Automated approaches to PT and SD indexing focus on titles and abstracts and are mostly limited to a small number of types or use somewhat noisy datasets as ground truth without consideration of this noise.
What This Paper Adds	We account for dataset quality through noise-aware training techniques as well as exploring the use of full-text in publication type and study design indexing.
Who Would Benefit	Researchers and clinicians who use research repositories (e.g., PubMed), especially those who use publication types and study designs for article screening and evidence synthesis.

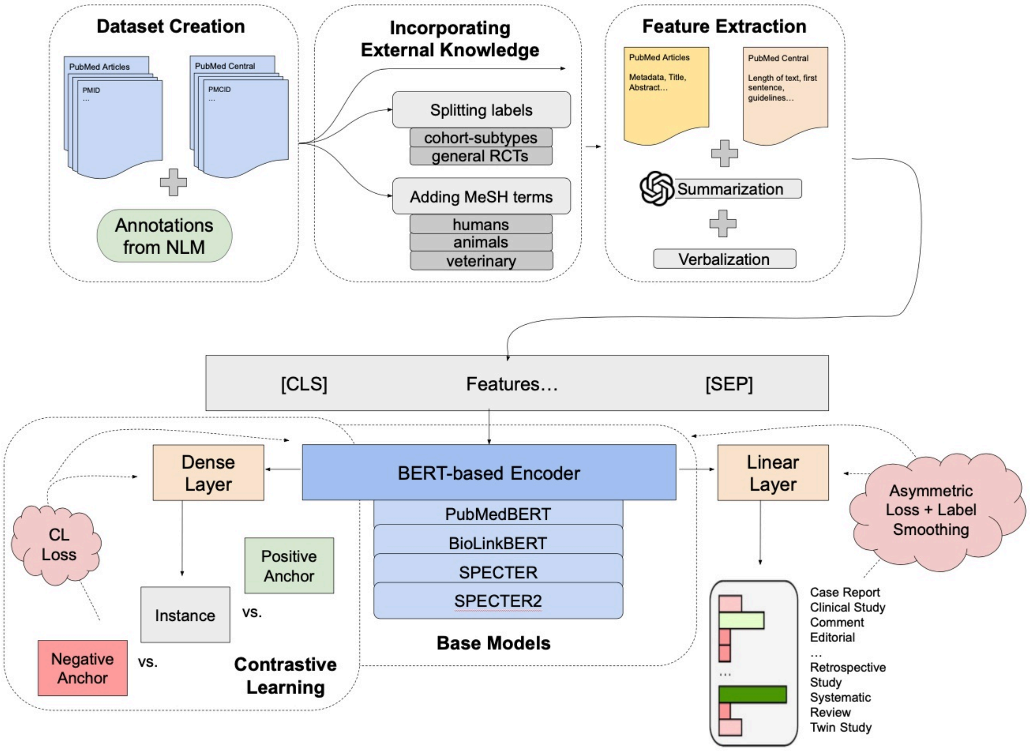
2. Materials and Methods

The components of our overall approach and the processing pipeline is illustrated in Figure 1. We describe each component below. We also discuss the experiments we performed and the evaluation methodology at the end of this section.

2.1. Dataset Construction

Candidate articles were initially selected using PubMed queries and downloaded using the NCBI e-utilities API [44]. We did not use the dataset from our previous work [33] as some queries were changed to enhance query consistency (e.g., use of ”mh:noexp” when the MeSH term had children). Additionally, we no longer require the “Humans” MeSH term in some queries (e.g., Clinical Study) and added additional “as topic” PTs (e.g., Clinical Trials as Topic and Meta-Analysis as Topic). All PT queries restricted results to (1) English articles or articles containing an English abstract, and (2)

Figure 1: Flow diagram of dataset and feature construction as well as experiments including feature extraction, incorporating external knowledge, base encoders, and contrastive learning.



articles published between 1987 and 2023, except for Retraction of Publication to increase the number of positively labeled articles. A majority of PTs also required the article to be human-indexed by NLM. We relaxed this restriction for Scientific Integrity Reviews, which have more easily discernible features where including non-human-indexed articles may not significantly lower dataset quality and for Systematic Reviews, where we used search terms rather than publication types or MeSH terms based on previous findings regarding the quality of articles tagged (or not tagged) as Systematic Reviews before 2019 [28]. Following previous work [28, 33], we excluded articles with Editorial, Letter, Comment, Practice Guideline, or Review PTs for a subset of clinically-focused PTs (e.g., RCTs) as these were commonly found to simply discuss the PT rather than be of that type based on initial exploration. Additionally, for RCTs, we also required the "Humans" MeSH term to better differentiate between Human RCTs and Veterinary RCTs, which we classify using a separate label similar to NLM. As an example, the query used for the Clinical Trial PT is provided below.

> *(1987:2023[dp] AND (english[Language] OR english abstract[pt]) NOT (indexingmethod_curated OR indexingmethod_automated)) AND "clinical trial"[pt] NOT case-control studies[mh:noexp] NOT cohort studies[mh:noexp] NOT editorial[pt] NOT letter[pt] NOT comment[pt] NOT "practice guideline"[pt] NOT review[pt]*

The queries used for dataset construction are available for download at the project GitHub repository. While potentially somewhat noisy, we used the query results as our gold-standard PT labels within this work. In total, labels were obtained for 9.6 million PubMed articles, featuring 61 distinct PTs. We sampled from this initial candidate set using a modified version of stratified sampling to restrict the dataset to a more reasonable size for training, while making efforts to prevent any majority class from dominating the dataset and retaining enough rare labels to effectively train the models. The stratified dataset was supplemented with randomly selected negative articles, i.e., articles containing no positive labels, totaling 20% of the entire dataset.

2.1.1. Alternate Label Splits

Some PT labels are hierarchical in nature and have high topical heterogeneity, making it difficult to learn effective representations with models. In a set of experiments, we aimed to evaluate whether adding more information related to fine-grained label splits could improve model performance by allowing it to better learn these heterogeneous labels. For these experiments, we added new labels into the dataset during the training process (and removed them before evaluation).

First, we added labels that represented overlaps between two heterogeneous PTs in efforts to steer the model to pay more attention to the correlation between these PTs. For example, Longitudinal Studies is a PT under Cohort Studies in MeSH, so a Longitudinal-Cohort Studies label was added. Positively labeled instances in this case are articles that were already labeled as both Longitudinal Studies and Cohort Studies. This was done for all children PTs of Cohort Studies (i.e., Follow-Up Studies, Longitudinal Studies, Prospective Studies, and Retrospective Studies). This experiment is referred to as COHORT-SPLIT. Similarly, for the GENERALIZED RCT experiment, we added a new label encompassing articles labeled as either Human RCTs or Veterinary RCTs to see if this could help the model better understand the study design, RCTs, which is independent of species.

We also performed experiments distinguishing HUMAN, ANIMAL, and VETERINARY studies. For HUMAN and ANIMAL experiments, we added a binary label to the multi-label dataset based on the presence of specific MeSH terms for the article ("Humans", "Animals"). We labeled articles during the VETERINARY experiment if they had one of the following MeSH terms: "Dogs", "Cats", "Cattle", "Horses", and "Swine". These terms were identified based on their frequency (>20) in articles tagged using the Randomized Controlled Trial, Veterinary PT.

For the COMBINED experiment, we added labels during training from these other experiments: COHORT-SPLIT, GENERALIZED RCT, HUMAN, ANIMAL, and VETERINARY.

Note that these terms were only used during training, and not in evaluation to ensure that the results are comparable to those of the models trained with the standard labels.

2.2. Feature Extraction

2.2.1. PubMed Features

We mainly followed our earlier work [33] for extracting PubMed features. For each article, features (title, abstract, and other metadata) were extracted from PubMed and verbalized as model input. Features and verbalizations are described in Table 1. Feature verbalization is meant to better contextualize features in efforts to improve BERT-based representations, which are trained to rely on context within their masked language modeling task. Previously, we extracted words spelled in all caps present within the title or abstract, however, in this work we only extract these from the title in efforts to primarily capture acronyms. Additionally, we added a feature related to the number of affiliations of an article and combined some features (e.g., number of authors and number of affiliations) to reduce input length. When a feature was missing, empty strings ("") were used instead for verbalization. Additionally, we verbalized numbers in the metadata features ("There are 4 authors" → "There are four authors") to prevent information loss during tokenization.

Feature	**Verbalization**
Title text	*This article's title is ...*
Abstract text	No verbalization performed on this feature.
Journal name & publication date	*This article was published in ... in ...*
Keywords	*This article's keywords are ...*
Number of references	*This article's cited ... references*
Number of authors & affiliations	*This article was written by ... authors from ... affiliations.*
Chemicals (number & list)	*This article used ... chemicals: ...*
Capitalized title words	*The title uses the abbreviations ...*

Table 1: The features extracted from PubMed that were used in experiments. In the cases where the feature is missing, an empty string replaces it.

2.2.2. Full-Text Features

For experiments involving full-text articles, a variety of features were extracted from PubMed Central (PMC) when available. The extracted features and examples of how they were verbalized are shown in Table 2. A subset of the features are explained below, others are self-explanatory:

- *Number features* refers to features related to (1) the number of figures detected, (2) the number of tables detected, and (3) the approximate article word count.

- *Guidelines* refers to a regular expression-based feature that captures the mentions of various reporting guidelines (e.g., "CONSORT" for RCTs and "STROBE" for observational studies). Mention of guidelines may provide a clue to its PT.

- *Ethics* refers to a regular expression-based feature that captures the mentions of ethical approval. This feature could help models differentiate between PTs that commonly use humans versus those that do not.

- *Identifiers* refers to a regular expression-based rule that determines the location and frequency of clinical identifiers (e.g., NCT# from clinicaltrials.gov detected in the methods section or a table).

- *1st sentence* refers to the first sentence in the Methods section (if detected). Otherwise, the first sentence in the full-text was used.

- *1st paragraph* refers to the first paragraph in the Methods section (if detected). Otherwise, the first paragraph in the full-text was used.

- *Label sentences* refers to sentences that mention any PT label (e.g., Randomized Controlled Trials), which are identified using regular expressions.

A few features verbalized missing features: caption features ("No figure/table detected."), guidelines ("No reporting guidelines were detected."), ethics ("No ethical approvals detected."), and label sentences ("No sentences containing labels were detected."). All other features verbalized missing features using an empty string.

Feature	**Verbalization**
Number of tables	*There are four tables.*
Number of figures	*There are three figures.*
Length of full-text	*The article is four hundred and seventy two words long.*
Section headings	*The section headings are Introduction, Methods, Results, and Discussion.*
Figure captions	*Figure captions are . . . and . . .*
Table captions	*Table captions are . . . and . . .*
Guidelines	*The following reporting guidelines are mentioned: STROBE.*
Ethics	*The following ethical approvals are mentioned: Institutional Review Board.*
Identifiers	*2 clinical identifiers found in the methods section. 6 clinical identifiers found in tables.*
1st sentence	*A multicenter, observational, prospective study was conducted in France...*
1st paragraph	*A multicenter, observational, prospective study was conducted in France... This study was...*
Label sentences	*The inverse-variance fixed-effects model was used for meta-analysis.*

Table 2: The full-text features used in full-text specific experiments, as well as examples showing how they were verbalized.

2.2.3. Summarization

We also explored summarizing relevant information from full-text as input to the model. Instead of processing the entire document, we focused on sentences from the Introduction and Methods sections only (referred to as "full-text" below for brevity). This approach is motivated by two considerations. First, the model's token limitations make it impractical to process lengthy full-text documents efficiently. Second, narrowing the input to these targeted sections provides the model with more relevant information. Our preliminary analysis of 50 articles showed that sentences containing key PT information often appear in the Introduction or Methods sections. This is important, as models often struggle with long-context memorization in summarization, which can result in hallucinations or the generation of irrelevant content [45].

We employed abstractive and extractive summarization methods to extract study design-related information from full-text. We augmented the PubMed and full-text features with summarization-based features to fine-tune the PT classification model (described below). Abstractive summarization provides a high-level overview of an entire article, often involving rewriting, paraphrasing, and reorganizing source document—processes that are prone to errors [46]. In contrast, extractive summarization ensures factual consistency by directly selecting relevant content from the source document.

As an extractive summarization baseline, we used TextRank[47], an unsupervised graph-based ranking algorithm that can be used for extractive text summarization. It represents text as a graph where sentences are nodes, and edges represent their similarity (i.e., content overlap). By applying the PageRank algorithm, TextRank identifies the most "important" sentences based on their centrality in the graph.

