# Efficient Training and Stacking Methods with BERTs-LightGBM for Paper Source Tracing

Xinlong Huang
Shenzhen University
Shenzhen, Guangdong, China
2210295079@email.szu.edu.cn

Qinglun Wang
Onewo Space-Tech Service Co., Ltd.
Shenzhen, Guangdong, China
wangql43@vanke.com

Guibin Wang*
Shenzhen University
Shenzhen, Guangdong, China
wanggb@szu.edu.cn

## Abstract

Despite the rapid development of large language models in recent years, tracing the source of scientific papers remains a challenging task due to the large-scale of citation relations between papers. To address this challenge, the PST-KDD-2024 competition is launched by the Knowledge Engineering Group (KEG) of Tsinghua University, in collaboration with ZhipuAI. In the competition, the Heart algorithm team has proposed an innovative method that includes three key strategies: adversarial weight pertubation (AWP), concatenate mean pooling (ConcatPool), and weighted ensemble. The core highlight of our approach is the application of AWP and ConcatPool to several BERT-based pretraining models, which significantly improves the performance and robustness of these models. In addition, the weighted ensemble strategy integrates the results of multiple models, including the BERT-based models and LightGBM, to leverage their strengths to produce more robust and accurate results. Through the test benchmark, we have achieved a significant improvement in the performance metric, with the test score increasing from 0.32042 to 0.41778. With our innovative solution, our team Heart won the 7th place in the final leaderboard of the PST-KDD-2024 competition.The implementation details and code are publicly available at this link: https://github.com/Heartttttt1/PST-KDD-2024-Heart-Rank7

## Keywords

Pretraining Model, Pooling Methods, Adversarial Training, Weighted Ensemble Models

**ACM Reference Format:**
Xinlong Huang, Qinglun Wang, and Guibin Wang. 2024. Efficient Training and Stacking Methods with BERTs-LightGBM for Paper Source Tracing. In *Proceedings of ACM KDD PST-OAG-Challenge (Conference KDD '2024).* ACM, Barcelona, Spain, 4 pages. https://doi.org/XXXXXXX.XXXXXXX

## 1 Introduction

In recent years, large language models (LLMs) have been extensively deployed across diverse fields, including natural language processing (NLP), image recognition, and decision-making systems. Despite their exceptional performance in multiple domains, an automatic algorithm for paper source tracing (PST) remains absent, impeding researchers from comprehensively understanding the evolution of science. The lack of such an algorithm not only hampers the innovation of science but also poses challenges to academic discourse and collaborative research endeavors [9]. Therefore, the objective of this competition is to enable developers to design an algorithm that can automatically identify the most significant references in a paper, i.e. source references, to facilitate the tracing of scientific progress.

Our team, leveraging our experience in text classification, achieved seventh place in the competition. Figure 1 illustrates our solution approach, which incorporates three key strategies:

(1) Adversarial weight perturbation (AWP) : A method that introduces small perturbations to model weights during training to enhance robustness and generalization.

(2) Concatenate mean pooling (ConcatPool): An approach that combines different pooling techniques to capture both local and global features of the text.

(3) Weighted Ensemble: A technique that integrates multiple models with different weights to improve overall performance.

## 2 Classification Models

### 2.1 SciBERT Model

Bidirectional Encoder Representations from Transformers (BERT) [4] is a transformers-based model that has significant impact on the field of NLP. It employs multiple layers of transformer block and is pre-trained on masked language modeling (MLM) and next sentence prediction (NSP) tasks, achieving exceptional performance in NLP tasks.

SciBERT, on the other hand, represents a specialized adaptation of BERT, which leverages unsupervised pretraining on a large multi-domain corpus of scientific publications [2]. The pretraining process enables SciBERT to effectively capture the nuances and complexities of scientific texts, exhibiting a profound comprehension of scientific jargon and concepts. Therefore, SciBERT is suitable for PST task and demonstrates superior performance over GLM-4 [9].

To further enhance the performance of SciBERT in the context of PST, two key improvements are introduced: ConcatPool method and AWP method.

