Efficient Few-shot Learning for Multi-label Classification of Scientific Documents with Many Classes

Tim Schopf, Alexander Blatzheim, Nektarios Machner, and Florian Matthes

Technical University of Munich, Department of Computer Science, Germany

`{tim.schopf,alexander.blatzheim,nektarios.machner,matthes}`
`@tum.de`

Abstract

Scientific document classification is a critical task and often involves many classes. However, collecting human-labeled data for many classes is expensive and usually leads to label-scarce scenarios. Moreover, recent work has shown that sentence embedding model fine-tuning for few-shot classification is efficient, robust, and effective. In this work, we propose FusionSent (**Fusion**-based **Sent**ence Embedding Fine-tuning), an efficient and prompt-free approach for few-shot classification of scientific documents with many classes. FusionSent uses available training examples and their respective label texts to contrastively fine-tune two different sentence embedding models. Afterward, the parameters of both fine-tuned models are fused to combine the complementary knowledge from the separate fine-tuning steps into a single model. Finally, the resulting sentence embedding model is frozen to embed the training instances, which are then used as input features to train a classification head. Our experiments show that Fusion-Sent significantly outperforms strong baselines by an average of 6.0 F_1 points across multiple scientific document classification datasets. In addition, we introduce a new dataset for multi-label classification of scientific documents, which contains 203,961 scientific articles and 130 classes from the arXiv category taxonomy. Code and data are available at https://github.com/sebischair/FusionSent.

1 Introduction

Scientific literature has grown exponentially over the last few decades, with countless new publications being added every year (Dong et al., 2017). To be searchable and accessible to researchers, policymakers, and the public, scientific literature must be managed and categorized in digital libraries (Toney and Dunham, 2022). However, this poses a significant challenge due to the huge volume of documents and the variety of topics they cover (Sadat and Caragea, 2022). In addition to the broad spectrum of possible topics, scientific documents often cannot be assigned to just one topic due to their interdisciplinary character. Consequently, automatically categorizing scientific documents must be approached as a multi-label classification problem over large label spaces. Previous works approach this task either in an unsupervised (Shen et al., 2018; Salatino et al., 2019; Mustafa et al., 2021; Toney and Dunham, 2022; Schopf and Matthes, 2024) or in a fully supervised (Gialitsis et al., 2022; Sadat and Caragea, 2022; E. Mendoza et al., 2022; Schopf et al., 2023) manner. While supervised approaches offer high prediction quality, they require a large corpus of annotated data to perform. Often, however, a large corpus of annotated data is unavailable, e.g., when a new categorization scheme is being developed for an emerging scientific field. Unsupervised approaches provide a possible circumvention of this limitation but are accompanied at the expense of prediction quality.

To improve classification performance in scenarios where labeled data is unavailable, domain experts may annotate a small part of the dataset. However, annotating many classes naturally leads to data scarcity, as collecting sufficient training data for all classes causes significantly higher costs (Xu et al., 2023a). Therefore, to support the classification of scientific documents in such scenarios, we consider the multi-label classification of scientific documents as a few-shot task. Few-shot approaches are designed to train an effective model with a few labeled examples, reducing the cost of developing models for new domains and tasks (Huang et al., 2023).

In recent work, SetFit (Tunstall et al., 2022) demonstrated strong few-shot classification performance by contrastively fine-tuning (Koch et al., 2015) sentence embedding models. Since this approach does not require prompts and is effective on relatively small models, it is much more efficient
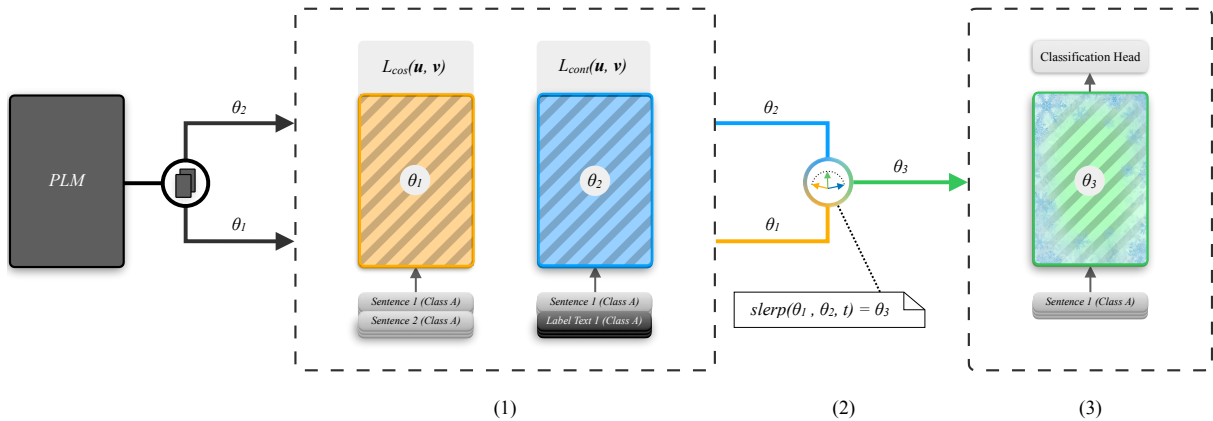
Figure 1: The training process of FusionSent comprises three steps: (1) Fine-tune two different sentence embedding models from the same Pre-trained Language Model (PLM), with parameters θ_1, θ_2 respectively. θ_1 is fine-tuned on pairs of training sentences using cosine similarity loss and θ_2 is fine-tuned on pairs of training sentences and their corresponding label texts, using contrastive loss. Label texts can consist of simple label/class names or of more extensive texts that semantically describe the meaning of a label/class. (2) Merge parameter sets θ_1, θ_2 into θ_3 using Spherical Linear Interpolation (SLERP). (3) Freeze θ_3 to embed the training sentences, which are then used as input features to train a classification head.

and consistent than common prompt-based methods such as In-Context Learning (ICL) (Brown et al., 2020) and Pattern-Exploiting Training (PET) (Schick and Schütze, 2021), which involve careful prompt engineering and large-scale model sizes.