We also utilized an off-the-shelf dense retriever, BMRetriever, to extract the most reliable grounding information, serving as an extractive summary [48]. This model leverages the capabilities of autoregressive large language models (LLMs) that can follow users' natural language instructions and is further fine-tuned on a combination of five biomedical tasks with labeled synthetic user query and response pairs. Given a task description, the model returns scores for each input sentence, indicating their informativeness and relevance. We experimented with three extractive prompts: EXTRACTIVE (SINGLE), EXTRACTIVE (MULTIPLE), and EXTRACTIVE (MULTIPLE+DEFINITION). The EXTRACTIVE (SINGLE) experiment used a consolidated query listing all target publication type labels with the prompt:

> "Given the following categories of interest and a biomedical article, retrieve sentences that are indicative of any of these study designs or publication types."

When all label names are presented together, the model may struggle to distinguish between the nuances of each label, especially if they are semantically similar or overlapping. This could result in retrieving sentences that are only somewhat related or not specific enough to each label. Therefore, we also employ the following multiple retrieval approaches to test our hypothesis. The EXTRACTIVE (MULTIPLE) experiment employed each category label as a distinct query. This approach assumes that the model understands the label's meaning and can retrieve sentences that are semantically similar to the full-text. Finally, EXTRACTIVE (MULTIPLE+DEFINITION) paired category labels with brief definitions to provide additional context and enhance the retrieval process. In both methods, sentences were ranked based on their similarity scores, and the top 20 sentences were selected for further use. 20 sentences were chosen to balance the context length limitations of the model (i.e., 512 tokens) with the need to ensure sufficient coverage of potentially relevant information.

Longformer [49], a sequence-to-sequence model, has been widely studied for the abstractive summarization of long documents, including biomedical articles [50]. PRIMERA [51], an encoder-decoder model based on the Longformer-Encoder-Decoder architecture, is further trained with a gap sentence generation objective, where salient sentences were masked to encourage the model to generate them. Since study design information can be implicit and may need to be inferred across sentences or sections, this model's ability to leverage global attention and pre-training on the DOC-SEP token, which was assigned global attention during pretraining and marks boundaries between sections, enables efficient aggregation of information across them, potentially improving memory retention and reducing hallucination issues.

Additionally, we prompted a large language model (META-LLAMA/LLAMA-3.2-3B-INSTRUCT [52]) to generate an open-ended, publication-type-focused summary as auxiliary input, following the concept of query-focused summarization [53], which aims to generate summaries based on a particular user's interest. The query was as follows:

> "Below is the title, journal, and a partial excerpt (introduction + methods) of a biomedical article. Your task is to summarize the article, focusing on the study design.
> Title: {title}
> Journal: {journal_title}
> Excerpt: {article_content}"

The article content includes the Introduction and Methods sections, the same as the input used in the extractive and abstractive approaches, BMRetriever and PRIMERA, respectively.

2.3. PT Classification Models

Our previous best model [33] used a PubMedBERT encoder as the base model. In this work, in an effort to enrich article representations, we evaluated several BERT-based models as base encoders for fine-tuning. Specifically, we experimented with the following BERT-based encoders:

- PubMedBERT [32]: pre-trained from scratch on PubMed abstracts and full-text using masked language modeling and next sentence prediction tasks. It was used in previous work and serves as the baseline.

- BioLinkBERT [34]: pre-trained using a document relation prediction task in lieu of next sentence prediction. It leverages citation links between articles for this task.

- SPECTER [35]: incorporates citation information through a custom contrastive loss that pulls together pairs of article representations if one cites the other; otherwise, representations are pushed apart.

- SPECTER2-Base [36]: extends SPECTER, by pre-training on a larger dataset (x10 as big; 6.2M training triplets across 23 fields of study).

- SPECTER2-Clf [36]: SPECTER2-Base pre-fine-tuned for classification. The model is pre-fine-tuned to predict an article's field of study (multi-label task) as well as to predict descriptors (e.g., "Brain","Breast Neoplasms") and qualifiers (e.g., "Complications", "Surgery") from the 30 most frequent top-level MeSH descriptors for articles with exactly one qualifier (multi-class task). This pre-fine-tuning drives acts similar to contrastive learning, driving article representations from similar fields and descriptors closer/further to improve differentiation between classes.

In our previous work [33], the model was trained using binary cross-entropy loss and AdamW optimizer. This model serves as the baseline in this work. We further experiment with other optimization and regularization techniques to improve the training process and mitigate the effect of noisy data. Specifically, we used RAdam (rectified Adam) [54], which uses warm-up in its implementation and was shown to be less sensitive to hyperparameters changes, making tuning less important. Instead of binary cross-entropy loss, we used asymmetric loss [41], an extension of focal loss [55], which aims to down-weight possibly mislabeled data in addition to down-weighting and hard-thresholding easy negative samples to focus the learning process on hard-to-classify examples. This can lead to faster convergence during training and improved performance when trained on noisy data. Additionally, we used label smoothing [42, 43] to further enhance model robustness against noise as well as to better calibrate the model. By taking a small probability from the correct labels and dispersing it among incorrect labels, label smoothing prevents the model from overfitting its predictions to potentially wrong labels. This regularization prevents model over-confidence and enhances performance on noisy datasets. We use the default parameters for asymmetric loss: $\gamma_- = 4$, $\gamma_+ = 1$, $m = 0.05$, and $\epsilon = $ 1e-8, where γ is a weighting term for positive and negative losses, m is the probability shifting hyperparameter and ϵ is a smoothing term. For label smoothing, we use $a = 0.05$, which is the amount of re-distributed probability.

2.4. Contrastive Learning

The basic idea behind contrastive learning is that similar instances (e.g., instances with the same labels) should have representations that are closer within the model embedding space, while dissimilar instances should be further apart. We experimented with contrastive learning (CL) in previous work [33]; however, the results were inconclusive, with minor improvements with unsupervised CL and performance degradation with supervised CL. In this work, we evaluate CL more extensively. As in prior work, we use unsupervised SimCSE method [37] as the CL baseline. In this setting, an instance's positive anchor is a dropout-augmented version of itself. Negative anchors are all other instances within the batch. This unsupervised loss is calculated as follows:

$$L_{unsup} = -\log \frac{e^{f(h_i^{z_i}, h_i^{z'_i})/\tau}}{\sum_{j=1}^{N} e^{f(h_i^{z_i}, h_j^{z_j'})/\tau}} \quad (1)$$

where N is the number of instances within the batch, $h_i^{z_i}$ is the hidden representation of the [CLS] token of an instance, i, $h_i^{z_i\prime}$ is a dropout-augmented representation, $f(\cdot,\cdot)$ is a similarity measurement function (in this case, cosine similarity), and τ is a temperature hyperparameter. As negative representations are weighed equally to calculate loss in SimCSE, outliers (or instances that are significantly different) have a greater effect, which may not be optimal. To mitigate this, ADjusted InfoNCE (ADNCE) [38] weighs negative instances using hyperparameters to focus on potentially more informative instances. This is done by Gaussian-like weighting negative anchors:

$$L_{unsup} = -\log \frac{e^{f(h_i^{z_i}, h_i^{z_i'})/\tau}}{\sum_{j=1}^{N} w e^{f(h_i^{z_i}, h_j^{z_j'})/\tau}} \tag{2}$$

$$w = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(e^{f(h_i^{z_i}, h_j^{z_j'})/\tau} - \mu)^2}{2\sigma^2}\right) \tag{3}$$

where σ controls the weight discrepancy among the samples and μ controls the region of weight allocation (i.e., samples closer to μ have larger weights).

While unsupervised CL focuses on improving uniformity (i.e., how evenly distributed all instances are within the representation space), supervised CL focuses on alignment, in which instances with the same label are brought closer together. In multi-label settings, CL is more difficult to implement as similarity is less clear due to the various levels of overlapping labels that may occur. To account for this, various contrastive loss functions have been proposed [56, 57], although these sometimes require specialized minibatching processes. HeroCon [39], a unified contrastive learning framework incorporating both unsupervised and supervised losses together, introduces a weighted supervised learning framework without the need for any specific batching procedures. The loss is weighed through the Hamming distance of each instance's set of binary labels as a form of supervised similarity. This is shown below:

$$L_{sup} = -\frac{1}{c}\sum_{a=1}^{c} \mathbb{E}_{X_i, X_j \epsilon \mathcal{P}^{\mathcal{L}}(a)}[log \frac{\sigma f(h_i^{z_i}, h_j^{z_j})}{\sigma f(h_i^{z_i}, h_j^{z_j}) + \sum_{x_k \epsilon N^{\mathcal{L}}(a)} \gamma f(h_i^{z_i}, h_k^{z_k})}] \tag{4}$$

$$\sigma = 1 - dist(Y_i^{\mathcal{L}}, Y_j^{\mathcal{L}})/c \tag{5}$$

$$\gamma = dist(Y_i^{\mathcal{L}}, Y_k^{\mathcal{L}}) \tag{6}$$

where $dist(Y_i^{\mathcal{L}}, Y_k^{\mathcal{L}})$ is used to weigh label similarity (i.e., Hamming distance), $\mathcal{P}^{\mathcal{L}}(a) = \{X_j | Y_j^{\mathcal{L}}(a) = 1\}$ is the set of positive instances with the a^{th} label, and $\mathcal{N}^{\mathcal{L}}(a) = \{X_k | Y_k^{\mathcal{L}}(a) \neq 1\}$ is the set of negative instances. Here, a is a PT (e.g., articles in a batch positively labeled as RCTs), c is the number of labels, and γ is the label similarity weighting, in this case, Hamming distance.

Additionally, Zheng et al., demonstrate additive benefits when using both unsupervised and supervised approaches together compared to either approach individually [39]. In supervised CL, HeroCon weights all label differences equally, which may not be optimal in our work. For example, intuitively, representations of cohort studies should probably be more closely related to prospective studies, whereas those of autobiographies should be more related to biographies. Lan et al. [40] introduce a new supervised contrastive loss, WeighCon, which learns optimal weights between labels rather than relying on Hamming distance. This follows the same formulation as above except with a learned weight:

$$\gamma_{ij} = \sigma(NN(y_i, y_j)) \tag{7}$$

where σ is a sigmoid activation function and NN is a one-layer, fully connected linear layer, taking in two label vectors and outputting a scalar value corresponding to the similarity between label vectors.

In efforts to improve these supervised contrastive loss functions, we also applied the asymmetric focusing modulation term from asymmetric loss[41]. This is shown below:

$$w_i = \{(1 - P_i Y_i^{\mathcal{L}}) - [(1 - P_{mi})(1 - Y_i^{\mathcal{L}})]\}^{[(1-Y_i^{\mathcal{L}})\mu_- + Y_i^{\mathcal{L}}\mu_+]} \tag{8}$$

where P_i is the sigmoid activated model prediction for a particular label in instance i, $Y_i^{\mathcal{L}}$ is the true label, $P_m i$ is the probability after asymmetric clipping ($P_{mi} = max(P_i + m, 1)$), and μ_- and μ_+ are focusing parameters. We use default values m=0.05, μ_-=4 and μ_+=1. Overall, this weighting term w_i is used to modulate the HeroCon and WeighCon weighting terms defined in equations 5, 6, and 7 to downweight easy instances and focus on harder ones, which may be more heterogeneous.

Additionally, we experiment with in-batch label correction wherein we consider any instance with a label probability $\geq$ 0.95 to be considered as positively labeled for that class, regardless of its true label, in efforts to better account for false negative labels.

In all experiments with unsupervised and supervised CL, contrastive loss terms were combined with the label loss (i.e., binary cross entropy) by multiplying the CL loss terms with a scaling factor before averaging these losses all together, as shown below:

$$L = mean(\alpha \cdot L_{unsup} + \beta \cdot L_{sup} + L_{ASL}) \tag{9}$$

where α and β are hyperparameters that weigh unsupervised and supervised contrastive loss respectively.

2.5. Experimental Setup

We use the following hyperparameters for model training: batch size (32), learning rate (transformer layers): 1e-4, learning rate (classification layer) (1e-2), epochs (25), optimizer (RAdam), dropout (0.1), label smoothing with $\alpha = 0.05$, asymmetric loss with γ_{-}=4, γ_{+}=1, and m=0.05, and finally, early stopping (when no improvement is observed on validation set macro-F1 over 4 epochs).

Unless the component is being studied or otherwise stated, experiments use these settings with the SPECTER2-Base encoder and ANIMAL MeSH split without contrastive loss (CL). Without considering CL, this combination represented the best performing model. The best CL (WeighCon) was not included to avoid unnecessary computational costs. All models were implemented using PyTorch (v.2.2.0). Model weights downloaded from the HuggingFace transformers library (v.4.40.2).