*2.1.1 ConcatPool method.* The ConcatPool method is an innovative pooling technique that enhances the performance and robustness of the model by combining features from different layers. As illustrated in Figure 2, it extracts and concatenates hidden states from the last four layers of the model, followed by mean pooling. This approach helps the model to better capture both local and

*Conference KDD '2024, 2024, Barcelona, Spain*

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

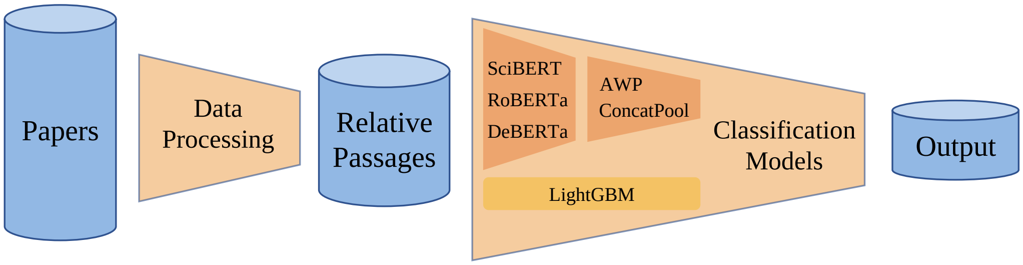

**Figure 1: The Overall Pipeline**

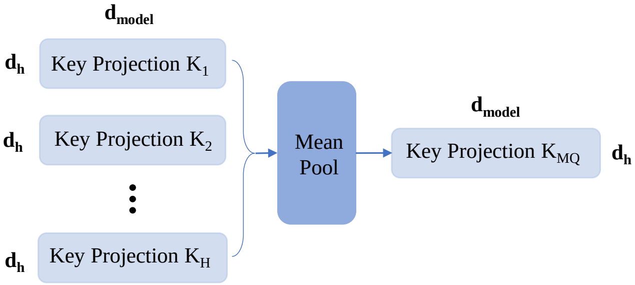

**Figure 2: The Process of Concatenate Mean Pooling**

global features of the text, thereby improving its performance on complex tasks [1].

Therefore, in the PST task, the application of the ConcatPool method allows SciBERT to have a better understanding of the structure and semantics of the text, enabling it to more accurately identify and trace source references within scientific papers.

#### 2.1.2 *AWP method.*

The AWP method represents another technique that improves the robustness of the model by introducing controlled perturbations to the model's weights during training [8]. These perturbations simulate adversarial scenarios and force the model to learn more robust and generalized representations of the input, thereby making it less susceptible to the problem of overfitting.

In the PST task, the weight perturbations make SciBERT able to adapt to input variations, which is crucial for improving the stability of the model when processing complex scientific literature.

By combining the ConcatPool and AWP methods, the performance of SciBERT in the PST task has been significantly enhanced, proving the effectiveness of these two methods. Consequently, we have extended the application of these two techniques to other BERT-based models, including RoBERTa and DeBERTa, to further optimize their performance in the PST task.

### 2.2 RoBERTa Model

RoBERTa [7] is an improved version of BERT. The model replaces the static masking in BERT with dynamic masking, thereby reducing data repetition and enhancing the adaptability of the model. Furthermore, RoBERTa employs complete sentences during pretraining rather than next sentence prediction (NSP) task, which allows the model to more effectively capture text coherence. Given the evidence that task-adaptive pretraining can enhance model performance [5], we have chosen dsp-roberta-base-dapt-cs-tapt-sciie-3219, and ConcatPool and AWP methods are utilized to finetune pretraining models.

### 2.3 DeBERTa Model

DeBERTa [6] introduces two novel techniques to BERT: the disentangled attention mechanism and the enhanced mask decoder. The disentangled attention mechanism meticulously separates the influence of content and positional information, while the enhanced mask decoder reconstructs the original input from masked tokens using these disentangled representations. These advancements significantly enhance the model's ability to capture complex dependencies and contextual nuances. Consequently, the model named deberta-v3-base is the chosen for the PST task, utilizing the ConcatPool and AWP methods.

### 2.4 LightGBM Model

In addition to the pre-trained models above, statistical methods can also be employed to identify source references. The importance of each reference can be indicated by statistical features such as citation counts, citation positions, author overlap, and text similarity. Therefore, we utilize LightGBM, a highly efficient algorithm with a gradient boosting framework, to determine the importance of each reference. LightGBM is capable of handling large datasets and providing fast training times, making it an ideal choice for the PST task.

