Empirical Study on Text Classification of Small Science Domain Datasets

Shanshan Liu[1] Masashi Ishii[2] Yuji Matsumoto[1]
[1]Center for Advanced Intelligence Project, RIKEN
[2]Material Database Group, MaDIS, NIMS
{shanshan.liu, yuji.matsumoto}@riken.jp
ishii.masashi@nims.go.jp

Abstract

As for the classification of scientific texts, there are no clear states on how much data is needed for model training and what performance mainstream text classifiers can achieve. This paper uses the number of training papers as a variable to train the models for sentence classification tasks on the thermoelectric material synthesis procedure dataset and the polymer biodegradability dataset, and analyzes the results of two classifiers based on LLaMA-2-7b and SciBERT-base. We aim to provide model baselines and corresponding data requirements for reference.

1 Introduction

For information extraction tasks on scientific texts, the first important step is to obtain sentences or text blocks in the literature that describe the target information.

In our previous work [1], we found that the performance of text classification[1)] has a huge impact on subsequent entity recognition and relation extraction tasks. However, further performance improvement is quite challenging.

The application scenarios of scientific information extraction are mostly to extract a specific type of information (such as the material synthesis procedure) in specific domains. It is difficult to collect enough labelled data to adequately train a classifier, and few language models have been pretrained with the domain text. Additionally, there is a lack of empirical research on this task. Advanced text classifiers are often only validated on general text datasets (e.g., the sentiment of financial news [2] or movie reviews [3]). Researchers have difficulty estimating data requirements and model performance.

To remove some hindrances to further research, we report the results of two mainstream text classifiers on two datasets we built – Thermoelectric Material Synthesis Procedure (TMSP) dataset and Polymer Biodegradability (P-BIO) dataset.

Scientific text classifiers based on SciBERT [4] were dominant before the emergence of the large language models (LLMs), and with the development of the LLMs, studies using LLaMA-2-7b [5] for text classification have produced top-of-the-line results on general benchmarks [6]. SciBERT-base is trained on computer science and broad biomedical domain, which has been widely used in various tasks addressing scientific texts. LLaMA-2-7b is an open LLM trained on 2 trillion tokens of data. Its promising performance in a variety of tasks has made it a hot topic in recent researches. Therefore, these two pretrained models can be used as suitable baselines.

Our works may help to answer two questions:

- How much annotated data (number of papers) is needed for supervised training of text classifier?
- How do the LLaMA-2-7b- and SciBERT-based classifiers perform on small scientific datasets?

Depending on the characteristics of the text, different amounts of training data are required. The more specialized texts demands more data and in-domain language models. There is huge potential for supervised learning using small datasets - 5 TMSP-type papers （around 20 positive sentences）/ 20 P-BIO-type papers （arround 330 positive sentences） can be trained to achieve acceptable performance. Even if LLMs perform well in general text classification, they may not be good for scientific text. SciBERT-based classifier is a trivial but effective solution for texts in which expertise terms are more common.

[1)] ”Text classification” in this paper refers to sentence classification.

Table 1 Statistics of Datasets

	Train			Dev			Test		
	#Paper	Avg. #Sent	Avg. #Pos Sent	#Paper	Avg. #Sent	Avg. #Pos Sent	#Paper	Avg. #Sent	Avg. #Pos Sent
TMSP	368	163.4	5.04	93	154.6	5.0	115	164.7	5.0
P-BIO	50	189.3	16.5*	5	125.4	6.4	5	100.0	6.8

#Paper, Avg. #Sent, Avg. #Pos Sent are the number of papers, the average number of sentences per paper, and the average number of positive sentences per paper.

* Training set includes several polymer biodegradability reviews, making the distribution different from the Test set.

Table 2 Instance examples

Index	Dataset	Label	Sentence
1	TMSP	Pos	Polycrystalline samples of **CaCu3Ti4-xRuxO12** were prepared by a solid-state reaction method.
2	TMSP	Pos	The product was finely **ground**, **pressed** into a pellet, and **sintered in air** at **1000-1050 •C** for **20 h**.
3	TMSP	Neg	In the solid solution of *CaCu3Ti4-xRuxO12*, one can observe how the localized electrons on the Cu2+ sites become itinerant with *Ru* substitution.
4	P-BIO	Pos	The enzymatic degradation of the rubber polymer **poly(cis-1,4-isoprene)**, e.g. by the latex clearing protein Lcp1VH2 of **Gordonia polyisoprenivorans VH2** has been demonstrated with latex milk or pure isoprene-rubber particles, recently.
5	P-BIO	Neg	*Polyethylene* degrading ability of the isolates has been assessed individually in a synthetic media containing *polyethylene* as a carbon source.

Note: Bold mentions are the salient information the subsequent extraction tasks focus on. Italicized mentions have the same expression as entities that need to be identified, but do not need to be extracted in subsequent tasks.