All contrastive learning experiments used the same temperature parameter ($\tau = 0.05$). We also set both ADNCE hyperparameters μ and σ to 1.0. For HeroCon and WeighCon experiments, we used the hyperparameters that Zheng et al. [39] used with the CelebA dataset, a multi-label dataset with 40 labels, the dataset most similar to ours that was evaluated in that work. Unsupervised CL experiments used $\alpha = 0.01$, supervised CL experiments used $\beta = 0.1$, and combined approaches used $\alpha = 0.1$ and $\beta = 0.1$.

All experiments were conducted on a single Tesla v100 GPU with 32GB of memory.

2.6. Evaluation

We used standard evaluation metrics: precision, recall, and F1 score, as well as area under ROC curve (AUC). Micro- and macro-averaged performance is reported for each metric. Micro-averaged metrics weigh all instances equally, while macro-averaged metrics weigh each class equally. Early stopping was based on macro-averaged F1, and model selection was implemented to maximize macro-averaged F1. This was done instead of loss or micro-F1 to avoid very poor performance on some rare classes. We calculated the optimal probabilistic thresholds for each label to maximize F1 score on the validation set. Using these thresholds, performance was evaluated on the test set. Bootstrap sampling on model predictions with 1,000 replicates was used to generate confidence intervals for each experiment (95% CI). We also calculated core-averaged performance, where "core" consisted of a set of PTs that the authors of this study deemed most important for evidence synthesis. Those PTs were as follows: Case Reports, Case-Control Studies, Clinical Studies as Topic, Clinical Study, Clinical Trial, Clinical Trial Protocol, Cohort Studies, Cross-Over Studies, Cross-Sectional Studies, Double-Blind Method, Evaluation Study, Follow-Up Studies, Longitudinal Studies, Meta-Analysis, Multicenter Study, Prospective Study, Random Allocation, Human Randomized Controlled Trials, Retrospective Studies, Systematic Review, Systematic Reviews as Topic, and Validation Study. Finally, we also calculated expected calibration error (ECE) [58] for experiments varying the loss function to attempt to quantify model calibration. In ECE, the model's probabilistic predictions are sorted and grouped into k bins (k=15 for our work) to calculate a weighted average of each bin's difference between the probabilities and the accuracy. We used three variants of ECE: standard ECE, which uses the absolute value to calculate difference in probability/accuracy (L1); root mean square calibration error, which uses root mean square average (L2); and maximum calibration error, which looks at the maximum difference. Primary comparisons were done using significance testing (one-sided t-test from the SciPy package *ttest_rel*).

For full-text experiments, models were trained and evaluated only on the subset of data where full-text articles were available in PMC Open Access Subset. Full-text experiments did not use optimized thresholding as it was found to harm performance due to the limited amount of full-text available data for some PTs in the validation set. We conducted ablation studies to isolate the impact of some features and architecture design choices on model performance (e.g., different base encoders).

To better understand how the input influences model predictions and enhance model interpretation, we implemented gradient-based saliency mapping [59]. Specifically, we employed Integrated Gradients [60], which assigns importance scores to input features by approximating the integral of the gradients of model outputs with respect to the input. We visualized these scores using the Captum package [61], highlighting keywords that influenced the classification outcome. We compared the model's predictions with the true labels to gain insights into how these terms contributed to both correct and incorrect predictions.

3. Results

3.1. Dataset Statistics

Our dataset included 166,232 articles, which were split into train (70%, n = 116,361), validation (10%, n = 16,624), and test (20%, n = 33,247) sets, while attempting to maintain the PT distribution in each split as seen in the overall dataset. The distribution of the 61 labels across the dataset is shown in Figure 2 (left panel) as well as the performance of the best performing model (right panel), which used SPECTER2-Base embeddings, asymmetric loss with label smoothing, and supervised contrastive loss using WeighCon. Of the 166,232 articles in our dataset, 24,398 made their full-text machine-accessible through the PMC Open Access Subset: 17,111 in training, 2,435 in validation, and 4,852 in testing.

3.2. PT Classification Models with PubMed-only Features

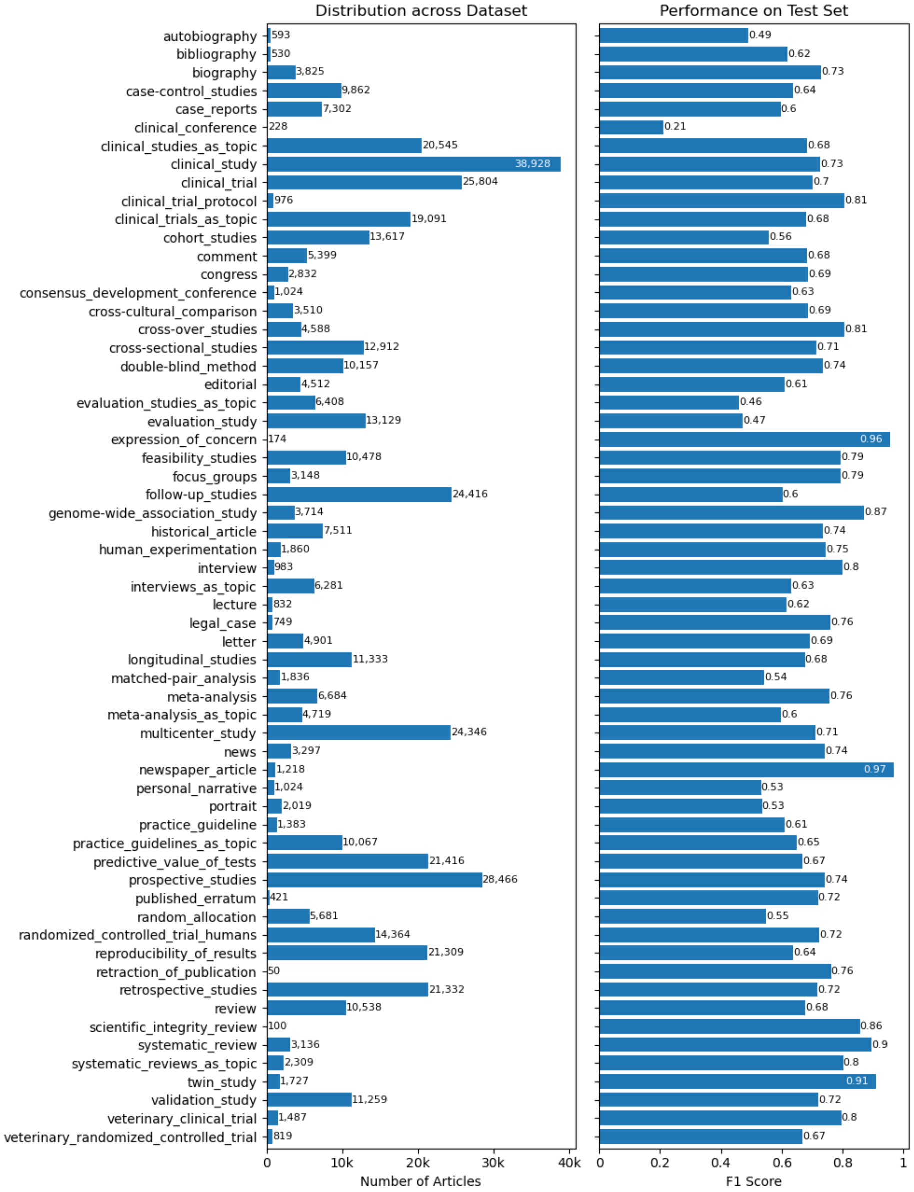
In our primary analysis, we compare the best model and features from our previous work [33] against the best-performing model from this work, which used SPECTER2-Base embeddings, asymmetric loss with label smoothing, and supervised contrastive loss using WeighCon. The results for these two models are provided in Table 3, which shows an improvement across all metrics. The best model in this work was significantly better than the model from previous work (both macro-F1 and micro-F1; $p < 0.0001$). Macro-F1 performance on core PTs were also higher in this work (0.697) compared to the previous work (0.690).

Model	**Precision ↑**		**Recall ↑**		**F_1 ↑**		**AUC ↑**	
	Micro	Macro	Micro	Macro	Micro	Macro	Micro	Macro
Previous architecture [33]	0.652 [0.648-0.655]	0.680 [0.672-0.687]	0.676 [0.673-0.680]	0.660 [0.652-0.670]	0.664 [0.661-0.667]	0.663 [0.656-0.670]	0.972 [0.971-0.972]	0.964 [0.963-0.965]
This work	**0.670 [0.666-0.673]**	**0.700 [0.693-0.707]**	**0.688 [0.685-0.692]**	**0.692 [0.684-0.700]**	**0.679 [0.676-0.682]**	**0.690 [0.683-0.696]**	**0.975 [0.974-0.975]**	**0.968 [0.967-0.969]**

Table 3: Performance comparison of our previous best model architecture and features [33] with the best model and features from this work. 95% CIs were calculated using bootstrap sampling from the test set (n = 33,247).

Ablation study performances are reported in Appendix A. Table A.5 presents the results of the ablation study assessing the impact of different loss functions. All models are trained using the best model as described in Section 2.5. Asymmetric loss variants generally outperformed binary cross

Figure 2: The left panel shows the PT label distribution for all articles in our full dataset. The right panel shows the individual label performances (F1 score) of the best-performing model (SPECTER2-base with WeighCon) on the PubMed test set (n = 33,247).



entropy variants in terms of F1, while performing worse across all expected calibration error (ECE) metrics. Label smoothing, as expected, lowered ECE, indicating better calibrated models.

The results of the ablation study measuring the impact of different base encoders are shown in Table A.6. In these experiments, all models are trained using the best experimental settings described in Section 2.5. We note PubMedBERT, here, was trained without early stopping as it was found to stop very early before it had its best performance. Generally, the citation-informed models outperformed PubMedBERT. There was no significant difference between the performances of the other models (BioLinkBERT, SPECTER, and SPECTER2 variants). SPECTER2-Base performed best overall (micro-F1: 0.679; macro-F1: 0.689).

We also performed experiments to examine the effects of using additional topical MeSH terms as labels (HUMAN, ANIMAL, VETERINARY), label splitting (COHORT-SPLIT and GENERALIZED RCT), as well as a combination of all of these approaches. The results of these experiments are shown in Table A.7. Here, the baseline is the SPECTER2-Base model without any added label splits. Overall, there was little difference between the performances of models trained on these variations; however, adding the "Animals" MeSH term during training had slightly better overall performance. Somewhat surprisingly, combining all approaches resulted in the worst macro-F1 performance, perhaps due to diluting the loss as a smaller proportion of the loss directly steers the model to improve label classification.

The results of our CL experiments are shown in Table A.8. These experiments used the SPECTER2-Base base encoder. Overall, SPECTER2-Base model trained with WeighCon contrastive loss performed best, although it was not significantly better than the other approaches, including training without CL. Performance of the unsupervised approaches (SimCSE, AD-NCE) and supervised approaches (HeroCon, WeighCon) were all roughly similar. All CL generally improved recall at the cost of precision, although this was magnified with unsupervised loss. Incorporating a noise-aware modulation term as well as in-batch label correction for supervised CL did not significantly improve performance. Performance was slightly improved when using these for HeroCon, while slightly worse for WeighCon, perhaps due to the static vs. learned weighting scheme. Combining unsupervised and supervised approaches did not additively improve the model over using these approaches independently.

3.3. PT Classification Models with Full-Text Features

We compared a model trained using PubMed-only features with a model trained using features derived from the full-text in addition to PubMed features. The results of this comparison are shown in Table 4. The best model using full-text features included EXTRACTIVE (MULTIPLE), label sentences, first sentence, NCT identifier information, ethics, number features (rough word count of article, # of tables, and # of figures), and primary section heading features. This model outperformed the baseline model using PubMed-only features (title, abstract, and metadata) across both micro- and macro-F1 metrics ($p < 0.0001$). Note that overall performances of both models are lower than those shown in Table 3, because these experiments used a smaller dataset with full-text articles only (17,111 training instances and and 4,852 test instances). Because of this smaller dataset, there were also cases where certain PTs had relatively few (or in some cases 0) instances within the validation or test set. Thus, we believe that micro-F1 should be given stronger consideration than macro-F1 when comparing full-text models in this work.

