Proceedings of Machine Learning Research vol 278:1–8, 2025 2025 2nd International Conference on Machine Learning and Intelligent Computing

Small Sample Patents Classification Task Based on Mengzi-BERT-base Single Model

Hao Huang
Quanzhou Industrial Inverstment Development Co..Ltd, Quanzhou, Fujian, China

Yi Liu*
2252070212@QQ.COM
College of Engineering, Huaqiao University, Quanzhou, Fujian, China

Editors: Nianyin Zeng, Ram Bilas Pachori and Dongshu Wang

Abstract

Small sample data classification faces challenges such as data scarcity, overfitting risks, and feature representation learning. In order to tackle these challenges, the present study proposes a transfer learning methodology that leverages the insights gained from extensive datasets or pre-trained models to enhance the model's capacity for generalization. Furthermore, meta-learning methodologies facilitate the rapid adaptation of models to novel tasks using a limited number of samples by employing strategies that enhance the learning process itself. Concurrently, data augmentation techniques enhance both the diversity and volume of samples through the synthesis, expansion, or transformation of small datasets, thereby augmenting the model's generalization capabilities. The paper also presents an active learning method that uses the uncertainty and information gain of the model to automatically select the most valuable samples for labeling to optimize the training effect of the model. It solves the problem of obtaining large-scale annotated data in many practical scenarios, and provides efficient classification and analysis of small amounts of annotated data. Moreover, it serves as the basis of zero-sample learning, which has important knowledge transfer and application value. The paper concludes by showing that the proposed approach outperforms existing methods on a benchmark dataset, demonstrating its effectiveness in addressing the challenges of small sample data classification.

Keywords: NLP; BERT; Mengzi.

1. Introduction

Over the past few years, bolstered by supportive policies and advancements in industrialization and educational attainment, China has experienced a significant increase in the number of patent applications. This has led to an increasing demand for patent retrieval, novelty search, and management. To improve the quality of patent services and meet these demands, it has become crucial to establish multidimensional patent classification systems. Widely utilized classification systems encompass the International Patent Classification (IPC), the Cooperative Patent Classification (CPC), and the European Classification (ECLA). However, using these complex classification systems can be challenging for non-IP professionals. In traditional natural language processing, words are typically represented as discrete symbols, which fail to capture the semantic relationships between words. However, the Word2Vec model transforms words into continuous vector representations by learning their distribution patterns in context. This methodology facilitates the positioning of words with analogous meanings at comparable distances within the vector space. Continuous vector representations offer several benefits, such as the ability to encapsulate semantic relationships, accommodate word compositionality, and facilitate reasoning processes.

The classification of text in small samples presents a significant challenge within the domain of natural language processing (Yan et al., 2018). In numerous practical scenarios, researchers frequently encounter circumstances in which the availability of labeled samples is limited, attributable to challenges associated with data collection or prohibitive costs. In such cases, traditional text classification algorithms often fail to achieve satisfactory performance. Consequently, enhancing the performance and generalization capabilities of models in small-sample text classification tasks has emerged as a significant area of interest for researchers. In recent years, the introduction of pre-trained models, such as BERT and GPT, has led to innovative methodologies for addressing small-sample text classification challenges (Xu and Du, 2020). These models are initially trained on extensive collections of unlabeled data, enabling them to develop comprehensive linguistic knowledge and semantic representations. When subsequently fine-tuned with a limited quantity of labeled data, these models have demonstrated exceptional performance on small-sample datasets.

In conclusion, establishing multidimensional patent classification systems is of great importance for patent classification tasks. This facilitates a more thorough and precise characterization of the content and attributes of patents, thereby enhancing patent retrieval and knowledge management processes. However, due to the diversity and complexity of the patent field, building multidimensional classification systems poses certain challenges. This places higher demands on language classification models, including strong language understanding and classification capabilities, recognition and understanding of domain-specific terminology and language, and efficient processing of large-scale datasets. The ongoing enhancement and optimization of language classification models can significantly improve the performance of multidimensional patent classification tasks, thereby increasing the efficacy of patent retrieval and knowledge management processes.