2 Method

We evaluate the text classifiers based on LLaMA-2-7b and SciBERT-base pretrained language models (PLMs). To investigate the quantitative demands on the training data, we randomly selected different numbers of papers from the training set for fine-tuning while keeping the validation and test sets unchanged. On the TMSP dataset, the experiments are carried out with [5, 10, 20, 30, 40, 50, 60, 80, 100, 120, 140, 150, 160, 180, 200, 250, 300, 350] papers, while on the P-BIO dataset, number of papers for training is [5, 10, 20, 30, 40, 50] respectively.

3 Experiment

3.1 Dataset

TMSP The Thermoelectric Materials Synthesis Procedure dataset contains 576 papers. Materials, processes, conditions of the process, and other relevant information are included in the target information. The sentences describing material synthesis procedures are positive.

P-BIO Polymer Biodegradability dataset, a dataset containing 60 papers. This dataset is built to extract the relation between the polymer and the microorganism (bacteria/fungi) that can degrade it from scientific papers. The sentences with both polymer names and microorganism names is positive.

Statistics of datasets are shown in Table 1. We show some sample sentences of both datasets in Table 2.

3.2 Model

The SciBERT-based classifier consists of the SciBERT-base[2], two MLP layers and one linear layer. The last hidden state of the first token - [CLS] token - is passed to two consecutive MLP layers, and then a linear layer for classification. To determine the effect of domain difference in pre-training, we also utilized the BERT-based[3] model on P-BIO dataset. The experimental setup using BERT is the same as for SciBERT.

As for the model using LLaMA-2-7b[4], we follow the work of Li et al., 2023 [6]: LS-LLaMA, and used the same fine-tuning strategy, i.e. fine-tuning with Low-Rank Adaptation (LoRA) [7] to minimize the cross-entropy loss. LS-LLaMA combines LLaMA-2-7b with a linear layer and classifies the sequence by the score over the last token.

[2] https://huggingface.co/allenai/scibert_scivocab_uncased

[3] https://huggingface.co/bert-base-uncased

[4] https://huggingface.co/meta-llama/Llama-2-7b-hf

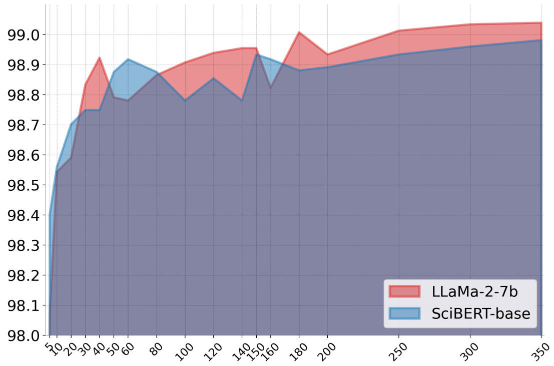

Figure 1 Accuracy on the TMSP dataset

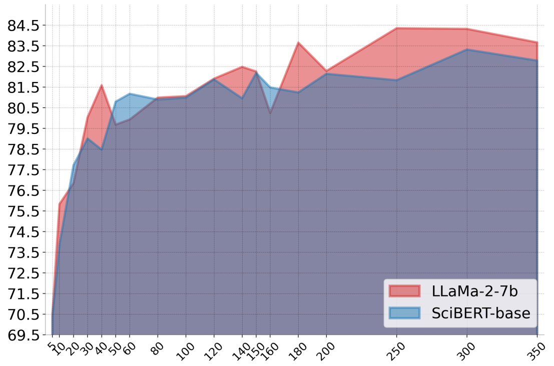

Figure 2 F1 score on the positive class of TMSP dataset

Considering that we are doing supervised fine-tuning on small datasets, only LLaMA-2-7b is utilized in this work.

We run 5- / 3-epoch fine-tuning on SciBERT- / LLaMA-2-7b-based classifiers and the model with the highest accuracy on the development set is used for testing.

3.3 Evaluation metric

We report the precision (P), recall (R) and micro F1 score (F) on the positive class and the accuracy (A) for the binary classification.

4 Results

4.1 Data requirement

Two datasets reflect difficulties in categorising scientific texts. One is that the same expression may appear in both positive and negative examples, such as the name of a synthetic material (e.g. CaCu3Ti4-xRuxO12) or a degraded polymer (e.g. polyethylene) (see Table 2), which requires the classifier to capture the implication of the entire sentence. Secondly, with limited data and even fewer positive examples, the classifier may not be adequately trained.