Features	**Precision ↑**		**Recall ↑**		**F_1 ↑**		**AUC ↑**	
	Micro	Macro	Micro	Macro	Micro	Macro	Micro	Macro
PubMed-only	0.505 [0.497-0.513]	0.485 [0.468-0.504]	0.791 [0.783-0.799]	0.684 [0.658-0.707]	0.616 [0.610-0.623]	0.556 [0.539-0.574]	0.972 [0.971-0.974]	0.938 [0.925-0.947]
Best full-text	**0.516** **[0.509-0.524]**	**0.489** **[0.471-0.509]**	**0.812** **[0.805-0.820]**	**0.702** **[0.676-0.724]**	**0.631** **[0.625-0.638]**	**0.566** **[0.548-0.584]**	**0.975** **[0.974-0.977]**	**0.943** **[0.931-0.952]**

Table 4: Performance comparison of models trained with PubMed-only features and with the inclusion of best full-text-based features.

We performed ablation studies to better isolate the impact of each feature. The results of these experiments on the full-text test set are provided in Table B.9. Removing the label sentences (sentences that contain one of the PT mentions) and TextRank had the greatest impact on performance (i.e., micro- and macro-F1). Somewhat surprisingly, the best combination of features determined with the validation set (micro-F1 in validation: 0.639; test: 0.631) did not perform best on the test set (removing first sentence - micro-F1 validation: 0.635; test: 0.632), which may indicate a need to evaluate on larger sets of full-text articles, which may result in more stable and generalizable feature selection. These results are also influenced by the 512 token limit inherent in BERT-based models as this combination of features generally exceeds this, requiring truncation.

3.4. Impact of Summarization of Full-Text Content

The results for summarization-specific experiments are detailed in Table B.10, where the PubMed-only features serve as the baseline approach. The integration of TextRank (B + TextRank) slightly improves the performance in terms of micro-F1 (0.619 vs. 0.616) and Micro-AUC (0.973 vs. 0.972) over the baseline. Incorporating summaries across all methods consistently enhances micro-F1 compared to the baseline, with the abstractive method B + primera yielding the highest improvement, increasing the micro-F1 score from 0.616 to 0.632. Additionally, it achieves the highest Micro- and Macro-AUC scores (0.974 and 0.941, respectively), outperforming the baseline (0.972 and 0.938, respectively). This indicates better discrimination and an improved ability to differentiate between classes.

However, the impact on macro-F1 is less consistent. Most summarization methods do not improve macro-F1 compared to the baseline, except for B + extractive (single), which shows a marginal increase from 0.556 to 0.558. This suggests that summaries enhance performance mostly on the frequent labels. Given the highly imbalanced nature of our dataset, the improvements in micro-F1 may be attributed to overfitting on the majority classes, leading to slight degradations in macro-F1 for some methods, including B + TextRank, B + llm, and B + primera.

4. Discussion

In this study, we aimed to improve PT classification performance by improving article representations, including full-text features, creating more homogeneous labels and accounting for label noise through advanced optimization techniques. Our best model improved on models reported in previous work [33], increasing macro-F1 from 0.663 to 0.690 and micro-F1 from 0.664 to 0.679 when evaluated on 61 labels across 33,247 articles. Furthermore, by adding full-text information, we were able to improve model performance on the full-text subset (n = 4,852) of the test set [micro-F1: 0.616 to 0.631; macro-F1 = 0.556 to 0.566]. While performance using the larger PubMed dataset is reasonably high for PTs such as Systematic Review (0.90 F1) and Genome-wide Association Study (0.87), it is lower for some important PTs (e.g., Random Allocation (0.55) and Cohort Studies (0.56)), indicating that PT classification remains a challenging task.

4.1. PT Classification Models with PubMed-only features

In prior work, we only used PubMed features (title, abstract, and metadata) as input [33]. Some of our experiments in this study used the same feature set so that we could assess whether enriched article representations, more fine-grained labels, and optimized training could enhance model performance. Our results suggest that a base encoder optimized for citation-aware biomedical document representation without any pre-fine-tuning (SPECTER2-base) provides enriched representations beneficial for PT classification. Additionally, our hypothesis that citation links between articles could provide additional signal for PT classification did seem to hold true with citation-aware models (BioLinkBERT, SPECTER, and SPECTER2 variants) generally outperforming PubMedBERT. The largest improvements appear to be in non-research related PTs, e.g., legal cases (PubMedBERT F1: 0.79 vs. SPECTER2-base F1: 0.97), however, there are some other research oriented PTs that improve as well (PubMedBERT to SPECTER2-base: predictive value of tests (+10.9 F1) and veterinary RCTs (+8.0 F1)). Our "core" metric, which is a macro-average of clinically important PTs, improved slightly as well (core-F1: PubMedBERT (0.668) vs. SPECTER2-base (0.697)). Future work could explore other auxiliary adaptive pre-training tasks that might benefit PT classification.

Noting the heterogeneity of some PTs (e.g., Cohort Studies, Clinical Study), we also tried to improve the quality of article representations in fine-tuning by splitting labels into more homogeneous classes as well as introducing topical labels for study populations. This was done to explicitly teach the model relationships between labels as well as correlations between populations and study design. However, benefits from this additional knowledge were negligible, which suggests that the model might be already learning to attend to information about study populations and leveraging label correlations in making the predictions.

We observed some benefits from optimizing training through the use of asymmetric loss, label smoothing, and CL. We explored asymmetric loss and label smoothing for regularization and mitigating the effect of noisy data on model performance. The preliminary results showed improvements with these optimizations (micro-F1: BCE (0.673) vs. ASL (0.679)). As another added benefit, label smoothing has been shown to calibrate predictions during training [43]. Our results showed similar results with label smoothing comparisons having lower ECE metrics than non-label smoothing experiments. ASL variants had much larger ECE, indicating worse calibration.

Based on the probability distribution shown in Appendix A.4, this is most likely related to the probability shifting in ASL. Future work may aim to address this in order to improve calibration of models using ASL.

Training with contrastive loss provided negligible benefits. Models trained with WeighCon had slightly higher performance overall, but these differences were not significant. This may suggest that the contrastive learning performed during pre-training within SPECTER2-base already achieves a somewhat optimal state for this task. Additionally, unlike in previous work focusing on computer vision [39], we did not see improvements by combining unsupervised and supervised CL approaches. Noise-aware weighting and in-batch label correction did not have a major impact as well. Given that CL adds significant training overhead and its benefits seem relatively limited for multi-label PT classification with this base encoder, models trained in the future may focus on other approaches to improve PT tagging, such as improving data quality.

4.2. PT Classification Models with Full-Text Features

Overall, we were able to improve model performance by adding full-text information. The difference in model performance with full-text features was statistically significant, and aligned with our general assumption that detailed information related to study design may not always be present in an article's abstract. The most impactful features during ablation studies included label sentences and TextRank. Most features were extracted using simple regular expressions, and further refinements to these expressions could show benefits. For example, a more systematic approach to identify study design information (e.g., information extraction of methodological characteristics [62] rather than simply using sentences with PT label mentions) could lead to further improvements. At the class-level, performance improved on certain PTs, perhaps highlighting that information may only be present in the full-text. For example, longitudinal studies (0.687 → 0.737), double-blind method (0.645 → 0.691), multicenter study (0.564 → 0.601), and human RCTs (0.629 → 0.671) all improved when using the best full-text features.

The improvements from incorporating summaries in full-text experiments over the baseline using PubMed-only features demonstrate the effectiveness of extractive, abstractive, and LLM-based approaches. Among them, the combination of EXTRACTIVE (MULTIPLE) with other full-text features achieved the best overall performance (Table B.9). This result aligns with expectations, as study designs in biomedical research (e.g., Randomized Controlled Trial, Cross-Sectional Study) are often explicitly mentioned in the Methods section. Extractive methods are particularly effective at identifying and directly retrieving these key phrases or sentences, which explains their strong performance. In contrast, when augmenting the PubMed-only features with individual summaries, B + PRIMERA achieved the best performance (Table B.10). This could be attributed to context length restrictions in our model. Our data analysis reveals that, in some cases, the generated summaries exceed 512 tokens, with an average length of 251 tokens for B + PRIMERA and 263 tokens for B + LLM, leading to truncation when appended to the PubMed-only features. This is even more problematic when using these summaries in conjunction with other full-text features as summaries were added as the last feature in the input (i.e., they are the first to be truncated). Extractive models re-rank retrieved sentences, placing the most relevant sentences first. This reduces the likelihood of losing key study design features, especially compared to abstractive summaries, which may not place the most useful information first. Future work may explore the effect of feature ordering on performance.

Additionally, our abstractive summarization approach relied on off-the-shelf models without task-specific adaptation. Fine-tuning an abstractive model on a dataset explicitly annotated with study design information could potentially enhance performance by reducing hallucinations and paraphrasing inaccuracies. However, to our knowledge, no sufficiently large dataset of this kind currently exists. Still, we believe that summarization method could be particularly helpful in three key scenarios: (1) when study design details are not explicitly stated but can be inferred from broader contextual clues; (2) when PubMed articles lack abstracts but have full-text; and (3) when the context size exceeds token limitations, summarization methods may help distill relevant details more effectively.

Despite the improved performance with full-text features, it is important to note that full-text content is often not machine-accessible. Our experiments focused on a relatively small subset of articles with full-text content available on PMC Open Access Subset for XML download. Some articles available in PDF format could be processed with PDF-to-text conversion tools, such as Grobid, and the models could be applied to the extracted text. However, a general solution to classify PTs for all biomedical articles using their full-text remains an open research area. We also note that NLM indexers typically have access to full-text during manual curation.

4.3. Dataset Issues

One concern with our models is the use of curated PT terms in PubMed as the ground truth. On one hand, these labels are manually assigned by NLM indexing experts and they are arguably hard to improve upon. Also given its scale, this dataset is an attractive resource for training broad-coverage PT classification models. On the other hand, previous work highlights the potential noisiness of the PT terms in PubMed [63, 29]. This noise could be caused by the nature of the task (multi-label with many labels), inconsistencies between experts, ambiguity in label descriptions, and semantic drift of the labels over time. Our preliminary analysis of the model outputs also highlights some potentially correct predictions that do not match the PT terms in PubMed. While some datasets have been manually annotated specifically to train NLP models for PT classification [31], these are limited in size and focus. We tried to partially address this issue purely as an optimization problem; however, future work should focus on a comprehensive qualitative evaluation of the model output to understand the extent and patterns of the noise in the PT terms and devise (semi-)automatic methods to improve label quality.

4.4. Error Analysis and Model Interpretation

We examined the most frequently misclassified PT pairs within some of our models. The most common misclassifications included false positive predictions of Historical Article when the true label is Biography (n = 586), Clinical Trials as Topic when it is Clinical Studies as Topic (n = 2,777), and Clinical Trial when it is a Clinical Study (n = 4,010). These PT pairs make sense intuitively as they are all hierarchically related within MeSH and are often predicted together. Unlike NLM, which assigns only the most specific tag and relies on back-end processing to retrieve broader results, our approach explicitly assigns all applicable tags to each paper, in efforts to be more consistent and comprehensive. Surprisingly, this false positive co-occurrence only occurred a single time with Human RCTs and Veterinary RCTs, indicating that these two PTs were not commonly confused. This was also the case within our model architecture from prior work, so this was not due to any additional MeSH term injection. However, it does highlight the model's ability to differentiate between research studying different species.

Future work may further explore our model's ability to differentiate between human, pre-clinical animal, and veterinary studies.