In this paper, we propose *FusionSent*, which builds on the idea of contrastive sentence embedding training for efficient few-shot classification. As illustrated in Figure 1, FusionSent uses the few annotated examples, as well as label texts, to contrastively fine-tune two separate sentence embedding models from the same Pre-trained Language Model (PLM) checkpoint. One model is fine-tuned to maximize similarities between training examples sharing the same class, and the other model is fine-tuned to maximize similarities between training examples and their corresponding label texts. After fine-tuning, the weights of both models are merged to obtain the model body of FusionSent. For subsequent classifier training, the model body is frozen to embed the few training examples, which are then used as input features to train a simple logistic regression head. This approach works effectively with relatively small model sizes, requires no prompts, and merging fine-tuned sentence embedding models incurs no additional inference or memory costs (Wortsman et al., 2022). Our experiments show that FusionSent consistently outperforms various baselines on different datasets for multi-label classification of scientific documents with many classes. Furthermore, we show that FusionSent can

improve few-shot performance in multi-class settings of different domains with a small number of classes.

In addition to FusionSent, we introduce a new dataset for multi-label classification of scientific documents. The dataset consists of 130 classes and 203,961 scientific articles that have been manually categorized by their authors into one or more topics from the arXiv category taxonomy[1].

2 Related Work

2.1 Classification of Scientific Documents

Unsupervised approaches typically use embeddings of topics as well as scientific documents and perform classification based on their similarities (Shen et al., 2018; Salatino et al., 2019; Mustafa et al., 2021; Toney and Dunham, 2022). More recently, classifying scientific documents has been regarded as a fully supervised task. SciNoBo (Gialitsis et al., 2022) uses the structural properties of publications and their citations and references organized in a multilayer graph network for predicting topics of scientific publications. HR-SciBERT (Sadat and Caragea, 2022) uses a multi-task learning approach for topic classification with keyword labeling as an auxiliary task. Finally, E. Mendoza et al. (2022) use ensemble models to classify scientific documents into multiple research themes.

---
[1] https://arxiv.org/category_taxonomy

2.2 Few-shot Classification

Prominent techniques for few-shot classification involve ICL, utilizing task-specific prompts with a few labeled examples (Brown et al., 2020). However, while avoiding gradient updates, ICL necessitates large model sizes for good performance, resulting in computationally expensive inference. Conversely, prompt-based fine-tuning proves to be effective with smaller models (Schick and Schütze, 2021; Tam et al., 2021; Gao et al., 2021). Additionally, Parameter Efficient Fine-Tuning (PEFT) can further reduce training costs by fine-tuning a considerably smaller module within a frozen PLM (Houlsby et al., 2019; Li and Liang, 2021; Hu et al., 2022; Karimi Mahabadi et al., 2022; He et al., 2022; Liu et al., 2022; Aly et al., 2023). In contrast to these methods, fine-tuning few-shot classification models via contrastive sentence embedding training provides two primary advantages: (1) it requires significantly smaller model sizes, and (2) eliminates the necessity for prompts or instructions (Tunstall et al., 2022; Huang et al., 2023; Bates and Gurevych, 2024), which can cause significant performance variance and require careful design (Perez et al., 2021).

2.3 Model Fusion

Model fusion, which involves the integration of capabilities from different models, can be mainly divided into two categories. Firstly, ensemble approaches combine the output of multiple models to enhance the overall prediction performance (Littlestone and Warmuth, 1994; Sagi and Rokach, 2018). Outputs are typically combined by weight averaging (Littlestone and Warmuth, 1994) or majority voting (Monteith et al., 2011). These ensemble approaches can improve the prediction performance of large-scale language models (Jiang et al., 2023). Secondly, weight merging approaches enable model fusion at the parameter level. Wortsman et al. (2022) show that weight averaging of multiple models fine-tuned with different hyperparameters improves prediction accuracy and robustness. Task vectors derived from model weights can be modified and combined together through arithmetic operations to steer the behavior of a resulting model (Ilharco et al., 2023). This approach can be enhanced by trimming task vectors and resolving sign conflicts before merging them (Yadav et al., 2023). In addition, Drop And Rescale (DARE) can be used as a general preprocessing technique for

existing model merging methods to merge multiple task-specific fine-tuned models into a single model with diverse abilities (Yu et al., 2023).

2.4 Datasets for Topic Classification of Scientific Documents

Various datasets for multi-label topic classification of scientific documents have been introduced. The Cora dataset (McCallum et al., 2000) contains about 50,000 computer science research papers categorized into 79 topics. Several datasets have been released based on the ACM Computing Classification System[2] (Santos and Rodrigues, 2009; Sadat and Caragea, 2022). Schopf et al. (2023) introduce a dataset of 179,349 scientific papers categorized into 82 different NLP-related topics. Yang et al. (2018) create a dataset of 55,840 arXiv[3] papers, in which each paper is assigned to several classes covering 54 different topics. However, this dataset is not publicly available.

3 Background

Sentence Embedding Model Fine-tuning for Few-shot Classification Tunstall et al. (2022) show that sentence embedding models can be used in a two-step training process for efficient few-shot classification. In the first step, a sentence embedding model is fine-tuned in a contrastive manner by sampling positive and negative sentence pairs from few-shot labeled examples. In the second step, the fine-tuned sentence embedding model is frozen to encode all available few-shot examples. The resulting embeddings are then used as input features to train a simple logistic regression classifier (Cox, 1958) as the model head.

