An Exploration of Transferring Domain Expertise

Stanford CS224N Custom Project

Jonathan Paul Hsu
Department of Computer Science
Stanford University
jphsu@stanford.edu

Abstract

This study investigates the ability of BERT, a pre-trained natural language processing (NLP) model, to transfer domain expertise through fine-tuning. We examine whether combining datasets from different sub-fields during fine-tuning improves performance in a specific domain compared to fine-tuning on a single dataset. We also examine whether fine-tuning on a domain different from the domain we test on yields positive results. Our contributions include a detailed analysis of the effects of dataset combination on model performance and an evaluation of cross-domain training and testing. Contrary to our expectations, our findings indicate that combining datasets results in a slightly worse performance than using a single dataset. However, training on a sentiment analysis dataset from a different domain than the test set still yielded significant improvements over baseline, demonstrating BERT's strong capability for domain knowledge transfer.

1 Key Information to include

- Mentor: Zhoujie (Jason) Ding

2 Introduction

The advent of pre-trained language models like BERT has brought significant advancements in natural language processing (NLP), enabling improvements across various applications such as sentiment analysis, text classification, and more. Fine-tuning these models on domain-specific datasets is a widely adopted practice to tailor their performance for specific tasks. However, the potential of leveraging multiple datasets from different sub-fields for fine-tuning, to enhance performance in a specific domain, remains underexplored.

Given that training language models is computationally expensive and not accessible to everyone, fine-tuning is the primary avenue many can turn to when utilizing language models for their own tasks. Thus, it is important for us to deepen our understanding of the fine-tuning process to democratize participation in this cutting edge technology.

This study aims to investigate two key aspects of domain expertise transfer in BERT: the effect of combining datasets from different sub-fields during fine-tuning, and the effectiveness of cross-domain training and testing. We will test domain expertise transfer through utilizing BERT to conduct sentiment analysis on different testing and training datasets. We hypothesize that combining datasets will provide a richer context, leading to improved model performance in a specific domain. Additionally, we explore whether fine-tuning BERT on a dataset from a different domain than the test set can still yield positive results, thus demonstrating the model's ability to transfer domain knowledge.

Our contributions are threefold:

1. We provide a detailed analysis of the effects of combining datasets on BERT's performance.
2. We evaluate the performance of BERT when fine-tuned on a domain different from the domain used for testing.
3. We highlight the challenges and potential pitfalls associated with multi-domain fine-tuning.

Contrary to our initial hypothesis, our findings reveal that combining datasets results in slightly worse performance compared to fine-tuning on a single dataset. This unexpected outcome underscores the complexities involved in multi-domain training. It also highlights potential problems during fine-tuning over over-fitting to the training dataset. However, our results also show that training on a sentiment analysis dataset from a different domain than the test set significantly improves performance over baseline, illustrating BERT's robust capability for domain knowledge transfer.

By examining these facets, this study contributes to a deeper understanding of how pre-trained language models like BERT can be fine-tuned for optimal performance across various domains, providing insights that could inform future research and practical applications in NLP.

3 Related Work

The development and success of BERT (Bidirectional Encoder Representations from Transformers) by Devlin et al. (2018) has marked a significant advancement in the field of natural language processing (NLP). BERT's ability to capture bidirectional context through its transformer architecture has set a new standard for NLP tasks, including sentiment analysis, text classification, and question answering.

Fine-tuning pre-trained models on domain-specific datasets has been widely adopted to enhance model performance on specific tasks. Sun et al. (2020) demonstrated that fine-tuning BERT on task-specific data significantly improves its performance across various NLP tasks. Their work highlights the importance of adapting pre-trained models to the nuances of domain-specific data.

The concept of leveraging multiple datasets from different domains for fine-tuning has also been explored in previous research. Liu et al. (2019) investigated the use of multi-domain training to improve model robustness and generalization. Their findings suggest that while multi-domain training can enhance generalization, it may also introduce complexities such as domain interference and overfitting.

In the realm of sentiment analysis, datasets like the Yelp datset and IMDb movie reviews dataset have been extensively used. The Yelp dataset, as described by Zhang et al. (2016), provides a rich source of customer reviews for sentiment classification tasks. Similarly, the IMDb dataset offers a large collection of movie reviews with sentiment labels, making it a valuable resource for training and evaluating sentiment analysis models.