The target sentences of TMSP contain fewer professional terms, and the way of expressing is easy to understand. Our experiments show that on the TMSP, both of classifiers can achieve an F1 of more than 80% when using 80 papers (Fig.2), and the accuracy is also above 98.7% (Fig.1). As the number of training papers increases, they can finally reach the highest value when using 300/250 papers (SciBERT: 83.32% F1; LLaMA-2-7b: 84.35% F1). (The whole experiment results is in Appendix A.1)

For P-BIO, a sentence in which the polymer name coexists with the name of the microorganism is regard as positive. Microorganism names are not common words (e.g. Gordonia polyisoprenivorans VH2), and the sentences are often complex and require expertise. To exceed 80% F1, SciBERT needs 20 papers (84.85% F1, 98.00% Acc) while 40 papers are as necessary for LLaMA-2-7b (91.43% F1, 98.80% Acc). The best results contributed by SciBERT-based model with 50 papers (91.67% F1, 98.80% Acc). As can be seen from Fig.3 that the SciBERT's curve is still in an upward phase and the LLaMA-2-7b's curve is oscillating, 50 papers are far from enough to get the best out of both models.

The results indicate that on tasks with text properties similar to TMSP - less professional nouns, wording and sentence construction similar to general text - only 5 papers (about 100 positive examples) for training achieves an F1 above 70% and an accuracy over 98%. However, on a narrow spectrum of text - P-BIO, achieving a 70% F1 score requires an in-domain PLM (SciBERT in our case) and more than 30 papers.

4.2 LLaMA-2-7b v.s SciBERT-base

Even though they are both scientific texts, the two datasets focus on two types of information, two different scientific domains. Therefore, the validity of the two models is quite different.

On TMSP, the two models have similar effects in the early stage, but as the data increases, LLaMA-2-7b shows off its "large" model capabilities (84.35% F1, 98.93% Acc, 250 papers) and clearly outperforms SciBERT.

What surprises us is: The results of LLaMA-2-7b are always significantly lower than those of SciBERT on P-BIO except with 5 or 40 training papers.

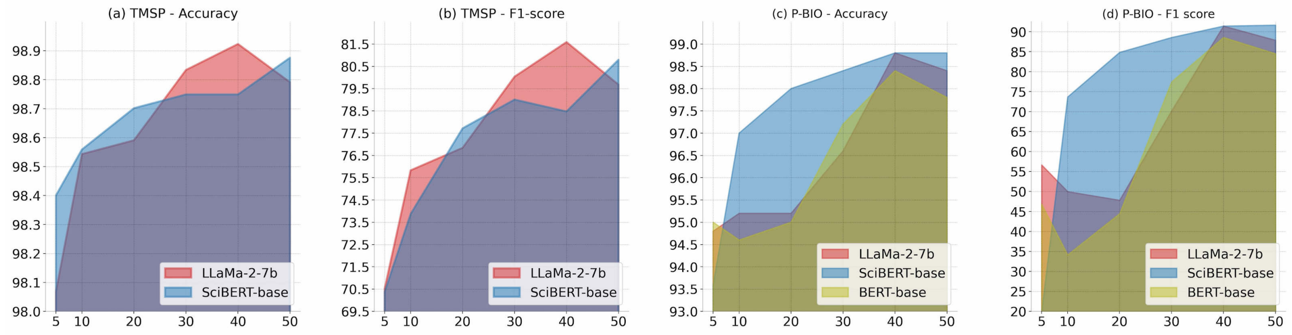

Figure 3 Accuracy and F1 score of positive classes on TMSP and P-BIO datasets

Even using the entire dataset, LLaMA-2-7b turns in less than ideal numbers. We repeated the experiment of training LLaMA-2-7b using 50 papers three times and obtained the same results in first two runs: only 2 out of 34 positive examples were successfully predicted. At the third time, the model was trained well (87.88% F1, 98.40% Acc) but still is inferior to SciBERT. We report the result of the last case in Figure 3 and Table 4.

We conjecture that when available data is not sufficient, changes in data distribution have a significant impact on LLaMA-2-7b. As shown in Table 1, the use of review papers (which lists a large number of polymers and corresponding decomposition microorganisms) makes obvious discrepancy in the distributions between the training set and development set / test set. When part of data is used, random sampling may not select review articles, allowing LLaMA-2-7b to achieve the same results as SciBERT with 40 articles. Compared with SciBERT, LLaMA-2-7b is not robust enough to distribution changes.

It is agnostic whether LLaMA-2-7b will surpass SciBERT after the training set exceeds more than 80 papers, as it did on TMSP. The annotation dataset, however, cannot be further expanded easily because there are few relevant studies in this domain, and fewer papers can be obtained. Annotating this type of information also requires a lot more effort from experts than generic text.