Label Texts for Document Classification in Label-scarce Scenarios Xu et al. (2023a) show that mapping representation spaces of training instances to their respective label descriptions in embedding space can be effective in label-scarce classification scenarios. They reformulate classification with many classes as a dense retrieval task and train a dual encoder that learns to maximize the similarity between embeddings of the training instances and their respective label descriptions. During inference, they use the top-k retrieved labels of each instance for classification. Similarly, WanDeR (Xu et al., 2023b) and FastFit (Yehudai and Bendel, 2024) use label names and dense retrieval for

---

[2]https://dl.acm.org/ccs
[3]https://arxiv.org

multi-class classification. However, dense retrieval approaches are challenging to apply in multi-label classification scenarios since the number of classes per instance can vary significantly.

4 Method

As illustrated in Figure 1, our few-shot classification method consists of separate training parts for the model body and the model head. We fine-tune a sentence embedding model as the model body, while the model head consists of a simple logistic regression classifier trained on the data encoded by the model body.

4.1 Model Body

Given a base PLM, we fine-tune the model body of FusionSent in three steps: (1) use SetFit's contrastive learning approach to construct positive and negative training pairs from the few training examples to fine-tune a sentence embedding model from the base PLM (2) construct positive and negative training pairs from the few training examples and their corresponding label texts to fine-tune a different sentence embedding model using the same base PLM, and (3) merge both fine-tuned sentence embedding models to obtain the model body of FusionSent.

In the first step, we fine-tune a sentence embedding model from the base PLM using contrastive learning and the few training examples. Specifically, from the few training examples, instances of the same class are selected as positive pairs, which are assigned a score of 1, and instances from different classes are selected as negative pairs, which are assigned a score of 0. These training pairs are then used to fine-tune a sentence embedding model with the Cosine Similarity Loss:

$$L_{cos} = \|y - cos\_sim(u, v)\|_2, \qquad (1)$$

where $u, v \in \mathbb{R}^D$ are the D-dimensional sentence embeddings of two sentences respectively, and $y \in \{0, 1\}$ is the pair label.

In the second step, we use a different contrastive training approach to fine-tune a separate sentence embedding model from the same base PLM, using the few training instances and their corresponding label texts. Specifically, positive pairs consist of training instances and the label texts of the class assigned to them. Negative pairs consist of training instances and label texts from different classes. Label texts can consist of simple label/class names,

which are usually available in datasets, or of more extensive texts that semantically describe the meaning of a label/class. We assign the positive pairs a score of 1 and the negative pairs a score of 0 to fine-tune a sentence embedding model with the Contrastive Loss (Hadsell et al., 2006):

$$L_{cont} = \frac{1}{2} \big[y \cdot cos\_dist(u, v)^2 +$$
$$(1 - y) \cdot max\{0, m - cos\_dist(u, v)\}^2 \big], \quad (2)$$

where $u, v \in \mathbb{R}^D$ are the D-dimensional sentence embeddings of two sentences respectively, $m = 0.5$ is a margin, and $y \in \{0, 1\}$ is the pair label.

To obtain the contrastive training pairs for steps one and two, we use an *oversampling* strategy. In this approach, an equal number of positive and negative training pairs are sampled, with the minority pair type (positive) being oversampled to align with the majority pair type (negative).

In the third step, the parameters of the fine-tuned sentence embedding models obtained in steps one and two are merged using Spherical Linear Interpolation (SLERP) (Shoemake, 1985). Specifically, let θ_1 be the parameters obtained from the first fine-tuning step and θ_2 the parameters obtained from the second fine-tuning step, we merge parameters with SLERP:

$$slerp(\theta_1, \theta_2; t) = \frac{\sin(1 - t)\Omega}{\sin \Omega} \theta_1 + \frac{\sin t\Omega}{\sin \Omega} \theta_2, \tag{3}$$

where $\theta_1 \cdot \theta_2 = \cos \Omega$ and $t = 0.5$ is an interpolation factor. Finally, the new parameters θ_3 obtained from SLERP merging are inserted into a sentence embedding model derived from the same architecture as the base PLM, resulting in the FusionSent model body.

4.2 Model Head & Inference

In the second part of FusionSent training, we first use the frozen model body to embed all available training instances. Then, we train a logistic regression model using the embedded training instances as input features. During inference, the model body embeds the inputs to provide features for the logistic regression head that subsequently classifies the unseen instances.

5 Experiments

5.1 Data

We construct a dataset of scientific documents derived from arXiv metadata (Clement et al., 2019). The arXiv metadata provides information about more than 2 million scholarly articles published in arXiv from various scientific fields. We use this metadata to create a dataset of 203,961 titles and abstracts categorized into 130 different classes. To this end, we first perform a stratified downsampling of the metadata to only 10% of all articles while retaining the original class distribution. Afterward, articles assigned to categories occurring less than 100 times in the downsampled dataset are removed. To obtain the final dataset, we then perform a stratified train/validation/test split of the processed dataset in an 80:10:10 ratio. The number of examples in each set are shown in Table 1.

Dataset split	Size
Train	163,168
Validation	20,396
Test	20,397

Table 1: Overview of the arXiv dataset.

Each article in the resulting arXiv dataset is categorized into one or more distinct categories. Figure 2 shows the distribution of papers across the 130 categories of the dataset.
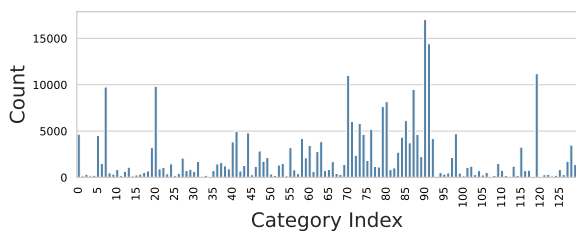

Figure 2: Number of papers in each category of the arXiv dataset.

In addition, we use the SciHTC dataset (Sadat and Caragea, 2022), which contains computer science papers categorized into one or more classes of the ACM Computing Classification System. We remove classes with less than 100 examples, resulting in 46,372 training samples and 5,838 test samples categorized into 62 different classes.