2. Related Work

2.1. Few-shot Classification

Few-shot text classification seeks to develop a classifier utilizing a minimal amount of labeled data. Conventional text classification techniques, including Support Vector Machines (SVM), Multi-Layer Perceptron (MLP), and Transformers, are not ideally suited for this context, as they necessitate a substantial quantity of labeled samples to effectively train and converge the model (Pang et al., 2021). In recent years, meta-learning methodologies have gained prominence as a framework for few-shot text classification, with the objective of addressing few-shot tasks through the acquisition of a set of generic "meta-tasks." There are two primary strategies for employing meta-learning in the context of the target task: (1) optimization-based methods and (2) metric-based methods. The former approach conceptualizes the meta-task as a broad parameter optimization process, which can be adapted to a variety of similar tasks. Conversely, the latter approach emphasizes the assessment of similarity between query and support samples by generating sample representations that exhibit enhanced clustering characteristics, subsequently utilizing this representation to model the classification probabilities of query samples (Wang et al., 2022).

Numerous research investigations have established the efficacy of meta-learning approaches in the context of few-shot text classification (Pan et al., 2019; Zheng et al., 2021; Huisman et al., 2021). Specifically, metric-based meta-learning has the potential to outperform text classification models that are trained from the ground up using traditional methods. Koch et al. (2015) introduced Siamese neural networks, a model comprising two identical neural networks, to tackle the challenge of one-shot learning. This approach enables the differentiation of whether query and support

samples originate from the same category. Drawing inspiration from the Siamese network architecture, Vinyals et al. (2016) developed Matching Networks, which leverage global information within each episode to characterize the query sample. They implement an attention mechanism to assign weights to the aggregated label information of support samples, thereby simulating the classification probability of the query sample. In a similar vein, Snell et al. (2017) introduced Prototypical Networks, a straightforward and effective probabilistic learning algorithm that constructs class prototypes as cyclic representations of the respective classes. This approach assesses the classification probability of query samples by averaging their similarity to each prototype representation derived from support samples within the same category. Furthermore, Sung et al. (2017) proposed Relation Networks, which employ deep neural networks to model the similarity metric function, thereby superseding traditional fixed-distance calculation methods such as L1 distance, L2 distance, and cosine similarity.

While metric-based meta-learning establishes a context of sparsely labeled samples through the clustering of features and text similarity for each meta-task, it remains necessary to filter and preserve a substantial quantity of representative labeled data for the purposes of training and debugging.

2.2. Prompt Learning

Prompt learning has emerged as a prominent methodology in the domain of text classification, leveraging the capabilities of pretrained language models (Zhang et al., 2021). This technique involves the fine-tuning of the model through the utilization of either manually constructed or automatically generated prompt sentences, with the incorporation of mask tokens into the original text. The primary aim is to employ the pretrained model to predict the labels corresponding to the masked positions, which are pertinent to the labels associated with the text classification task. Prompt learning can be categorized into two principal types based on the design of the prompts: discrete prompts and continuous prompts. Discrete prompt learning employs a collection of natural language prompt tokens for the classification of text, exemplified by the PET framework (Devlin et al., 2019). For instance, when presented with an input text, the input embedding sequence can be structured to include both the input text and prompt tokens, such as "This is a [MASK] topic." Discrete prompting entails a local search mechanism, as neural networks operate within a continuous framework while the search occurs in a discrete domain. Conversely, continuous prompt learning posits that prompt templates can be subject to training, thereby optimizing continuous prompt embeddings. This methodology involves training the input sequence, which consists of a series of trainable, continuous embeddings, on the input text, functioning as placeholders for the mask tokens predicted by the model. For instance, the EFL method (Sun et al., 2019) employs the T5 model to generate optimal discrete prompt templates, thereby obviating the necessity for manual search. Another methodology, known as P-tuning (Brown et al., 2020; Liu et al., 2023), treats prompt templates as trainable entities and optimizes continuous prompt embeddings, achieving performance levels that are comparable to fine-tuning with BERT in supervised learning contexts.