## 3 Related Works

### 3.1 Dataset Description

A comprehensive understanding of the dataset is paramount in competition. The OAG-PST dataset mainly encompasses papers from the domains of artificial intelligence (AI), database and data mining, and high-performance computing (HPC) [9]. The dataset can be divided into four parts, including the training set, valid set, test set, and the documents of the papers, as detailed in Table 1.

The specific explanations are as follows:

- pid: the unique id of the paper from documents.
- pids: the list of each pid.
- refs_trace: the most important references of the papers.
- full text: full paper, including pid, title, abstract, body, etc.

**Table 1: Dataset Description**

| Type | File name | Fields |
|---|---|---|
| Train dataset | PST_train_ans.json | pids, refs_trace |
| Valid dataset | PST_valid_wo_ans.json | pids, refs_trace |
| Test dataset | PST_test_wo_ans.json | pids, refs_trace |
| Documents | pid.xml | full texts |

## 3.2 Data Processing

The development of LLMs has been significantly influenced by the use of unsupervised pretraining with large-scale, high-quality text datasets. Therefore, it is necessary to preprocess the data and filter out irrelevant content, enhancing the performance of the model.

In XML-formatted text, there are often tags that do not contribute to the understanding of the text, such as "<ref type="bibr" target="#b19">". Consequently, we utilized regular expressions to clean these tags, converting the text into a more concise format.

Additionally, the PST task involves determining whether a reference is a source paper by extracting the context surrounding the citation position. Through testing, it has been found that a context width of 400 characters yields the best results.

## 3.3 Weighted Ensemble Models for PST

In the realm of NLP, the ensemble learning approach is a prevalent strategy employed to enhance overall predictive performance. The combination of different models can produce more robust results and preform better than a single model, due to these model can learn distinct patterns from the data [3]. Moreover, the ensemble approach mitigates the risk of converging to local optima, thereby enhancing the generalization capability of the model [10].

In our works, we have chosen three BERT-based pretraining models, namely scibert-scivocab-uncased, dsp-roberta-base-dapt-cs-tapt-sciie-3219, and deberta-v3-base. After data preprocessing, the clean text is fed into three models along with the corresponding labels. Through a series of transformer blocks and concatenate mean pooling, the last hidden state is passed through a final linear layer to yield the output, which represents the probability distribution of each reference being a source paper. In addition, we also utilized cross-validation to train SciBERT and RoBERTa, which helps to improve the performance of the weighted ensemble model. Meanwhile, LightGBM generates outputs based on the clean text and additional statistical features extracted from the DBLP citation network. Based on the online validation scores, a weighted ensemble method is utilized to integrate the outputs from six different models. The entire process is depicted in Figure 3.

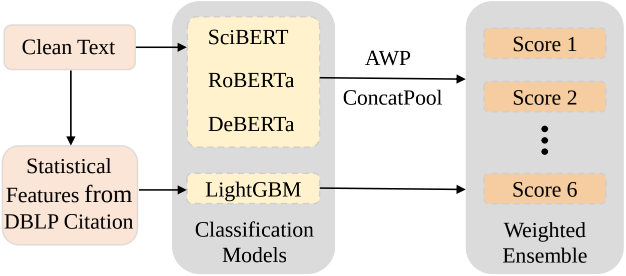

**Figure 3: Weighted Ensemble Method**

## 3.4 Experiments

For each paper $V_q$, the average precision (AP) will be calculated according to the following formula:

$$AP\left(V_q\right) = \frac{1}{R_q}\sum_{k=1}^{M} P_q(k)\mathbf{1}_k$$

where $R_q$ is the number of positive samples, i.e., important references, and $P_q(k)$ represents the precision up to the $k$-th item in the ranked list for paper $V_q$. $\mathbf{1}_k$ is an indicator function, if the $k$-th reference is a important reference, then $\mathbf{1}_k = 1$, otherwise $\mathbf{1}_k = 0$; $M$ denotes the total number of corpus papers.