As a third dataset for scientific document classification, we use the NLP taxonomy dataset (Schopf et al., 2023), which contains papers from the ACL Anthology[4], the arXiv cs.CL category, and Scopus[5], categorized into one or more Natural Language Processing (NLP)-related classes. We perform a stratified 90:10 split between training and test examples, resulting in 161,414 training and 17,935 test instances categorized into 82 different classes.

5.2 Models

We experiment with two baselines and four different few-shot learning approaches for multi-label classification of scientific documents.

FineTune The first baseline consists of a standard encoder-only transformer that is fine-tuned for text classification.

SetFit The second baseline consists of the SetFit approach without any changes to the architecture or the training procedure.

Label Embedding (LE) As an initial few-shot learning approach, we experiment with only training one sentence embedding model that uses few-shot examples and their corresponding label texts in a contrastive learning approach. This approach consists of training a model body, as described in step 2 in Section 4.1, and training a logistic regression head on top of it.

SetFit→LE We also experiment with combining contrastive learning approaches to directly fine-tune a single sentence embedding model rather than separate models that are merged later. For this purpose, we perform training steps one and two as described in Section 4.1 sequentially on the same sentence embedding model. For classification, we then train a logistic regression head.

LE→SetFit This approach also only trains a single sentence embedding model. However, we first perform training step two followed by step one as described in Section 4.1 on the same model. We then train a logistic regression model head for the classification.

FusionSent Finally, we experiment with the FusionSent approach as described in Section 4.

5.3 Experimental Setup

Systematically evaluating few-shot performance is challenging due to the potential instability arising from fine-tuning on small datasets (Zhang et al.,

---

[4]https://aclanthology.org
[5]https://www.scopus.com

Dataset →	arXiv			SciHTC			NLP Taxonomy			Average
Method ↓	F_1	P	R	F_1	P	R	F_1	P	R	F_1
					$\lvert N \rvert = 2^*$					
FineTune	-	-	-	-	-	-	$13.1_{3.2}$	$78.3_{5.5}$	$7.2_{2.0}$	-
SetFit	$37.5_{1.6}$	$45.5_{1.7}$	$32.0_{2.4}$	$31.4_{0.8}$	$39.9_{0.1}$	$26.0_{1.2}$	$58.7_{2.0}$	$55.7_{4.2}$	$62.3_{0.9}$	$42.5_{1.5}$
Label Embedding	$43.2_{0.1}$	$45.3_{1.9}$	$40.4_{0.5}$	$34.1_{4.3}$	$46.1_{6.3}$	$27.1_{3.3}$	$65.8_{2.2}$	$66.0_{5.5}$	$66.0_{1.6}$	$47.7_{2.2}$
SetFit→LE	$41.5_{0.9}$	$45.3_{1.5}$	$38.3_{1.7}$	$32.3_{1.9}$	$41.6_{3.2}$	$26.3_{1.4}$	$64.6_{1.8}$	$62.7_{4.4}$	$66.8_{2.1}$	$46.1_{1.5}$
LE→SetFit	$39.2_{0.5}$	$41.5_{1.1}$	$37.1_{0.5}$	$34.0_{1.6}$	$38.9_{1.6}$	$30.2_{2.0}$	$61.3_{1.6}$	$55.0_{3.4}$	$69.3_{1.5}$	$44.8_{1.2}$
FusionSent	$\mathbf{44.4_{0.4}}$	$50.5_{1.2}$	$39.6_{0.1}$	$\mathbf{36.7_{1.9}}$	$48.0_{4.2}$	$29.7_{0.9}$	$\mathbf{66.2_{2.1}}$	$67.4_{4.6}$	$65.2_{0.8}$	$\mathbf{49.1_{1.5}}$
					$\lvert N \rvert = 4^*$					
FineTune	-	-	-	$10.3_{1.6}$	$59.3_{3.6}$	$5.6_{0.9}$	$43.5_{1.8}$	$85.2_{2.1}$	$29.2_{1.9}$	-
SetFit	$45.6_{1.0}$	$46.4_{1.6}$	$45.1_{0.9}$	$35.0_{0.3}$	$42.6_{1.1}$	$29.7_{0.3}$	$63.6_{1.5}$	$60.7_{2.3}$	$66.9_{1.8}$	$48.1_{0.9}$
Label Embedding	$47.0_{1.1}$	$47.7_{0.6}$	$46.3_{2.0}$	$32.7_{1.1}$	$41.9_{2.3}$	$26.9_{0.6}$	$71.9_{0.6}$	$70.8_{0.5}$	$73.0_{1.2}$	$50.5_{0.9}$
SetFit→LE	$46.0_{0.3}$	$47.4_{0.4}$	$44.7_{0.7}$	$30.7_{2.4}$	$36.9_{4.4}$	$29.5_{2.1}$	$70.0_{0.6}$	$68.1_{0.8}$	$72.1_{0.4}$	$48.9_{1.7}$
LE→SetFit	$45.8_{1.0}$	$45.0_{1.4}$	$46.7_{0.9}$	$32.0_{2.9}$	$40.6_{2.3}$	$27.8_{1.9}$	$66.1_{0.7}$	$62.1_{0.8}$	$70.7_{0.9}$	$48.0_{1.5}$
FusionSent	$\mathbf{48.3_{1.1}}$	$51.0_{1.0}$	$46.0_{1.5}$	$\mathbf{38.5_{2.3}}$	$45.4_{1.0}$	$33.5_{2.9}$	$\mathbf{72.6_{0.5}}$	$72.1_{0.3}$	$73.2_{0.7}$	$\mathbf{53.1_{1.3}}$
					$\lvert N \rvert = 8^*$					
FineTune	$18.7_{2.8}$	$72.5_{2.4}$	$10.8_{1.9}$	$26.6_{3.9}$	$55.0_{3.7}$	$17.5_{3.1}$	$67.1_{1.1}$	$88.0_{1.6}$	$54.2_{1.1}$	$37.5_{2.6}$
SetFit	$46.0_{0.6}$	$44.1_{0.5}$	$48.1_{1.1}$	$37.5_{3.7}$	$46.2_{3.2}$	$31.5_{3.8}$	$66.2_{0.2}$	$66.4_{0.4}$	$66.0_{0.4}$	$49.9_{1.5}$
Label Embedding	$47.8_{1.4}$	$46.3_{1.5}$	$49.5_{1.4}$	$36.4_{3.2}$	$41.3_{2.5}$	$32.5_{3.5}$	$77.2_{0.5}$	$74.0_{0.7}$	$80.8_{0.7}$	$53.8_{1.7}$
SetFit→LE	$45.3_{1.2}$	$45.5_{1.0}$	$45.1_{1.4}$	$30.5_{1.6}$	$31.4_{2.0}$	$29.7_{1.9}$	$72.7_{0.4}$	$70.5_{0.5}$	$75.1_{0.4}$	$49.5_{1.1}$
LE→SetFit	$44.1_{1.0}$	$42.1_{1.8}$	$46.3_{0.5}$	$32.9_{3.0}$	$41.8_{4.1}$	$27.1_{2.3}$	$66.7_{0.4}$	$65.8_{1.3}$	$67.6_{1.1}$	$47.9_{1.5}$
FusionSent	$\mathbf{49.0_{1.3}}$	$49.0_{1.8}$	$49.0_{0.9}$	$\mathbf{41.2_{4.6}}$	$43.4_{5.4}$	$39.2_{3.9}$	$\mathbf{78.3_{0.3}}$	$76.0_{0.2}$	$80.7_{0.5}$	$\mathbf{56.2_{2.1}}$
					$\lvert N \rvert = Full^{**}$					
FineTune	71.6	78.2	66.1	57.9	73.5	47.8	95.9	96.2	95.7	75.1