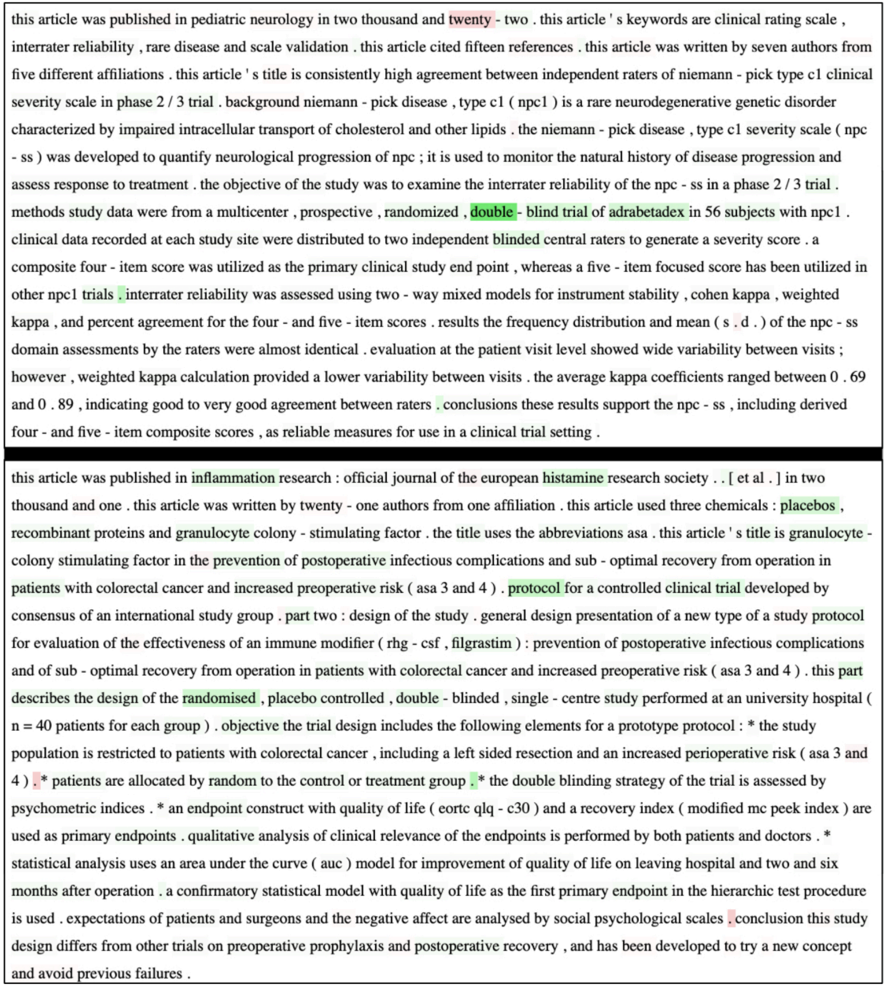
To better understand how the model makes predictions, we generated visual representations from example instances, highlighting keywords that influenced the classification outcomes. Figure 3 shows the gradient-based saliency mappings of two instances. Words in green have a positive association with predicting "True", while red words have a negative association. As an example of what we would expect, the top panel in Figure 3 shows a correctly predicted instance of Double-Blind Method. Below, we show a false positive instance of Human RCTs. We can see that the article is actually an RCT protocol. This most likely occurred due to inconsistencies within NLM indexing. RCT protocols were previously tagged as Human RCTs instead of as Clinical Trial Protocols. NLM indexing is not consistent since their tagging before 2019 and after 2019 changed [64]. Over time, new PTs have been added or their definitions changed. It is likely that our model is more consistent and accurate compared to NLM indexing, although the performance might still be improved by removing RCT protocol articles from the RCT training set where possible in future work.

4.5. Limitations

There are several limitations related to this study. The major challenge relates to the noise in the dataset, as discussed above. Despite acknowledging the issue and trying to mitigate it using optimization techniques, we still need to use a somewhat noisy test set for our evaluation. Therefore, there are cases where the model prediction is correct and is evaluated as incorrect (and vice versa). Future work needs to better understand the extent and patterns of noise in the PubMed PTs and try to establish a true gold standard, at least for evaluation. Second, while our list of publication types and study designs is extensive, it is not exhaustive. Our current model does not classify some PTs initially deemed to have relatively little utility within the biomedical community (e.g., Directory, Journal Article). At the same time, PubMed PTs are also not exhaustive (e.g., Case Series is not a PubMed PT). In future work, we aim to expand our classification scheme to include all PTs indexed in PubMed as well as new PTs that would serve unmet needs of the diverse biomedical research community.

Our base models are based on encoder-only architectures. In this work, we have not fully explored autoregressive LLMs, which have shown impressive performance on some NLP tasks in recent years, except using one for

Figure 3: Gradient-based saliency maps for two instances within the test set. Green indicates a positive association with the model predicting a specific label, while red indicates a negative association with a particular label. The first article is PMID: 34952292 and the second is PMID:11392607.



this article was published in pediatric neurology in two thousand and twenty - two . this article ' s keywords are clinical rating scale , interrater reliability , rare disease and scale validation . this article cited fifteen references . this article was written by seven authors from five different affiliations . this article ' s title is consistently high agreement between independent raters of niemann - pick type c1 clinical severity scale in phase 2 / 3 trial . background niemann - pick disease , type c1 (npc1) is a rare neurodegenerative genetic disorder characterized by impaired intracellular transport of cholesterol and other lipids . the niemann - pick disease , type c1 severity scale (npc - ss) was developed to quantify neurological progression of npc ; it is used to monitor the natural history of disease progression and assess response to treatment . the objective of the study was to examine the interrater reliability of the npc - ss in a phase 2 / 3 trial . methods study data were from a multicenter , prospective , randomized , double - blind trial of adrabetadex in 56 subjects with npc1 . clinical data recorded at each study site were distributed to two independent blinded central raters to generate a severity score . a composite four - item score was utilized as the primary clinical study end point , whereas a five - item focused score has been utilized in other npc1 trials . interrater reliability was assessed using two - way mixed models for instrument stability , cohen kappa , weighted kappa , and percent agreement for the four - and five - item scores . results the frequency distribution and mean (s . d .) of the npc - ss domain assessments by the raters were almost identical . evaluation at the patient visit level showed wide variability between visits ; however , weighted kappa calculation provided a lower variability between visits . the average kappa coefficients ranged between 0 . 69 and 0 . 89 , indicating good to very good agreement between raters . conclusions these results support the npc - ss , including derived four - and five - item composite scores , as reliable measures for use in a clinical trial setting .

this article was published in inflammation research : official journal of the european histamine research society . . [et al .] in two thousand and one . this article was written by twenty - one authors from one affiliation . this article used three chemicals : placebos , recombinant proteins and granulocyte colony - stimulating factor . the title uses the abbreviations asa . this article ' s title is granulocyte - colony stimulating factor in the prevention of postoperative infectious complications and sub - optimal recovery from operation in patients with colorectal cancer and increased preoperative risk (asa 3 and 4) . protocol for a controlled clinical trial developed by consensus of an international study group . part two : design of the study . general design presentation of a new type of a study protocol for evaluation of the effectiveness of an immune modifier (rhg - csf , filgrastim) : prevention of postoperative infectious complications and of sub - optimal recovery from operation in patients with colorectal cancer and increased preoperative risk (asa 3 and 4) . this part describes the design of the randomised , placebo controlled , double - blinded , single - centre study performed at an university hospital (n = 40 patients for each group) . objective the trial design includes the following elements for a prototype protocol : * the study population is restricted to patients with colorectal cancer , including a left sided resection and an increased perioperative risk (asa 3 and 4) . * patients are allocated by random to the control or treatment group . * the double blinding strategy of the trial is assessed by psychometric indices . * an endpoint construct with quality of life (eortc qlq - c30) and a recovery index (modified mc peek index) are used as primary endpoints . qualitative analysis of clinical relevance of the endpoints is performed by both patients and doctors . * statistical analysis uses an area under the curve (auc) model for improvement of quality of life on leaving hospital and two and six months after operation . a confirmatory statistical model with quality of life as the first primary endpoint in the hierarchic test procedure is used . expectations of patients and surgeons and the negative affect are analysed by social psychological scales . conclusion this study design differs from other trials on preoperative prophylaxis and postoperative recovery , and has been developed to try a new concept and avoid previous failures .

abstractive summarization in full-text experiments. Our initial explorations with zero-shot and few-shot learning using such models yielded poor results; however, we aim to explore this direction to provide a more comprehensive comparison with encoder-only models.

5. Conclusions

In this study, we trained and validated Transformer-based models for PT classification on a dataset of biomedical articles labeled with 61 PTs using a combination of PubMed queries and indexing terms. Specifically, we compared different base encoders for article representation, investigated whether more fine-grained labels created through label splitting and injecting relevant MeSH terms could enhance performance, and experimented with regularization and optimization techniques for enhanced training, reducing overfitting, and mitigating label noise to some extent. Additionally, we investigated the use of full-text features within PT classification. Our model performance improves upon that reported in previous work [33]; in particular, our results demonstrate the value of incorporating enriched article representations and full-text information into automated PT classification models. This model, while imperfect, could help to ensure consistent PT indexing within repositories like PubMed and beyond when deployed (e.g., Semantic Scholar, etc.). Future work will analyze model output more extensively through manual analysis, continue to expand to include PTs not currently considered, as well as deploy these models into existing systems (e.g., Anne O'Tate [65]) to increase accessibility to both academia and research repositories. We will also explore autoregressive LLMs with larger context sizes in more depth for PT classification.

Data Availability

All article data is publicly available from PubMed or PubMed Central through the open archives initiative. Label data (including PMIDs) as well as all code necessary to replicate experiments is available on Github (`https://github.com/ScienceNLP-Lab/MultiTagger-v2`).

Acknowledgments

This work was supported by the National Library of Medicine of the National Institutes of Health under the award number R01LM14292. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The funder had no role in considering the study design or in the collection, analysis, interpretation of data, writing of the report, or decision to submit the article for publication.

Author Contributions

Joe D. Menke: Conceptualization, Methodology, Software, Visualization, Writing – original draft, and Writing – review and editing; Shufan Ming: Conceptualization, Methodology, Software, Visualization, Writing – original draft, and Writing – review and editing; Shruthan Radhakrishna: Software; Halil Kilicoglu: Conceptualization, Funding acquisition, Methodology, Supervision, Writing – original draft, and Writing – review and editing; Neil R. Smalheiser: Conceptualization, Funding acquisition, Supervision, and Writing – review and editing.

References

[1] D. L. Sackett, W. M. Rosenberg, J. M. Gray, R. B. Haynes, W. S. Richardson, Evidence based medicine: what it is and what it isn't, BMJ 312 (7023) (1996) 71–72.

[2] A. M. Cohen, C. E. Adams, J. M. Davis, C. Yu, P. S. Yu, W. Meng, L. Duggan, M. McDonagh, N. R. Smalheiser, Evidence-based medicine, the essential role of systematic reviews, and the need for automated text mining tools, in: Proceedings of the 1st ACM International Health Informatics Symposium, 2010, pp. 376–380.

[3] S. Khangura, K. Konnyu, R. Cushman, J. Grimshaw, D. Moher, Evidence summaries: the evolution of a rapid review approach, Systematic Reviews 1 (1) (2012) 1–9.

[4] J. Clark, P. Glasziou, C. Del Mar, A. Bannach-Brown, P. Stehlik, A. M. Scott, A full systematic review was completed in 2 weeks using automation tools: a case study, Journal of Clinical Epidemiology 121 (2020) 81–90.

[5] National Library of Medicine (US), Incorporating Values for Indexing Method in MEDLINE/PubMed XML, NLM Tech Bulletin e2 (2018) 423, accessed on 02.26.2024.
URL nlm.nih.gov/pubs/techbull/ja18/ja18_indexing_method.html

[6] A. R. Aronson, J. G. Mork, C. W. Gay, S. M. Humphrey, W. J. Rogers, The NLM indexing initiative's medical text indexer, in: MEDINFO 2004, IOS Press, 2004, pp. 268–272.

[7] J. G. Mork, A. Jimeno-Yepes, A. R. Aronson, et al., The NLM Medical Text Indexer System for Indexing Biomedical Literature, BioASQ@ CLEF 1 (2013).

[8] J. Mork, A. Aronson, D. Demner-Fushman, 12 years on–Is the NLM medical text indexer still useful and relevant?, Journal of Biomedical Semantics 8 (1) (2017) 1–10.

[9] A. R. Rae, J. G. Mork, D. Demner-Fushman, A Neural Text Ranking Approach for Automatic MeSH Indexing., in: CLEF (Working Notes), 2021, pp. 302–312.

[10] National Library of Medicine (US), Frequently Asked Questions about Indexing for MEDLINE, accessed on 02.26.2024 (2010).
URL nlm.nih.gov/bsd/indexfaq.html

[11] G. Tsatsaronis, G. Balikas, P. Malakasiotis, I. Partalas, M. Zschunke, M. R. Alvers, D. Weissenborn, A. Krithara, S. Petridis, D. Polychronopoulos, et al., An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition, BMC Bioinformatics 16 (1) (2015) 1–28.

[12] A. Nentidis, A. Krithara, K. Bougiatiotis, M. Krallinger, C. Rodriguez-Penagos, M. Villegas, G. Paliouras, Overview of BioASQ 2020: The eighth bioASQ challenge on large-scale biomedical semantic indexing and question answering, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction: 11th International Conference of the CLEF Association, CLEF 2020, Thessaloniki, Greece, September 22–25, 2020, Proceedings 11, Springer, 2020, pp. 194–214.