Given the AP of each paper $V_q$, we can calculate the mean average precision (MAP) as follows:

$$MAP = \frac{1}{n}\sum_{q=1}^{n} AP\left(V_q\right)$$

In model selection, we conduct experiments on the validation leaderboard. Based on the online evaluation metrics, we determine the optimal classification models for our application. Table 2 primarily illustrates the improvements in validation and test scores following the processing by some models.

As shown in Table 2, the initial performance of the baseline SciBERT model indicates a moderate level of accuracy in the PST task, with a validation score of 0.36484 and a test score of 0.32042. The utilization of the ConcatPool and AWP methods leads to a significant increase in performance, with the validation score increasing to 0.39859 and the test score to 0.38373. This improvement demonstrates the effectiveness of these two techniques in enhancing the performance and robustness of the model.

Further improvements by incorporating RoBERTa and DeBERTa models show incremental gains in the validation score, reaching 0.40864 and 0.41527, respectively. However, the test scores for these models decrease slightly, indicating potential overfitting to the validation set or differences in the test set distribution. The addition of LightGBM, a gradient boosting framework, results in a significant jump in both validation and test scores, with values of 0.42970 and 0.41312, respectively. This suggests that LightGBM complements the BERT-based models well, probably by capturing different aspects of the text through its tree-based learning approach.

**Table 2: Score in Validation and Test Leaderboard**

| Model | Validation Score | Test Score |
|---|---|---|
| SciBERT | 0.36484 | 0.32042 |
| + ConcalPool and AWP | 0.39859 | 0.38373 |
| + RoBERTa | 0.40864 | 0.38267 |
| + DeBERTa | 0.41527 | 0.38152 |
| + LightGBM | 0.42970 | 0.41312 |
| + Weighted Ensemble | **0.43255** | **0.41778** |

Finally, the application of a weighted ensemble method achieves the highest scores, with a validation score of 0.43255 and a test score of 0.41778. This result highlights the effectiveness of ensemble techniques in aggregating the strengths of multiple models to achieve superior performance on both validation and unseen test data.

## 4 Conclusion

The PST-KDD-2024 competition presents an opportunity for us to explore automated algorithms for tracing source references. In this competition, we have successfully developed an advanced algorithm for PST by intergrating AWP, ConcatPool, and a weighted ensemble of models including SciBERT, RoBERTa, DeBERTa, and LightGBM. The introduction of AWP has enhanced the robustness and generalization of our models, while ConcatPool has effectively captured both local and global features of the text, leading to more accurate identification of source references. The ensemble approach, which combines outputs from different models, has further improved the overall performance and reliability of our algorithm. The experiments conducted on the OAG-PST dataset have shown that our proposed method achieves a competitive ranking in the competition.

Although our work has achieved certain advancements, the limitations still remain. Notably, we have yet to explored the potential of models with billions of parameters, such as GLM-4, in the context of PST.

Finally, our team has won the 7th place in the final leaderboard, demonstrating the effectiveness of our approach in identifying and tracing source references within scientific papers.

## References

[1] Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Sanghai. 2023. Gqa: Training generalized multi-query transformer models from multi-head checkpoints. *arXiv preprint arXiv:2305.13245* (2023).

[2] Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676* (2019).

[3] Chris Deotte, Jean-Francois Puget, Benedikt Schifferer, Gilberto Titericz, et al. 2023. Winning Amazon KDD Cup'23. In *Amazon KDD Cup 2023 Workshop*.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[5] Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. *arXiv preprint arXiv:2004.10964* (2020).

[6] Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543* (2021).

[7] 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. *arXiv preprint arXiv:1907.11692* (2019).

[8] Dongxian Wu, Shu-Tao Xia, and Yisen Wang. 2020. Adversarial weight perturbation helps robust generalization. *Advances in neural information processing systems* 33 (2020), 2958–2969.

[9] Fanjin Zhang, Kun Cao, Yukuo Cen, Jifan Yu, Da Yin, and Jie Tang. 2024. PST-Bench: Tracing and Benchmarking the Source of Publications. *arXiv preprint arXiv:2402.16009* (2024).

[10] Huimin Zhao, Jinyu Zhu, and Wu Deng. 2024. A new weighted ensemble model-based method for text implication recognition. *Multimedia Tools and Applications* (2024), 1–16.