Table 2: FusionSent performance scores and standard deviations for few-shot classification of scientific documents compared to different approaches across three test datasets and four training set sizes $\lvert N \rvert$. Micro F_1, Precision (P), and Recall (R) scores are reported. *Number of training samples per class. **Entire available training data used. In some cases, insufficient training examples were provided for the FineTune model to learn, resulting in no class predictions during testing.

2021). In our multi-label scientific document classification experiments, we use three random training splits for each dataset and sample size to mitigate this issue. For each method, we report the average measure and the standard deviation across these splits. We use SciNCL (Ostendorff et al., 2022) as the base PLM for each model. While we train the FineTune model for 50 epochs in the few-shot setting, we use the same approach to train a model on the full training datasets for 3 epochs. In both cases, we use a batch size of 12. For sentence embedding model training according to step 1 in Section 4.1, we use a batch size of 4 for all models, and for training according to step 2 in Section 4.1, we use a batch size of 1. Both steps are trained for 1 epoch for all models. In addition, each model is trained with a learning rate of $2e^{-5}$.

For the arXiv dataset, we use the publicly available category descriptions as label texts. For the SciHTC and NLP taxonomy datasets, we generate short descriptive texts from the provided label names with GPT-4 (OpenAI et al., 2023) and use them as label texts. Table 5 shows examples of the used label names and label descriptions.

6 Results

Table 2 shows a comparison between FusionSent and the other few-shot approaches for $\lvert N \rvert \in \{2, 4, 8\}$ labeled training samples per class. We observe that FusionSent consistently outperforms F_1 scores of all approaches investigated for each dataset and training set size. Further, FusionSent significantly outperforms SetFit across all training set sizes by an average of 6.0 F_1 points. While the other approaches using label texts for sentence embedding training on a single model can perform better than SetFit, they fall short of the FusionSent approach. The Label Embedding (LE) approach shows consistent improvements over SetFit on average, while the SetFit→LE and LE→SetFit approaches only outperform SetFit in a few cases.

The results demonstrate that using label texts for sentence embedding training can help to separate instances of different classes in the embedding space, providing a crucial property for the classification head to perform well. However, combining the contrastive sentence embedding training approaches of SetFit and LE in a single model does not significantly increase performance. Using a two-step contrastive training approach does not enable a single sentence embedding model to effec-

tively encode information from both training steps, as it may suffer from forgetting previously acquired knowledge (Biesialska et al., 2020). Conversely, the FusionSent results indicate that this limitation can be circumvented by training separate sentence embedding models with different contrastive learning approaches and subsequently merging their parameters. This approach ensures that the individually trained models encode different information, whereas merging allows their respective knowledge to complement each other, resulting in improved model performance.

Dataset →	SST-5	AGNews	Emotion	EnronSpam	CR	Average
Method ↓	Multi-class Classification			Binary Classification		
	$|N| = 8^*$					
FineTune[†]	$33.5_{2.1}$	$81.7_{3.8}$	$28.7_{6.8}$	$85.0_{6.0}$	$58.8_{6.3}$	$57.5_{5.0}$
SetFit	$41.7_{2.0}$	$82.6_{3.6}$	$49.5_{3.8}$	$91.0_{3.2}$	$89.6_{1.2}$	$70.9_{2.8}$
FusionSent	$43.0_{3.2}$	$84.4_{2.2}$	$57.1_{2.5}$	$91.4_{3.9}$	$89.8_{1.0}$	$73.1_{2.6}$
	$|N| = 64^*$					
FineTune[†]	$45.9_{6.9}$	$88.4_{0.9}$	$65.0_{17.2}$	$95.9_{0.8}$	$88.9_{1.9}$	$76.8_{5.5}$
SetFit	$48.1_{4.4}$	$87.7_{0.8}$	$78.5_{2.0}$	$96.1_{0.5}$	$90.6_{0.7}$	$80.2_{1.7}$
FusionSent	$50.0_{2.8}$	$88.3_{0.8}$	$78.7_{1.6}$	$96.5_{0.6}$	$90.7_{0.6}$	$80.8_{1.3}$
	$|N| = Full^*$					
FineTune[†]	59.8	93.8	92.6	99.0	92.4	87.5

Table 3: FusionSent accuracy scores and standard deviations for few-shot classification with few classes compared to the baselines across five test datasets for three training set sizes. *Number of training samples per class. **Entire available training data used. [†]Results from Tunstall et al. (2022).