[13] Overview of B]ioASQ 2023: The eleventh bioASQ challenge on large-scale biomedical semantic indexing and question answering, author=Nentidis, Anastasios and Katsimpras, Georgios and Krithara, Anastasia and Lima López, Salvador and Farré-Maduell, Eulália and Gasco, Luis and Krallinger, Martin and Paliouras, Georgios, booktitle=International Conference of the Cross-Language Evaluation Forum for European Languages, pages=227–250, year=2023, organization=Springer.

[14] K. Liu, S. Peng, J. Wu, C. Zhai, H. Mamitsuka, S. Zhu, MeSHLabeler: improving the accuracy of large-scale MeSH indexing by integrating diverse evidence, Bioinformatics 31 (12) (2015) i339–i347.

[15] S. Peng, R. You, H. Wang, C. Zhai, H. Mamitsuka, S. Zhu, DeepMeSH: deep semantic representation for improving large-scale MeSH indexing, Bioinformatics 32 (12) (2016) i70–i79.

[16] R. You, Y. Liu, H. Mamitsuka, S. Zhu, BERTMeSH: deep contextual representation learning for large-scale high-performance MeSH indexing with full text, Bioinformatics 37 (5) (2021) 684–692.

[17] S. Dai, R. You, Z. Lu, X. Huang, H. Mamitsuka, S. Zhu, FullMeSH: improving large-scale MeSH indexing with full text, Bioinformatics 36 (5) (2020) 1533–1541.

[18] G. Xun, K. Jha, Y. Yuan, Y. Wang, A. Zhang, MeSHProbeNet: a self-attentive probe net for MeSH indexing, Bioinformatics 35 (19) (2019) 3794–3802.

[19] X. Wang, R. Mercer, F. Rudzicz, KenMeSH: Knowledge-enhanced End-to-end Biomedical Text Labelling, in: S. Muresan, P. Nakov, A. Villavicencio (Eds.), Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 2941–2951. doi:10.18653/v1/2022.acl-long.210.

[20] K. Knight, S. Wade, L. Balducci, Prevalence and outcomes of anemia in cancer: a systematic review of the literature, The American Journal of Medicine 116 (7) (2004) 11–26.

[21] N. L. Wilczynski, R. B. Haynes, Consistency and accuracy of indexing systematic review articles and meta-analyses in medline, Health Information & Libraries Journal 26 (3) (2009) 203–210.

[22] L. Peirson, D. Fitzpatrick-Lewis, D. Ciliska, R. Warren, Screening for cervical cancer: a systematic review and meta-analysis, Systematic Reviews 2 (2013) 1–14.

[23] X. I. Yao, X. Wang, P. J. Speicher, E. S. Hwang, P. Cheng, D. H. Harpole, M. F. Berry, D. Schrag, H. H. Pang, Reporting and guidelines in propensity score analysis: a systematic review of cancer and cancer surgical studies, JNCI: Journal of the National Cancer Institute 109 (8) (2017) djw323.

[24] A. M. Cohen, N. R. Smalheiser, M. S. McDonagh, C. Yu, C. E. Adams, J. M. Davis, P. S. Yu, Automated confidence ranked classification of randomized controlled trial articles: an aid to evidence-based medicine, Journal of the American Medical Informatics Association 22 (3) (2015) 707–717.

[25] B. C. Wallace, A. Noel-Storr, I. J. Marshall, A. M. Cohen, N. R. Smalheiser, J. Thomas, Identifying reports of randomized controlled trials (RCTs) via a hybrid machine learning and crowdsourcing approach, Journal of the American Medical Informatics Association 24 (6) (2017) 1165–1168.

[26] I. J. Marshall, A. Noel-Storr, J. Kuiper, J. Thomas, B. C. Wallace, Machine learning for identifying randomized controlled trials: an evaluation and practitioner's guide, Research Synthesis Methods 9 (4) (2018) 602–614.

[27] N. R. Smalheiser, A. M. Cohen, Design of a generic, open platform for machine learning-assisted indexing and clustering of articles in PubMed, a biomedical bibliographic database, Data and Information Management 2 (1) (2018) 27–36.

[28] A. M. Cohen, J. Schneider, Y. Fu, M. S. McDonagh, P. Das, A. W. Holt, N. R. Smalheiser, Fifty ways to tag your pubtypes: Multi-tagger, a set of probabilistic publication type and study design taggers to support biomedical indexing and evidence-based medicine, medRxiv (2021) 2021–07.

[29] J. Schneider, L. Hoang, Y. Kansara, A. M. Cohen, N. R. Smalheiser, Evaluation of publication type tagging as a strategy to screen randomized controlled trial articles in preparing systematic reviews, JAMIA Open 5 (1) (2022) ooac015.

[30] R. Proescholdt, T.-K. Hsiao, J. Schneider, A. M. Cohen, M. S. McDonagh, N. R. Smalheiser, Testing a filtering strategy for systematic reviews: evaluating work savings and recall, AMIA Summits on Translational Science Proceedings 2022 (2022) 406.

[31] M. Neves, A. Klippert, F. Knöspel, J. Rudeck, A. Stolz, Z. Ban, M. Becker, K. Diederich, B. Grune, P. Kahnau, et al., Automatic classification of experimental models in biomedical literature to support searching for alternative methods to animal experiments, Journal of Biomedical Semantics 14 (1) (2023) 13.

[32] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, H. Poon, Domain-specific language model pretraining for biomedical natural language processing, ACM Transactions on Computing for Healthcare (HEALTH) 3 (1) (2021) 1–23.

[33] J. Menke, H. Kilicoglu, N. Smalheiser, Publication type tagging using transformer models and multi-label classification, in: AMIA Annual Symposium Proceedings, American Medical Informatics Association, 2024.

[34] M. Yasunaga, J. Leskovec, P. Liang, LinkBERT: Pretraining Language Models with Document Links, in: S. Muresan, P. Nakov, A. Villavicencio (Eds.), Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 8003–8016. doi:10.18653/v1/2022.acl-long.551.
URL aclanthology.org/2022.acl-long.551

[35] 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. doi:10.18653/v1/2020.acl-main.207.
URL aclanthology.org/2020.acl-main.207

[36] A. Singh, M. D'Arcy, A. Cohan, D. Downey, S. Feldman, SciRepEval: A Multi-Format Benchmark for Scientific Document Representations, in: H. Bouamor, J. Pino, K. Bali (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Singapore, 2023, pp. 5548–5566. doi:10.18653/v1/2023.emnlp-main.338.
URL aclanthology.org/2023.emnlp-main.338

[37] T. Gao, X. Yao, D. Chen, SimCSE: Simple Contrastive Learning of Sentence Embeddings, in: M.-F. Moens, X. Huang, L. Specia, S. W.-t. Yih (Eds.), Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 6894–6910.

[38] J. Wu, J. Chen, J. Wu, W. Shi, X. Wang, X. He, Understanding contrastive learning via distributionally robust optimization, Advances in Neural Information Processing Systems 36 (2024).

[39] L. Zheng, J. Xiong, Y. Zhu, J. He, Contrastive learning with complex heterogeneity, in: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2022, pp. 2594–2604.

[40] M. Lan, L. Zheng, S. Ming, H. Kilicoglu, Multi-label Sequential Sentence Classification via Large Language Model, in: Findings of the Association for Computational Linguistics: EMNLP 2024, 2024, pp. 16086–16104.

[41] T. Ridnik, E. Ben-Baruch, N. Zamir, A. Noy, I. Friedman, M. Protter, L. Zelnik-Manor, Asymmetric loss for multi-label classification, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 82–91.

[42] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2818–2826.

[43] R. Müller, S. Kornblith, G. E. Hinton, When does label smoothing help?, Advances in Neural Information Processing systems 32 (2019).

[44] National Center for Biotechnology Information (US), Entrez Programming Utilities HelpEntrez Programming Utilities Help, accessed on 01.10.2024 (2010).
URL ncbi.nlm.nih.gov/books/NBK25501/

[45] N. F. Liu, K. Lin, J. Hewitt, A. Paranjape, M. Bevilacqua, F. Petroni, P. Liang, Lost in the middle: How language models use long contexts, Transactions of the Association for Computational Linguistics 12 (2024) 157–173.

[46] J. Maynez, S. Narayan, B. Bohnet, R. McDonald, On Faithfulness and Factuality in Abstractive Summarization, 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. 1906–1919. doi:10.18653/v1/2020.acl-main.173.
URL aclanthology.org/2020.acl-main.173/

[47] R. Mihalcea, P. Tarau, TextRank: Bringing Order into Text, in: D. Lin, D. Wu (Eds.), Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Barcelona, Spain, 2004, pp. 404–411.
URL aclanthology.org/W04-3252/

[48] R. Xu, W. Shi, Y. Yu, Y. Zhuang, Y. Zhu, M. D. Wang, J. C. Ho, C. Zhang, C. Yang, BMRetriever: Tuning Large Language Models as Better Biomedical Text Retrievers, in: Y. Al-Onaizan, M. Bansal, Y.-N. Chen (Eds.), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 22234–22254. doi:10.18653/v1/2024.emnlp-main.1241.
URL aclanthology.org/2024.emnlp-main.1241/

[49] I. Beltagy, M. E. Peters, A. Cohan, Longformer: The long-document transformer, arXiv preprint arXiv:2004.05150 (2020).

[50] A. Cohan, F. Dernoncourt, D. S. Kim, T. Bui, S. Kim, W. Chang, N. Goharian, A Discourse-Aware Attention Model for Abstractive Summarization of Long Documents, in: M. Walker, H. Ji, A. Stent (Eds.), Proceedings of the 2018 Conference of the North American Chapter of the Asso-

ciation for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 615–621. doi:10.18653/v1/N18-2097.
URL aclanthology.org/N18-2097/

[51] W. Xiao, I. Beltagy, G. Carenini, A. Cohan, PRIMERA: Pyramid-based Masked Sentence Pre-training for Multi-document Summarization, in: S. Muresan, P. Nakov, A. Villavicencio (Eds.), Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 5245–5263. doi:10.18653/v1/2022.acl-long.360.
URL aclanthology.org/2022.acl-long.360/

[52] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan, et al., The Llama 3 Herd of Models, arXiv preprint arXiv:2407.21783 (2024).

[53] J. Vig, A. Fabbri, W. Kryscinski, C.-S. Wu, W. Liu, Exploring Neural Models for Query-Focused Summarization, in: M. Carpuat, M.-C. de Marneffe, I. V. Meza Ruiz (Eds.), Findings of the Association for Computational Linguistics: NAACL 2022, Association for Computational Linguistics, Seattle, United States, 2022, pp. 1455–1468. doi:10.18653/v1/2022.findings-naacl.109.
URL aclanthology.org/2022.findings-naacl.109/

[54] L. Liu, H. Jiang, P. He, W. Chen, X. Liu, J. Gao, J. Han, On the Variance of the Adaptive Learning Rate and Beyond, in: Proceedings of the Eighth International Conference on Learning Representations (ICLR 2020), 2020.

[55] T.-Y. Ross, G. Dollár, Focal loss for dense object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2980–2988.

[56] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, D. Krishnan, Supervised Contrastive Learning, Advances in Neural Information Processing Systems 33 (2020) 18661–18673.

[57] N. Lin, G. Qin, G. Wang, D. Zhou, A. Yang, An effective deployment of contrastive learning in multi-label text classification, in: Findings of the Association for Computational Linguistics: ACL 2023, 2023, pp. 8730–8744.

[58] M. Pakdaman Naeini, G. Cooper, M. Hauskrecht, Obtaining Well Calibrated Probabilities Using Bayesian Binning, Proceedings of the AAAI Conference on Artificial Intelligence 29 (1) (Feb. 2015). doi:10.1609/aaai.v29i1.9602.
URL `ojs.aaai.org/index.php/AAAI/article/view/9602`

[59] J. Bastings, K. Filippova, The elephant in the interpretability room: Why use attention as explanation when we have saliency methods?, in: Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP, 2020, pp. 149–155.

[60] M. Sundararajan, A. Taly, Q. Yan, Axiomatic attribution for deep networks, in: International Conference on Machine Learning, PMLR, 2017, pp. 3319–3328.

[61] N. Kokhlikyan, V. Miglani, M. Martin, E. Wang, B. Alsallakh, J. Reynolds, A. Melnikov, N. Kliushkina, C. Araya, S. Yan, O. Reblitz-Richardson, Captum: A unified and generic model interpretability library for PyTorch (2020). arXiv:2009.07896.

[62] L. Hoang, Y. Guan, H. Kilicoglu, Methodological information extraction from randomized controlled trial publications: a pilot study, in: AMIA Annual Symposium Proceedings, Vol. 2022, 2022, p. 542.