3. Methods

In this study, we present a methodology that leverages the pre-trained BERT model in conjunction with the AdamW optimizer to tackle the challenge of small-sample text classification, as illustrated in Figure 1. The proposed approach encompasses several key components, including data preprocessing, the utilization of the pre-trained BERT model, the application of the AdamW optimizer,

fine-tuning and training procedures, parameter configuration, and bias correction measures. Initially, we conduct preprocessing of the raw text data, which involves tokenization, the removal of stop words, and the transformation of text into word vector representations. This preprocessing step is essential for converting the text data into a format that is compatible with machine learning algorithms, thereby facilitating subsequent model training. Subsequently, we employ the pre-trained BERT model as the foundational model for our classification task. The BERT model, which has been pre-trained on a substantial unlabeled dataset, possesses extensive language knowledge and semantic representations. In the context of small-sample text classification, we fine-tune this pre-trained BERT model to adapt it specifically to the classification task at hand.

The pretrained BERT model, designated as Mengzi-BERT-base, is employed for the task of text classification. The architecture of the model is depicted in Figure 2. BERT Tokenizer is used for text tokenization and encoding to input the text into the BERT model. Based on proprietary techniques involving linguistic knowledge, knowledge graphs, and domain data augmentation, comprehensive improvements have been made to the model architecture (including the base embedding representation and the interaction layer attention mechanism) and pre-training strategies. In particular, with regard to the model architecture, linguistic attributes, including semantic roles and part-of-speech tagging, are incorporated into the embedding representation. Additionally, attention mechanisms are implemented in accordance with syntactic constraints to improve the model's capacity to represent linguistic knowledge effectively. In terms of training strategies, a Mask mechanism based on entity knowledge and discourse is introduced to strengthen the representation of linguistic components and discourse relations.

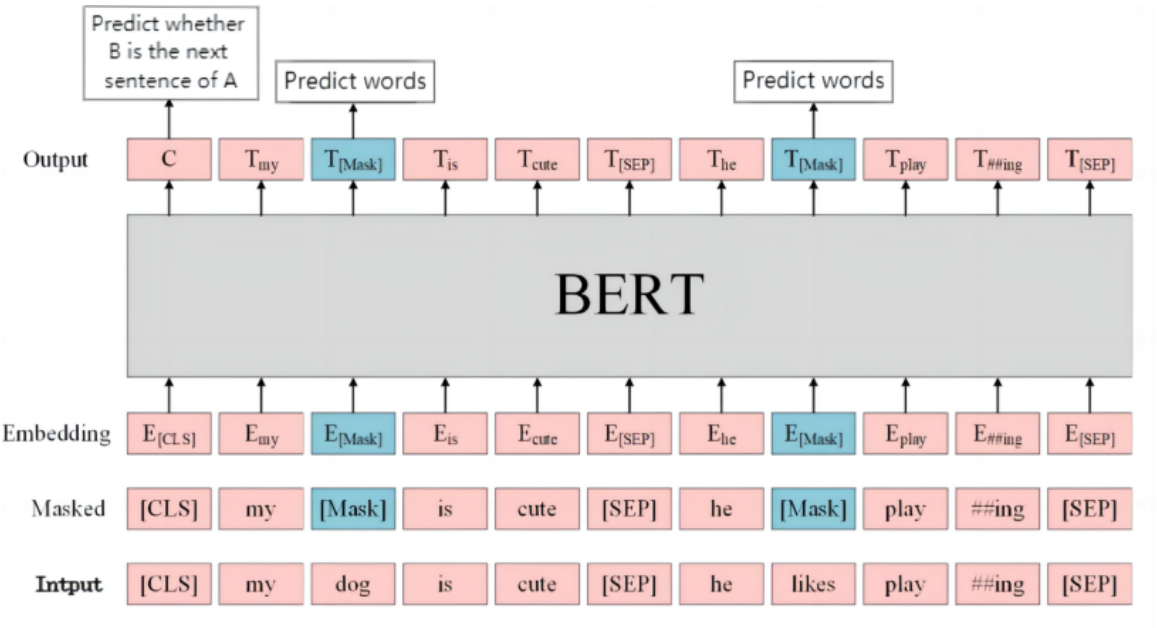

Figure 1: The pre-trained BERT model, in conjunction with the AdamW optimizer, is employed to tackle the challenge of text classification in scenarios characterized by limited sample sizes.