7 Experiments with Few Classes

To determine the generalizability of FusionSent to few-shot settings with a low number of classes and different domains, we perform experiments on the SST-5 (Socher et al., 2013), CR (Hu and Liu, 2004), Emotion (Saravia et al., 2018), AGNews (Zhang et al., 2015), and EnronSpam (Metsis et al., 2006) datasets as processed for few-shot classification by Tunstall et al. (2022). These datasets each comprise two to six classes and cover the fields of sentiment classification, opinion detection from customer reviews, emotion detection from *Twitter* tweets, news article classification, and e-mail spam detection. For all datasets, we generate short descriptive texts from the provided label names with GPT-4 and use them as label texts. Table 5 shows examples of the used label names and label descriptions. We experiment with all ten available randomized training splits for each dataset and sample size using $|N| = 8$ and $|N| = 64$ few-shot examples. We report the average accuracy and standard deviation across the training splits for each method.

We use the paraphrase-mpnet-base-v2[6] model (Reimers and Gurevych, 2019) as base PLM for SetFit and FusionSent. Additionally, we use RoBERTa$_{LARGE}$ (Liu et al., 2019) for the Fine-Tune baseline. The other training parameters remain the same as in Section 5.3.

Results Table 3 shows the results of SetFit and FusionSent training on the binary and multi-class datasets. On average, FusionSent outperforms Set-Fit for $|N| = 8$ by an average of 2.2 accuracy points. However, as the number of training samples increases to $|N| = 64$, the gap decreases to 0.6 accuracy points. In addition, the improvements

---

[6]https://huggingface.co/sentence-transformers/paraphrase-mpnet-base-v2

are more substantial for multi-class classification, whereas they are only minimal in the binary case.

For binary classification, instances of different classes must be pushed apart from each other as far as possible to allow the logistic regression classifier to find a good decision boundary. For this relatively simple problem, the results indicate that SetFit can already effectively separate instances of different classes in embedding space. For multiple classes, however, it is more difficult to find positions in the embedding space that separate instances of different classes from each other. In these cases in particular, FusionSent shows its strengths and helps the classification head to find good decision boundaries between classes.

8 Robustness Against Label Text Variations

To evaluate the robustness of FusionSent against different label text variations, we conduct experiments on the previous datasets using simple label names instead of extensive label descriptions. In these experiments, we simply use the label names as provided by the respective datasets and compare the classification results with those obtained by using detailed label descriptions. Table 5 shows examples of the used label names and label descriptions. We use $|N| = 8$ few-shot examples and report the average F_1 performance over the respective training splits. Furthermore, we use the same training parameters as in Section 5.3 and Section 7.

Results Figure 3 shows the performance of FusionSent using simple label names, as provided by the respective datasets, compared to using exten-
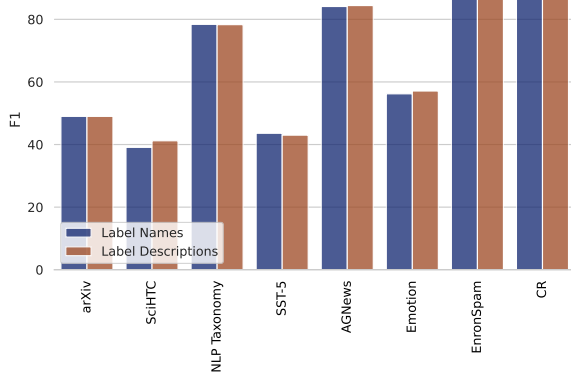
Figure 3: FusionSent micro F_1 scores for few-shot classification on 8 different datasets using either extensive label descriptions or simple label names. We report the average score over the random training splits of each dataset using $|N| = 8$ training examples per class.

sive textual label descriptions generated by GPT-4. We obtain similar performances across the different label text variants with a mean performance difference of 0.48 and a standard deviation of 0.39 F_1 points. Furthermore, there is no clear pattern as to whether the use of extensive label descriptions or simple label names leads to significantly improved performance. In comparison, performance variations of 10 accuracy points and more when using different prompts on the same model are characteristic for prompt-based few-shot classification approaches (Perez et al., 2021). Therefore, we conclude that FusionSent is relatively robust to label text variations and the use of simple label names is already sufficient to achieve good classification performance.

9 Computational Costs

To compare the relative computational costs of FusionSent and SetFit, we follow the approach of Liu et al. (2022) and use FLOPs-per-token estimates, which can be obtained from Kaplan et al. (2020). Specifically, encoder-only models with N parameters have approximately $2N$ FLOPs-per-token for inference and $6N$ FLOPs-per-token for training. The resulting cost for inference and training is then given by:

$$C_{inf} = 2N \cdot \ell_{seq}, \qquad (4)$$

$$C_{train} = 6N \cdot \ell_{seq} \cdot n_{steps} \cdot n_{batch}, \qquad (5)$$

where ℓ_{seq} is the input sequence length, n_{steps} is the number of training steps, and n_{batch} is the

batch size. Since we are training two model bodies for FusionSent, we calculate the training costs for each model body separately and then add them up. For inference, we can use the formula as provided, since we only use one model body.

We estimate the costs using the scientific document classification datasets from Table 2 and SciNCL as base PLM with $N = 110M$ parameters. Based on the median number of tokens per instance in all datasets, we use $\ell_{seq} = 194$ to estimate the costs for training approaches that do not use label texts. Since we perform inference on these instances, we also use this value to estimate the inference cost for all approaches. Taking into account the shorter label texts, we use $\ell_{seq} = 130$ to estimate the costs for training approaches that utilize label texts. Additionally, we use fixed values of $n_{steps} = 1,000$, and $n_{batch} = 8$ for all training estimates.