[63] A. M. Cohen, Z. O. Dunivin, N. R. Smalheiser, A probabilistic automated tagger to identify human-related publications, Database 2018 (2018) bay079.

[64] S. Tybaert. MEDLINE Data Changes—2019. NLM Tech Bull. 2018 Nov-Dec;(425):e4a.

[65] N. R. Smalheiser, D. P. Fragnito, E. E. Tirk, Anne O'Tate: Value-added PubMed search engine for analysis and text mining, PloS one 16 (3) (2021) e0248335.

Appendix A. Ablation studies and analyses with PubMed-only features

Table A.5 shows the results of the ablation study using different loss functions.

Model	Precision ↑		Recall ↑		F_1 ↑		ECE ↓		
	Micro	Macro	Micro	Macro	Micro	Macro	L1	L2	Max
BCE	0.672 [0.668-0.675]	0.688 [0.678-0.695]	0.674 [0.671-0.677]	0.667 [0.659-0.676]	0.673 [0.670-0.675]	0.670 [0.662-0.677]	0.003 [0.003-0.003]	0.012 [0.011-0.012]	0.080 [0.071-0.089]
BCE-LS	0.651 [0.647-0.654]	0.674 [0.667-0.680]	0.683 [0.680-0.687]	0.672 [0.664-0.681]	0.666 [0.664-0.669]	0.663 [0.656-0.669]	0.002 [0.001-0.002]	0.004 [0.003-0.004]	0.033 [0.024-0.044]
ASL	0.672 [0.669-0.676]	0.703 [0.696-0.709]	0.685 [0.682-0.689]	0.685 [0.677-0.692]	0.679 [0.676-0.681]	0.689 [0.682-0.694]	0.180 [0.179-0.180]	0.205 [0.205-0.206]	0.366 [0.364-0.367]
ASL-LS (Main model)	0.674 [0.671-0.677]	0.703 [0.696-0.710]	0.683 [0.680-0.687]	0.685 [0.678-0.691]	0.679 [0.676-0.681]	0.689 [0.683-0.695]	0.176 [0.175-0.176]	0.201 [0.200-0.201]	0.357 [0.355-0.359]

Table A.5: Ablation study using different loss functions. ASL refers to asymmetric loss. BCE refers to binary cross entropy. LS is label smoothing. All LS experiments used a=0.05, the hyperparameter to control the amount of smoothing.

Table A.6 shows the results of the ablation study using different base models.

Model	Precision ↑		Recall ↑		F_1 ↑		AUC ↑	
	Micro	Macro	Micro	Macro	Micro	Macro	Micro	Macro
PubMedBERT	0.645 [0.642-0.649]	0.665 [0.658-0.672]	0.645 [0.641-0.648]	0.636 [0.629-0.643]	0.645 [0.642-0.648]	0.642 [0.636-0.648]	0.968 [0.968-0.969]	0.959 [0.958-0.961]
BioLinkBERT	0.664 [0.660-0.667]	0.687 [0.680-0.694]	0.690 [0.687-0.694]	0.696 [0.689-0.703]	0.677 [0.674-0.679]	0.686 [0.680-0.692]	0.974 [0.974-0.975]	0.967 [0.966-0.968]
SPECTER	0.671 [0.668-0.675]	0.706 [0.699-0.713]	0.682 [0.679-0.685]	0.677 [0.669-0.684]	0.677 [0.674-0.679]	0.686 [0.679-0.693]	0.974 [0.973-0.974]	0.967 [0.966-0.967]
SPECTER2-Base (main model)	0.674 [0.671-0.677]	0.703 [0.696-0.710]	0.683 [0.680-0.687]	0.685 [0.677-0.692]	0.679 [0.676-0.681]	0.689 [0.682-0.695]	0.974 [0.974-0.975]	0.968 [0.967-0.969]
SPECTER2-Clf	0.676 [0.673-0.679]	0.707 [0.700-0.713]	0.681 [0.678-0.684]	0.677 [0.669-0.685]	0.679 [0.676-0.681]	0.686 [0.679-0.692]	0.974 [0.974-0.975]	0.968 [0.967-0.969]

Table A.6: Ablation study using different base models. SPECTER2-Base refers to the model without multi-task learning adapters, whereas SPECTER2-Clf refers to the model using classification-specific adapters. SPECTER2-Base is marked (main model) to indicate it was used in the best performing model.

Table A.7 shows the performance comparison of models trained using supplemented label sets.

Table A.8 shows the performance comparison of models trained using contrastive loss.

Model	**Precision ↑**		**Recall ↑**		**F_1 ↑**		**AUC ↑**	
	Micro	Macro	Micro	Macro	Micro	Macro	Micro	Macro
No Splits	0.672 [0.669-0.675]	0.696 [0.689-0.702]	0.683 [0.680-0.686]	0.687 [0.680-0.694]	0.677 [0.674-0.680]	0.687 [0.681-0.693]	0.974 [0.973-0.974]	0.967 [0.966-0.968]
+ COHORT-SPLIT	0.664 [0.661-0.667]	0.699 [0.693-0.706]	0.692 [0.689-0.695]	0.688 [0.681-0.695]	0.678 [0.675-0.681]	0.688 [0.682-0.694]	0.974 [0.974-0.975]	0.967 [0.967-0.968]
+ GENERALIZED RCT	0.674 [0.671-0.678]	0.701 [0.694-0.708]	0.682 [0.679-0.686]	0.685 [0.677-0.692]	0.678 [0.675-0.681]	0.687 [0.681-0.694]	0.974 [0.974-0.974]	0.967 [0.967-0.968]
+ HUMAN	0.667 [0.664-0.670]	0.696 [0.689-0.703]	0.687 [0.684-0.691]	0.689 [0.682-0.696]	0.677 [0.674-0.680]	0.686 [0.680-0.692]	0.974 [0.974-0.975]	0.968 [0.967-0.968]
+ ANIMAL (main model)	0.674 [0.671-0.677]	0.703 [0.696-0.710]	0.683 [0.680-0.687]	0.685 [0.677-0.692]	0.679 [0.676-0.681]	0.689 [0.682-0.695]	0.974 [0.974-0.975]	0.968 [0.967-0.969]
+ VETERINARY	0.669 [0.665-0.672]	0.702 [0.695-0.709]	0.687 [0.683-0.690]	0.685 [0.678-0.693]	0.678 [0.675-0.680]	0.688 [0.681-0.694]	0.974 [0.974-0.975]	0.968 [0.967-0.969]
+ COMBINED	0.674 [0.671-0.678]	0.717 [0.705-0.727]	0.680 [0.676-0.683]	0.656 [0.649-0.664]	0.677 [0.674-0.680]	0.672 [0.665-0.679]	0.974 [0.974-0.974]	0.967 [0.967-0.968]

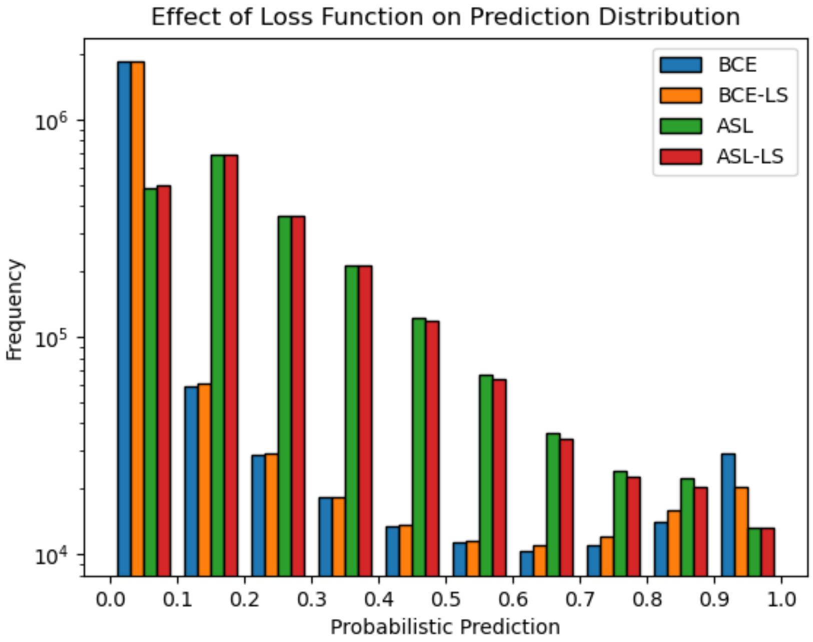
Table A.7: Performance comparison of models trained using supplemented label sets: label splitting (COHORT-SPLIT and GENERALIZED RCT), non-PT MeSH terms (HUMAN, ANIMAL, and VETERINARY), or a combination of all of these. The baseline is the SPECTER2-base model with the standard label set (61 PT labels). The ANIMAL experiment is marked to indicate it is used in the best performing model.

Model	**Precision ↑**		**Recall ↑**		**F_1 ↑**		**AUC ↑**	
	Micro	Macro	Micro	Macro	Micro	Macro	Micro	Macro
SPECTER2-Base (Base)	0.674 [0.671-0.677]	0.703 [0.696-0.710]	0.683 [0.680-0.687]	0.685 [0.677-0.692]	0.679 [0.676-0.681]	0.689 [0.682-0.695]	0.974 [0.974-0.975]	0.968 [0.967-0.969]
Base + SimCSE	0.653 [0.650-0.656]	0.704 [0.697-0.711]	0.699 [0.695-0.702]	0.687 [0.678-0.695]	0.675 [0.672-0.678]	0.687 [0.680-0.694]	0.975 [0.974-0.975]	0.968 [0.967-0.969]
Base + ADNCE	0.656 [0.653-0.660]	0.707 [0.699-0.713]	0.695 [0.691-0.698]	0.683 [0.674-0.691]	0.675 [0.672-0.678]	0.686 [0.679-0.693]	0.975 [0.974-0.975]	0.968 [0.967-0.969]
Base + HeroCon	0.669 [0.666-0.673]	0.698 [0.691-0.705]	0.688 [0.684-0.691]	0.691 [0.682-0.698]	0.678 [0.675-0.681]	0.688 [0.681-0.694]	0.975 [0.974-0.975]	0.968 [0.967-0.969]
Base + HeroCon + Noise-Aware	0.669 [0.666-0.673]	0.699 [0.693-0.706]	0.689 [0.686-0.693]	0.691 [0.683-0.699]	0.679 [0.676-0.682]	0.689 [0.683-0.695]	0.975 [0.974-0.975]	0.968 [0.967-0.969]
Base + WeighCon (main model)	0.670 [0.666-0.673]	0.700 [0.693-0.707]	0.688 [0.685-0.692]	0.692 [0.684-0.700]	0.679 [0.676-0.682]	0.690 [0.683-0.696]	0.975 [0.974-0.975]	0.968 [0.967-0.969]
Base + WeighCon + Noise-Aware	0.669 [0.666-0.673]	0.699 [0.693-0.706]	0.689 [0.686-0.693]	0.691 [0.683-0.699]	0.679 [0.676-0.682]	0.689 [0.683-0.695]	0.975 [0.974-0.975]	0.968 [0.967-0.969]
Base + SimCSE + WeighCon	0.654 [0.651-0.657]	0.705 [0.697-0.712]	0.697 [0.693-0.700]	0.684 [0.675-0.693]	0.675 [0.672-0.677]	0.686 [0.678-0.693]	0.975 [0.974-0.975]	0.968 [0.967-0.969]

Table A.8: Performance comparison of models trained using SPECTER2-Base and various CL approaches, including a noise-aware component, which utilized a modulation term and in-batch label correction for contrastive loss. The WeighCon experiment is marked (main model) to indicate it was used in the best performing model.

A figure plotting the probability distribution of model predictions for each loss function is shown in Figure A.4. In a well calibrated model, 90% of instances predicted with 0.9 would be positively labeled; 10% of instances with 0.1 would be positively labeled. With ASL, it appears the probability shifting results in a model that is more poorly calibrated than BCE. Despite this, the overall performance, including AUC (micro-AUC [95% CI]: BCE=0.973 [0.973-0.974], ASL=0.974 [0.974-0.975]) is better when using ASL. To ensure a high recall/sensitivity while limiting false positives, optimal thresholds should be determined empirically rather than simply using low probability cut-offs (0.01) as is more common in well calibrated models.

Figure A.4: The distribution of probabilistic predictions for articles in the test set for models in the loss function ablation study, which evaluates binary cross entropy, binary cross entropy with label smoothing, asymmetric loss, and asymmetric loss with label smoothing.



Appendix B. Ablation studies and analyses with full-text features

Figure B.5 shows the PT label distribution for the full-text dataset and the performances of the base model and the best-performing full-text model on the full-text test set.