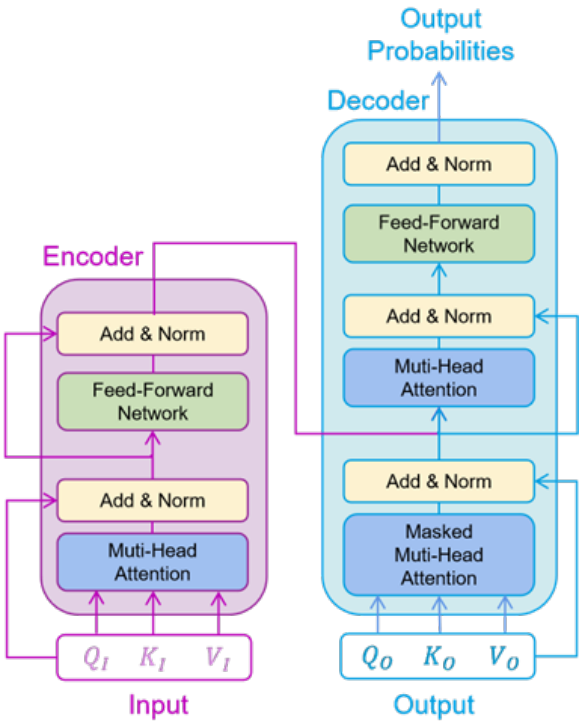

Figure 2: Pretrained BERT model is used for the text classification task, with the model named Mengzi-BERT-base.

To enhance training efficiency, the methodology of large model distillation alongside the initialization of a smaller model is utilized. To tailor the Mengzi model for specialized sectors such as finance and marketing, domain-specific datasets are employed for further training, accompanied by the development of relevant prompt templates (Schick and Schütze, 2021). This strategy has resulted in notable enhancements in performance. The BERT model is utilized for sequence classification tasks through the BERT for sequence classification class, which encompasses 36 labeled categories. The architecture of the BERT model is depicted in Figure 3.

4. Results

The results of the experiments are presented in Table 1. It is evident that the proposed Induction Networks demonstrate superior classification performance across all four experimental conditions. In contrast, the distance metric learning models, including Matching Networks, Prototypical Networks, Graph Networks, and Relation Networks, primarily focus on feature representation and distance measurement at the sample-wise level. Our research introduces an induction module that emphasizes class-wise representation, which we argue is more resilient to variations in the samples within the support set. Furthermore, our model surpasses the most recent optimization-based method, SNAIL. The performance disparity between Induction Networks and SNAIL, as indicated in the table, is statistically significant at the 99% confidence level when assessed using a paired