The weak performance of LLaMA-2-7b on P-BIO emphasizes the importance of specific domain pretrained models for scientific text. In the range of 5 - 50 papers, SciBERT is just on par with the LLaMA-2-7b on TMSP. However, on P-BIO, benefiting from pretrained on biomedical domain corpus, SciBERT has reached an F1 score and accuracy (91.71% F1, 98.80% Acc, 50 papers) that is much higher than that on TMSP (Best: 80.79% F1, 98.88% Acc, 50 papers). SciBERT's effectiveness demonstrates the potential of small scientific datasets. Training with 50 papers can obtain an F1 score of up to 91.67%, which is enough to promote subsequent information extraction tasks, such as entity recognition and relation extraction.

4.3 Specific domain PLM matters

The performance of BERT is close to that of LLaMA-2-7b but far from that of SciBERT in (c) and (d) of Fig.3. While their pre-training data and model sizes are not in the same order of magnitude at all, LLaMA-2-7b has only a slight advantage over BERT on P-BIO. This is strong evidence that SciBERT's strong performance on the P-BIO dataset is due to the domain of the pre-training dataset rather than its model structure. Despite the pre-training corpus containing the biomedical domain, which does not exactly overlap with the domain of P-BIO, the similarities in the textual characteristics of the two domains are sufficient to produce more desirable results. Choosing a similar domain dataset based on text characteristics for unsupervised pre-training of PLMs may be a workable option if no big corpus is available from that domain.

5 Conclusion

We have conducted a study on the amount of data required to train a scientific sentence classifier and the effect of classifiers with different pretrained language models on two types of texts. For highly specialized texts, more training data is required, with 20 papers and in-domain pretrained models needed to achieve results above 80% F1 score. For near-generic scientific texts, only 5 papers are necessitated to exceed the qualifying line of 80% F1 score, and the large generalized language model performs well with more data. We encourage researchers to construct datasets and select base models for text classification tasks based on the text characteristics.

6 Acknowledgement

We thank Chowdhury Mohammad Mahir Asef and Mutsunori Uenuma for their contribution to the construction of the Thermoelectric Material Synthesis Procedure (TMSP) dataset.

References

[1] Tatsuya Ishigaki, Yui Uehara, Shanshan Liu, Topic Goran, and Hiroya Takamura. Machine learning-based knowledge acquisition from material science literature. 2021.

[2] The twitter financial news dataset. https://huggingface.co/datasets/zeroshot/twitter-financial-news-sentiment.

[3] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In David Yarowsky, Timothy Baldwin, Anna Korhonen, Karen Livescu, and Steven Bethard, editors, **Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing**, pp. 1631–1642, Seattle, Washington, USA, October 2013. Association for Computational Linguistics.

[4] Iz Beltagy, Kyle Lo, and Arman Cohan. Scibert: A pretrained language model for scientific text, 2019.

[5] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023.

[6] Zongxi Li, Xianming Li, Yuzhang Liu, Haoran Xie, Jing Li, Fu lee Wang, Qing Li, and Xiaoqin Zhong. Label supervised llama finetuning, 2023.

[7] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021.

Table 3 Results on the TMSP dataset

#Paper	Accuracy (%)		F1 score on the positive class (%)	
	LLaMA-2-7b	SciBERT-base	LLaMA-2-7b	SciBERT-base
5	98.07	98.40	70.53	70.38
10	98.54	98.56	75.83	73.88
20	98.59	98.70	76.84	77.72
30	98.83	98.75	80.04	79.01
40	98.92	98.75	81.59	78.47
50	98.79	98.88	79.68	80.79
60	98.78	98.92	79.93	81.18
80	98.87	98.88	80.99	80.90
100	98.91	98.78	81.06	80.99
120	98.94	98.85	81.91	81.87
140	98.95	98.78	82.48	80.96
150	98.95	98.93	82.26	82.19
160	98.82	98.92	80.25	81.48
180	99.01	98.88	83.65	81.24
200	98.93	98.89	82.28	82.14
250	99.01	98.93	**84.35**	81.83
300	99.03	98.96	84.32	**83.32**
350	**99.04**	**98.98**	83.66	82.78

Table 4 Results on the P-BIO dataset

#Paper	Accuracy (%)			F1 score on the positive class (%)		
	LLaMA-2-7b	SciBERT-base	BERT-base	LLaMA-2-7b	SciBERT-base	BERT-base
5	94.80	93.60	95.00	56.67	20.00	46.81
10	95.20	97.00	94.60	50.00	73.68	34.15
20	95.20	98.00	95.00	47.83	84.85	44.44
30	96.60	98.40	97.20	70.18	88.57	77.42
40	**98.80**	**98.80**	**98.40**	**91.43**	91.43	**88.57**
50	98.40	**98.80**	97.80	87.88	**91.67**	84.51

A Appendix

A.1 Results on the TMSP dataset

Please refer to Table 3 for the experimental results on the TMSP dataset.

A.2 Results on the P-BIO dataset

Please refer to Table 4 for the experimental results on the TMSP dataset.