Method	Inf. FLOPs	Train FLOPs	Avg. F_1
SetFit	$4.3e10$	$1.0e15$	$49.9_{1.5}$
FusionSent	$4.3e10$	$1.7e15$	$56.2_{2.1}$

Table 4: Computational costs and average micro F_1 scores of FusionSent and SetFit using $|N| = 8$ training samples on the scientific document classification datasets listed in Table 2.

As shown in Table 4, the increase in F_1 performance is accompanied by increased training costs. This is the result of training two sentence embedding models instead of one. However, by merging the models, the inference efficiency remains the same as when using the base PLM. Although FusionSent incurs higher training costs, it can significantly improve prediction performance while maintaining SetFit's inference efficiency.

10 Conclusion

We introduce FusionSent, a new approach for efficient and prompt-free few-shot classification of scientific documents. FusionSent uses label texts and contrastive learning to improve classification performances over several other few-shot approaches. We show that FusionSent is particularly effective in scenarios with many classes while being computationally efficient during inference. Additionally, FusionSent is robust against label text variations. Finally, we introduce a new arXiv dataset for multi-label classification of scientific documents.

References

Rami Aly, Xingjian Shi, Kaixiang Lin, Aston Zhang, and Andrew Wilson. 2023. Automated few-shot classification with instruction-finetuned language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2414–2432, Singapore. Association for Computational Linguistics.

Luke Bates and Iryna Gurevych. 2024. Like a good nearest neighbor: Practical content moderation and text classification. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 276–297, St. Julian's, Malta. Association for Computational Linguistics.

Magdalena Biesialska, Katarzyna Biesialska, and Marta R. Costa-jussà. 2020. Continual lifelong learning in natural language processing: A survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6523–6541, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Colin B. Clement, Matthew Bierbaum, Kevin P. O'Keeffe, and Alexander A. Alemi. 2019. On the use of arxiv as a dataset.

D. R. Cox. 1958. The regression analysis of binary sequences. *Journal of the Royal Statistical Society. Series B (Methodological)*, 20(2):215–242.

Yuxiao Dong, Hao Ma, Zhihong Shen, and Kuansan Wang. 2017. A century of science: Globalization of scientific collaborations, citations, and innovations. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, page 1437–1446, New York, NY, USA. Association for Computing Machinery.

Óscar E. Mendoza, Wojciech Kusa, Alaa El-Ebshihy, Ronin Wu, David Pride, Petr Knoth, Drahomira Herrmannova, Florina Piroi, Gabriella Pasi, and Allan Hanbury. 2022. Benchmark for research theme classification of scholarly documents. In *Proceedings of the Third Workshop on Scholarly Document Processing*, pages 253–262, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3816–3830, Online. Association for Computational Linguistics.

Nikolaos Gialitsis, Sotiris Kotitsas, and Haris Papageorgiou. 2022. Scinobo: A hierarchical multi-label classifier of scientific publications. In *Companion Proceedings of the Web Conference 2022*, WWW '22, page 800–809, New York, NY, USA. Association for Computing Machinery.

R. Hadsell, S. Chopra, and Y. LeCun. 2006. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742.

Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. 2022. Towards a unified view of parameter-efficient transfer learning. In *International Conference on Learning Representations*.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, page 168–177, New York, NY, USA. Association for Computing Machinery.

Yongxin Huang, Kexin Wang, Sourav Dutta, Raj Patel, Goran Glavaš, and Iryna Gurevych. 2023. AdaSent: Efficient domain-adapted sentence embeddings for few-shot classification. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3420–3434, Singapore. Association for Computational Linguistics.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. 2023. Editing models with task arithmetic. In *The Eleventh International Conference on Learning Representations*.

Dongfu Jiang, Xiang Ren, and Bill Yuchen Lin. 2023. LLM-blender: Ensembling large language models with pairwise ranking and generative fusion. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14165–14178, Toronto, Canada. Association for Computational Linguistics.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models.

Rabeeh Karimi Mahabadi, Luke Zettlemoyer, James Henderson, Lambert Mathias, Marzieh Saeidi, Veselin Stoyanov, and Majid Yazdani. 2022. Prompt-free and efficient few-shot learning with language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3638–3652, Dublin, Ireland. Association for Computational Linguistics.

Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov. 2015. Siamese neural networks for one-shot image recognition. volume 2 of *ICML deep learning workshop*.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597, Online. Association for Computational Linguistics.

N. Littlestone and M.K. Warmuth. 1994. The weighted majority algorithm. *Information and Computation*, 108(2):212–261.

Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin A Raffel. 2022. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. In *Advances in Neural Information Processing Systems*, volume 35, pages 1950–1965. Curran Associates, Inc.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Andrew Kachites McCallum, Kamal Nigam, Jason Rennie, and Kristie Seymore. 2000. Automating the construction of internet portals with machine learning. *Inf. Retr.*, 3(2):127–163.

Vangelis Metsis, Ion Androutsopoulos, and Georgios Paliouras. 2006. Spam filtering with naive bayes-which naive bayes? In *CEAS*, volume 17, pages 28–69. Mountain View, CA.

Kristine Monteith, James L. Carroll, Kevin Seppi, and Tony Martinez. 2011. Turning bayesian model averaging into bayesian model combination. In *The 2011 International Joint Conference on Neural Networks*, pages 2657–2663.

Ghulam Mustafa, Muhammad Usman, Lisu Yu, Muhammad Tanvir afzal, Muhammad Sulaiman, and Abdul Shahid. 2021. Multi-label classification of research articles using word2vec and identification of similarity threshold. Scientific Reports.

OpenAI, :, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mo Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak,

Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2023. Gpt-4 technical report.