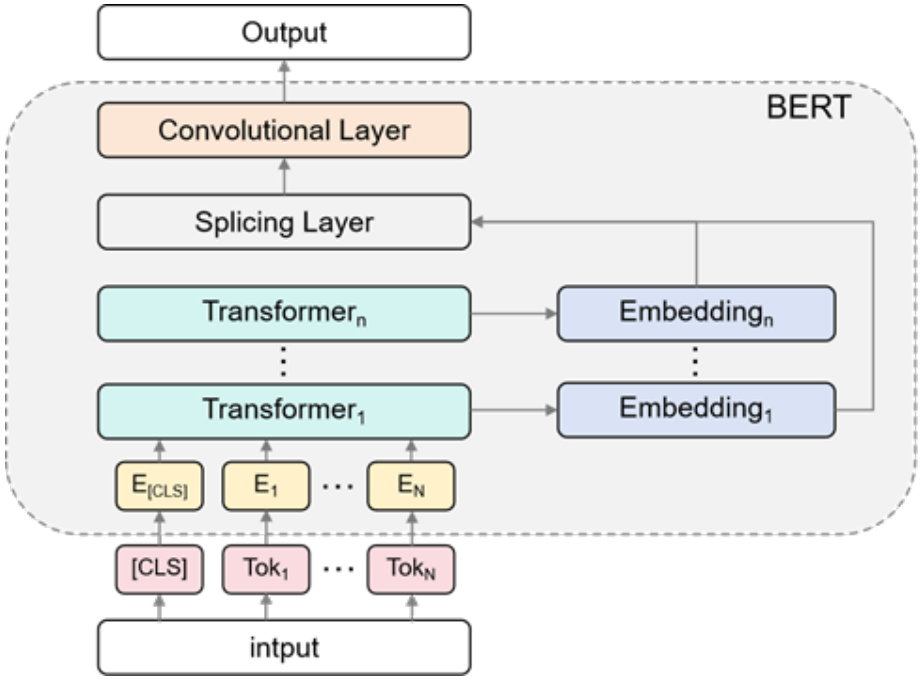

Figure 3: The structure of the BERT model.

comparison. Additionally, the performance gap between our model and other baseline models in the 10-shot scenario is more pronounced than in the 5-shot scenario. This observation can be attributed to the fact that, in the 10-shot scenario, the advantages of increased data size for the baseline models are counterbalanced by a greater presence of sample-level noise.

Table 1: The study encompasses an analysis of temperature variations and wildlife populations across the three designated areas

Model	5-way Acc.		10-way Acc.	
	5-shot	10-shot	5-shot	10-shot
Matching Networks	82.54 $\pm$ 0.12	84.63 $\pm$ 0.08	73.64 $\pm$ 0.15	76.72 $\pm$ 0.07
Prototypical Networks	81.82 $\pm$ 0.08	85.83 $\pm$ 0.06	73.31 $\pm$ 0.14	75.97 $\pm$ 0.11
Graph Network	84.15 $\pm$ 0.16	87.24 $\pm$ 0.09	75.58 $\pm$ 0.12	78.27 $\pm$ 0.10
Relation Network	84.41 $\pm$ 0.14	86.93 $\pm$ 0.15	75.28$\pm$ 0.13	78.61$\pm$ 0.06
SNAIL	84.62 $\pm$ 0.16	87.31 $\pm$ 0.11	75.74 $\pm$ 0.07	79.26 $\pm$ 0.09
ours	87.16 $\pm$ 0.09	88.49 $\pm$ 0.17	78.27 $\pm$ 0.14	81.64 $\pm$ 0.08

In the context of small-sample text classification, we undertook empirical research on the methodologies put forth and documented the actual performance outcomes. The subsequent section presents our analysis and discussion of the experimental findings.

Initially, we examined the influence of data preprocessing on the outcomes of the experiments. During the preprocessing phase, we employed standard text processing methodologies, including tokenization and the removal of stop words. These preprocessing procedures serve to cleanse and

convert the raw text data, resulting in more precise and comprehensive feature representations. Consequently, effective data preprocessing can enhance the model's comprehension of the textual content, thereby augmenting classification performance.

The pre-trained BERT model exerts a considerable influence on the outcomes of the experiments conducted. This model is initially trained on extensive unlabeled datasets, enabling it to acquire rich semantic representations. During the fine-tuning phase, the BERT model is further trained using a limited amount of labeled data, allowing it to be tailored for specific classification tasks. The results of the experiments indicate that the pre-trained BERT model serves as an effective foundational model for text classification tasks involving small sample sizes, leading to a notable enhancement in classification performance. By employing the AdamW optimizer to update the parameters of the BERT model, we can achieve more effective optimization and enhance the model's performance.