Table B.9 shows the results of the ablation study of different combinations of full-text features.

Model	**Precision ↑**		**Recall ↑**		**F_1 ↑**		**AUC ↑**	
	Micro	Macro	Micro	Macro	Micro	Macro	Micro	Macro
Best Combination	0.516 [0.509-0.524]	0.489 [0.471-0.509]	0.812 [0.805-0.820]	0.702 [0.676-0.724]	0.631 [0.625-0.638]	0.566 [0.548-0.584]	0.975 [0.974-0.977]	0.943 [0.931-0.952]
- First sentence	0.518 [0.510-0.526]	0.495 [0.477-0.514]	0.810 [0.803-0.818]	0.699 [0.672-0.721]	0.632 [0.625-0.639]	0.567 [0.549-0.585]	0.975 [0.974-0.976]	0.942 [0.929-0.951]
- Label sentences	0.512 [0.504-0.519]	0.483 [0.464-0.502]	0.807 [0.799-0.814]	0.696 [0.672-0.718]	0.626 [0.619-0.633]	0.559 [0.541-0.577]	0.975 [0.974-0.976]	0.942 [0.929-0.951]
- Ethics	0.518 [0.510-0.525]	0.489 [0.471-0.509]	0.811 [0.803-0.818]	0.698 [0.674-0.721]	0.632 [0.624-0.638]	0.564 [0.546-0.582]	0.975 [0.974-0.977]	0.942 [0.930-0.952]
- NCT identifiers	0.518 [0.510-0.526]	0.487 [0.469-0.505]	0.812 [0.805-0.820]	0.694 [0.669-0.718]	0.632 [0.625-0.639]	0.561 [0.544-0.579]	0.975 [0.974-0.977]	0.942 [0.931-0.951]
- Number features	0.515 [0.508-0.523]	0.485 [0.466-0.504]	0.812 [0.804-0.820]	0.695 [0.669-0.720]	0.630 [0.624-0.638]	0.560 [0.543-0.579]	0.976 [0.974-0.977]	0.943 [0.931-0.953]
- Section headings	0.516 [0.508-0.523]	0.485 [0.466-0.504]	0.812 [0.804-0.819]	0.696 [0.670-0.719]	0.631 [0.623-0.637]	0.561 [0.543-0.578]	0.975 [0.974-0.977]	0.943 [0.931-0.951]
- TextRank	0.514 [0.506-0.522]	0.486 [0.468-0.506]	0.811 [0.803-0.818]	0.703 [0.677-0.725]	0.623 [0.621-0.636]	0.564 [0.546-0.583]	0.975 [0.974-0.977]	0.943 [0.930-0.953]
- EXTRACTIVE (MULTIPLE)	0.515 [0.508-0.522]	0.486 [0.468-0.506]	0.812 [0.804-0.820]	0.698 [0.673-0.723]	0.630 [0.624-0.637]	0.563 [0.545-0.581]	0.975 [0.974-0.977]	0.943 [0.930-0.952]

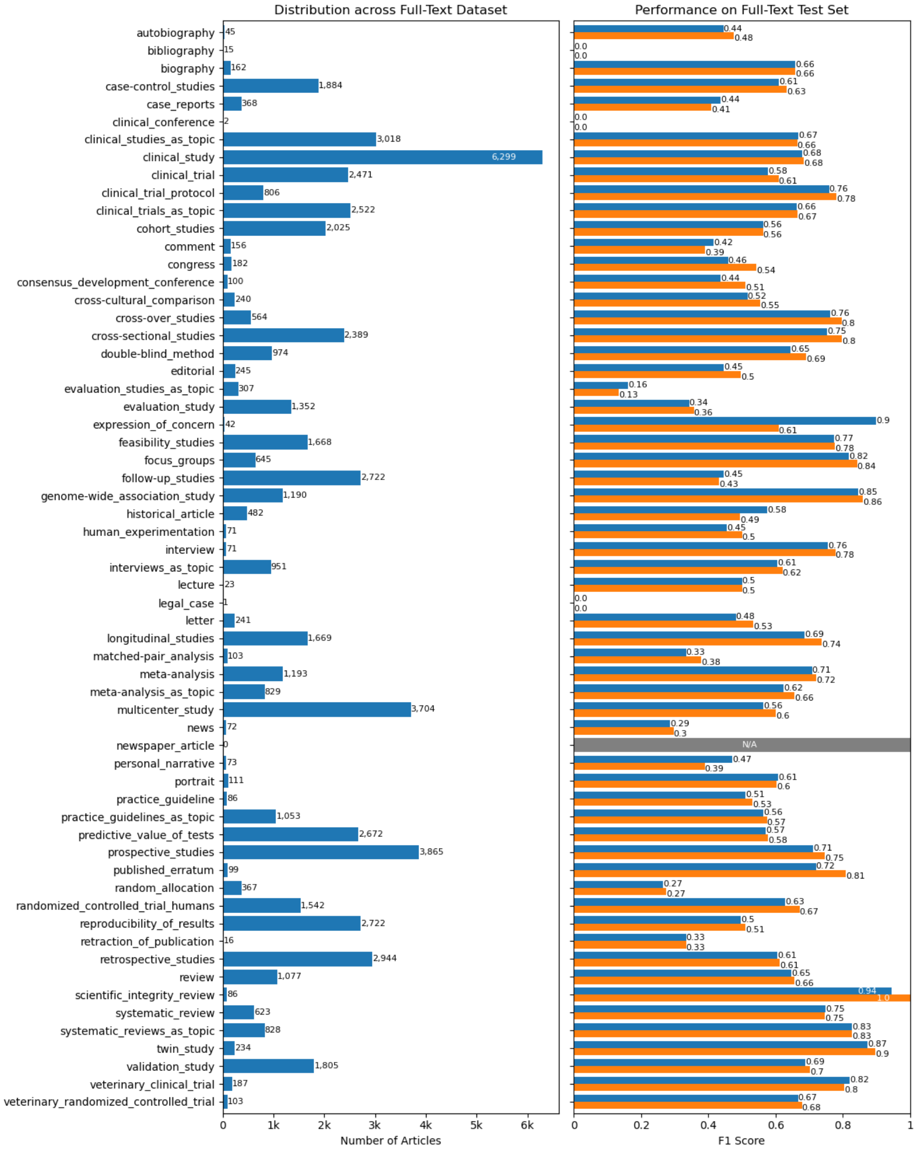
Table B.9: An ablation study of different combinations of full-text features. Models are trained on the full-text training set (n = 17,111) and evaluated on the full-text test set (n = 4,852) are reported. The best combination model uses EXTRACTIVE (MULTIPLE) summaries, label sentences, first sentence, NCT identifier information, ethics, number features (rough word count of article, # of tables, and # of figures), and primary section heading features. "-" indicates a feature was removed.

Table B.10 shows the performance comparison of models using full-text features and summarization techniques.

Model	**Precision ↑**		**Recall ↑**		**F$_1$ ↑**		**AUC ↑**	
	Micro	Macro	Micro	Macro	Micro	Macro	Micro	Macro
PubMed features only	0.505 [0.497-0.513]	0.485 [0.468-0.504]	0.791 [0.783-0.799]	0.684 [0.658-0.707]	0.616 [0.610-0.623]	0.556 [0.539-0.574]	0.972 [0.971-0.974]	0.938 [0.925-0.947]
B + TextRank	0.505 [0.498-0.513]	0.477 [0.459-0.499]	0.797 [0.789-0.806]	0.677 [0.650-0.701]	0.619 [0.612-0.625]	0.549 [0.530-0.568]	0.973 [0.972-0.974]	0.938 [0.927-0.947]
B + extractive (single)	0.509 [0.501-0.517]	0.485 [0.463-0.505]	0.797 [0.789-0.805]	0.689 [0.665-0.712]	0.621 [0.614-0.628]	0.558 [0.538-0.577]	0.973 [0.972-0.975]	0.938 [0.926-0.947]
B + extractive (multiple)	0.510 [0.503-0.518]	0.485 [0.463-0.506]	0.796 [0.788-0.804]	0.680 [0.656-0.702]	0.622 [0.615-0.628]	0.555 [0.534-0.575]	0.973 [0.972-0.975]	0.938 [0.926-0.948]
B + extractive (multiple+definition)	0.507 [0.500-0.515]	0.484 [0.465-0.505]	0.797 [0.789-0.805]	0.687 [0.662-0.709]	0.620 [0.613-0.627]	0.556 [0.536-0.575]	0.973 [0.972-0.975]	0.939 [0.927-0.948]
B + llm	0.515 [0.507-0.521]	0.483 [0.462-0.505]	0.802 [0.795-0.809]	0.681 [0.655-0.705]	0.627 [0.620-0.633]	0.554 [0.534-0.574]	0.974 [0.972-0.975]	0.938 [0.925-0.948]
B + primera	0.521 [0.513-0.529]	0.485 [0.465-0.506]	0.802 [0.794-0.810]	0.680 [0.654-0.703]	0.632 [0.624-0.639]	0.554 [0.534-0.574]	0.974 [0.973-0.976]	0.941 [0.928-0.950]

Table B.10: Performance comparison of models using full-text features and various summarization techniques on the full-text test articles (n = 4,852).

Figure B.5: The left panel shows the PT label distribution for articles in our full-text dataset. The right panel shows the individual label performances (F1 score) of the base model (blue) and the best-performing full-text model (orange) on the full-text test set (n = 4,852). The base model uses no full-text features, just features derived from PubMed (e.g., title, abstract, etc.). The best model uses the following full-text features: EXTRACTIVE (MULTIPLE), label sentences, first sentence, NCT identifier information, ethics, number features (rough word count of article, # of tables, and # of figures), and primary section heading features.


Distribution across Full-Text Dataset
Performance on Full-Text Test Set
autobiography 45 0.44 0.48
bibliography 15 0.0 0.0
biography 162 0.66 0.66
case-control_studies 1,884 0.61 0.63
case_reports 368 0.44 0.41
clinical_conference 2 0.0 0.0
clinical_studies_as_topic 3,018 0.67 0.66
clinical_study 6,299 0.68 0.68
clinical_trial 2,471 0.58 0.61
clinical_trial_protocol 806 0.76 0.78
clinical_trials_as_topic 2,522 0.66 0.67
cohort_studies 2,025 0.56 0.56
comment 156 0.42 0.39
congress 182 0.46 0.54
consensus_development_conference 100 0.44 0.51
cross-cultural_comparison 240 0.52 0.55
cross-over_studies 564 0.76 0.8
cross-sectional_studies 2,389 0.75 0.8
double-blind_method 974 0.65 0.69
editorial 245 0.45 0.5
evaluation_studies_as_topic 307 0.16 0.13
evaluation_study 1,352 0.34 0.36
expression_of_concern 42 0.9 0.61
feasibility_studies 1,668 0.77 0.78
focus_groups 645 0.82 0.84
follow-up_studies 2,722 0.45 0.43
genome-wide_association_study 1,190 0.85 0.86
historical_article 482 0.58 0.49
human_experimentation 71 0.45 0.5
interview 71 0.76 0.78
interviews_as_topic 951 0.61 0.62
lecture 23 0.5 0.5
legal_case 1 0.0 0.0
letter 241 0.48 0.53
longitudinal_studies 1,669 0.69 0.74
matched-pair_analysis 103 0.33 0.38
meta-analysis 1,193 0.71 0.72
meta-analysis_as_topic 829 0.62 0.66
multicenter_study 3,704 0.56 0.6
news 72 0.29 0.3
newspaper_article 0 N/A
personal_narrative 73 0.47 0.39
portrait 111 0.61 0.6
practice_guideline 86 0.51 0.53
practice_guidelines_as_topic 1,053 0.56 0.57
predictive_value_of_tests 2,672 0.57 0.58
prospective_studies 3,865 0.71 0.75
published_erratum 99 0.72 0.81
random_allocation 367 0.27 0.27
randomized_controlled_trial_humans 1,542 0.63 0.67
reproducibility_of_results 2,722 0.5 0.51
retraction_of_publication 16 0.33 0.33
retrospective_studies 2,944 0.61 0.61
review 1,077 0.65 0.66
scientific_integrity_review 86 0.94 1.0
systematic_review 623 0.75 0.75
systematic_reviews_as_topic 828 0.83 0.83
twin_study 234 0.87 0.9
validation_study 1,805 0.69 0.7
veterinary_clinical_trial 187 0.82 0.8
veterinary_randomized_controlled_trial 103 0.67 0.68
0 1k 2k 3k 4k 5k 6k
Number of Articles
F1 Score