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

Ethan Perez, Douwe Kiela, and Kyunghyun Cho. 2021. True few-shot learning with language models. In *Advances in Neural Information Processing Systems*, volume 34, pages 11054–11070. Curran Associates, Inc.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Mobashir Sadat and Cornelia Caragea. 2022. Hierarchical multi-label classification of scientific documents. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages

8923–8937, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Omer Sagi and Lior Rokach. 2018. Ensemble learning: A survey. *WIREs Data Mining and Knowledge Discovery*, 8(4):e1249.

Angelo A. Salatino, Francesco Osborne, Thiviyan Thanapalasingam, and Enrico Motta. 2019. The cso classifier: Ontology-driven detection of research topics in scholarly articles. In *Digital Libraries for Open Knowledge*, pages 296–311, Cham. Springer International Publishing.

António Paulo Santos and Fátima Rodrigues. 2009. Multi-label hierarchical text classification using the acm taxonomy. In *14th portuguese conference on artificial intelligence (epia)*, volume 5, pages 553–564. Springer Berlin.

Elvis Saravia, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu, and Yi-Shin Chen. 2018. CARER: Contextualized affect representations for emotion recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3687–3697, Brussels, Belgium. Association for Computational Linguistics.

Timo Schick and Hinrich Schütze. 2021. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269, Online. Association for Computational Linguistics.

Tim Schopf, Karim Arabi, and Florian Matthes. 2023. Exploring the landscape of natural language processing research. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 1034–1045, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Tim Schopf and Florian Matthes. 2024. NLP-KG: A system for exploratory search of scientific literature in natural language processing. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 127–135, Bangkok, Thailand. Association for Computational Linguistics.

Zhihong Shen, Hao Ma, and Kuansan Wang. 2018. A web-scale system for scientific knowledge exploration. In *Proceedings of ACL 2018, System Demonstrations*, pages 87–92, Melbourne, Australia. Association for Computational Linguistics.

Ken Shoemake. 1985. Animating rotation with quaternion curves. In *Proceedings of the 12th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '85, page 245–254, New York, NY, USA. Association for Computing Machinery.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for

semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Derek Tam, Rakesh R. Menon, Mohit Bansal, Shashank Srivastava, and Colin Raffel. 2021. Improving and simplifying pattern exploiting training. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4980–4991, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Autumn Toney and James Dunham. 2022. Multi-label classification of scientific research documents across domains and languages. In *Proceedings of the Third Workshop on Scholarly Document Processing*, pages 105–114, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Lewis Tunstall, Nils Reimers, Unso Eun Seo Jo, Luke Bates, Daniel Korat, Moshe Wasserblat, and Oren Pereg. 2022. Efficient few-shot learning without prompts.

Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. 2022. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 23965–23998. PMLR.

Nan Xu, Fei Wang, Mingtao Dong, and Muhao Chen. 2023a. Dense retrieval as indirect supervision for large-space decision making. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15021–15033, Singapore. Association for Computational Linguistics.

Ran Xu, Yue Yu, Joyce Ho, and Carl Yang. 2023b. Weakly-supervised scientific document classification via retrieval-augmented multi-stage training. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, page 2501–2505, New York, NY, USA. Association for Computing Machinery.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin Raffel, and Mohit Bansal. 2023. TIES-merging: Resolving interference when merging models. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Pengcheng Yang, Xu Sun, Wei Li, Shuming Ma, Wei Wu, and Houfeng Wang. 2018. SGM: Sequence generation model for multi-label classification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3915–3926, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Asaf Yehudai and Elron Bendel. 2024. When llms are unfit use fastfit: Fast and effective text classification with many classes.

Le Yu, Bowen Yu, Haiyang Yu, Fei Huang, and Yongbin Li. 2023. Language models are super mario: Absorbing abilities from homologous models as a free lunch.

Tianyi Zhang, Felix Wu, Arzoo Katiyar, Kilian Q Weinberger, and Yoav Artzi. 2021. Revisiting few-sample {bert} fine-tuning. In *International Conference on Learning Representations*.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Appendix

Dataset	Label Names	Label Descriptions
arXiv	General Relativity and Quantum Cosmology ...	General relativity and quantum cosmology focuses on gravitational physics, including experiments and observations related to the detection and interpretation of gravitational waves, experimental tests of gravitational theories, computational general relativity, relativistic astrophysics, solutions to Einstein's equations and their properties, alternative theories of gravity, classical and quantum cosmology, and quantum gravity. ...
SciHTC	Information retrieval ...	The "Information Retrieval" class within the 2012 ACM Computing Classification System encompasses the study and design of systems for indexing, searching, and retrieving information from large datasets. It includes the development of algorithms and techniques for processing and querying textual and multimedia data, as well as evaluating the effectiveness of retrieval systems. Key topics within this class involve search engine architectures, query representation, relevance feedback, and information extraction. The field also addresses challenges such as handling unstructured data, understanding user context, and ensuring privacy and security in the retrieval process. ...
NLP Taxonomy	Named Entity Recognition ...	Named Entity Recognition is the identification and classification of entities (e.g., names of people, organizations) in text. ...
SST-5	very positive ...	'very positive' is used for data samples that express strong or intense positive sentiments, enthusiasm, or approval. ...
AGNews	Sports ...	'Sports' represents data samples related to sports news, events, scores, and athlete performances. ...
Emotion	sadness ...	'sadness' is characterized by feelings of hopelessness, disappointment, melancholy, and vulnerability, often accompanied by a sense of isolation or being overwhelmed. ...
EnronSpam	spam ...	'spam' is an unsolicited and often irrelevant or inappropriate message sent over the internet, typically to a large number of users, for the purpose of advertising, phishing, spreading malware, or other malicious activities. ...
CR	negative ...	'negative' corresponds to criticisms, complaints, or expressions of dissatisfaction with products or services. ...

Table 5: Examples of label names and extensive label descriptions for different datasets.