5. Conclusions

This paper examines methodologies for tackling the challenge of small-sample text classification. It commences with an introduction to the implementation of pre-trained BERT models alongside the AdamW optimizer for this specific task. The paper meticulously outlines the procedures involved in updating the parameters of the BERT model utilizing the AdamW optimizer within the coding framework. Additionally, it elucidates the concepts of learning rate scheduling and bias correction throughout the training phase. Moreover, the paper explores the roles of self-supervised learning and transfer learning in small-sample text classification endeavors. By leveraging pre-training on unlabeled datasets and subsequently fine-tuning on a limited set of labeled data, the paper illustrates how the extensive use of unlabeled data and the generalizable features acquired from pre-trained models can enhance model performance.

In conclusion, this study introduces a methodology for addressing small-sample text classification challenges through the utilization of pre-trained BERT models in conjunction with the AdamW optimizer, as well as various enhancement strategies and empirical validation. The primary contribution of this research is the provision of targeted solutions for small-sample text classification tasks, while also investigating techniques such as the application of pre-trained models, data augmentation, and model fine-tuning to enhance overall model efficacy. This work holds considerable significance for the advancement of solutions to small-sample text classification issues within the field of natural language processing.

References

Tom B. Brown, Benjamin Mann, Nick Ryder, and et al. Language models are few-shot learners. *CoRR*, abs/2005.14165, 2020.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. pages 4171–4186, June 2019. doi: 10.18653/v1/N19-1423.

Mike Huisman, Jan N. van Rijn, and Aske Plaat. A survey of deep meta-learning. *Artificial Intelligence Review*, 54(6):4483–4541, 2021.

Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov. Siamese neural networks for one-shot image recognition. 2015.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, and et al. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.*, 55(9), January 2023. doi: 10.1145/3560815.

Chongyu Pan, Jian Huang, Jianxing Gong, and Xingsheng Yuan. Few-shot transfer learning for text classification with lightweight word embedding based models. *IEEE Access*, 7:53296–53304, 2019. doi: 10.1109/ACCESS.2019.2911850.

Ning Pang, Xiang Zhao, Wei Wang, and et al. Few-shot text classification by leveraging bi-directional attention and cross-class knowledge. *Science China Information Sciences*, 64(3): 130103, 2021. doi: 10.1007/s11432-020-3055-1.

Timo Schick and Hinrich Schütze. It's not just size that matters: Small language models are also few-shot learners. pages 2339–2352, June 2021. doi: 10.18653/v1/2021.naacl-main.185.

Jake Snell, Kevin Swersky, and Richard S. Zemel. Prototypical networks for few-shot learning. *CoRR*, abs/1703.05175, 2017.

Yu Sun, Shuohuan Wang, Yu-Kun Li, and et al. ERNIE: enhanced representation through knowledge integration. *CoRR*, abs/1904.09223, 2019.

Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip H. S. Torr, and Timothy M. Hospedales. Learning to compare: Relation network for few-shot learning. *CoRR*, abs/1711.06025, 2017.

Oriol Vinyals, Charles Blundell, Timothy P. Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. *CoRR*, abs/1606.04080, 2016.

Depei Wang, Zhuowei Wang, Lianglun Cheng, and Weiwen Zhang. Few-shot text classification with global–local feature information. *Sensors*, 22(12), 2022. doi: 10.3390/s22124420.

Jincheng Xu and Qingfeng Du. Learning transferable features in meta-learning for few-shot text classification. *Pattern Recognition Letters*, 135:271–278, 2020. doi: https://doi.org/10.1016/j.patrec.2020.05.007.

Leiming Yan, Yuhui Zheng, and Jie Cao. Few-shot learning for short text classification. *Multimedia Tools and Applications*, 77(22):29799–29810, 2018. doi: 10.1007/s11042-018-5772-4.

Ningyu Zhang, Luoqiu Li, Xiang Chen, and et al. Differentiable prompt makes pre-trained language models better few-shot learners. *CoRR*, abs/2108.13161, 2021.

Jianming Zheng, Fei Cai, Wanyu Chen, and et al. Taxonomy-aware learning for few-shot event detection. page 3546–3557, 2021. doi: 10.1145/3442381.3449949.