The challenges of cross-domain learning and domain adaptation have been addressed by various studies. Pan and Yang (2010) shows a comprehensive survey on transfer learning, emphasizing the importance of domain adaptation techniques to improve model performance across different domains. Their work underscores the potential benefits and challenges of applying transfer learning in NLP tasks. Their work also serves as inspiration for the hypothesis we put forth in this research.

Our study builds on these foundational works by investigating the effects of combining datasets from different sub-domains during fine-tuning and evaluating BERT's cross-domain generalization capabilities. By exploring these aspects, we aim to contribute to the understanding of how pre-trained language models can be optimized for performance across various domains and how we can think about tailoring training datasets to generalized or specific tasks.

4 Approach

In this study, we utilize the BERT Model transformer with a sequence classification/regression head on top, a pre-trained transformer architecture known for its bidirectional context capturing ability, which makes it well-suited for fine-tuning on domain-specific datasets to transfer domain knowledge. We implement the BERT base model through the transformers library from huggingface. To investigate the impact of dataset combination and cross-domain training, we fine-tuned BERT under three different configurations: exclusively on the Yelp dataset containing customer reviews,

exclusively on the IMDb dataset consisting of movie reviews, and on a merged dataset of Yelp and IMDb reviews. We then test the three fine-tuned models as well as a base model without fine-tuning against test datasets from Yelp, IMDb, and the merged dataset.

We pre-process the Yelp dataset to transform its labels from a scale of 0-4 to a scale of 0-1 to ensure compatibility with the IMDb dataset. The preprocessing steps are crucial to ensure compatibility with BERT's input requirements. We use BERT's tokenizer to split text into tokens. The fine-tuning process involves training BERT on the respective datasets with a learning rate of 2e-5, a batch size of 8, and training over three epochs. We utilize the AdamW optimizer and a linear learning rate scheduler.

To evaluate BERT's ability to transfer domain knowledge, we conducted experiments where the model trained on one dataset (e.g., IMDb) is tested on another (e.g., Yelp). This setup assesses BERT's generalization capability across different domains. Models are evaluated using accuracy. As a baseline, we use the BERT Base model with a sequence classification/regression head on top without finetuning to evaluate the improvements achieved through fine-tuning BERT.

The experiments are conducted on Google Cloud Platform with a machine equipped with an NVIDIA T4 GPU, ensuring adequate computational power for training.

5 Experiments

5.1 Data

We utilize three datasets.

1. **Yelp:** The yelp_review_full dataset from Zhang et al. (2016) consists of reviews from Yelp and was extracted from the Yelp Dataset Challenge 2015 data. This dataset is on huggingface and contains 650k rows of data. We have modified the scale of the labels to be 0-1 from the original scale of 0-4, with 0 being a negative sentiment and 1 being a positive sentiment.
2. **IMDb:** The imdb dataset from Maas et al. (2011) consists of a large collection of movie reviews, with ratings similarly converted into binary sentiment labels. The dataset has 25k rows of data with a 0-1 scale for labels.
3. **Combined:** The combined dataset is initially constructed with all 25k data points from the imdb dataset as well as a randomly sampled selection of 25k rows from the yelp dataset to ensure equal weight from both domains.

5.2 Evaluation method

Model performance is evaluated using standard classification metrics, with a primary focus on accuracy. The evaluation involves running the model in evaluation mode on the test datasets, where predictions are generated and compared to the true labels. Accuracy is computed by measuring the ratio of correctly predicted instances to the total instances. This metric provides a straightforward and reliable measure of the model's effectiveness in predicting sentiment label.

5.3 Experimental details

We implement BERT with BertModelForSequenceClassification based on the bert-base-cased model. We then fine-tune on the respective datasets using the following configuration: a learning rate of 5e-5, a batch size of 8, and training over three epochs. We also utilize the AdamW optimizer and a linear learning rate scheduler.

In constructing the final datasets we use for evaluation and training, we take random samples of 1,000 from each dataset we are using to reduce the time of both evaluation and training.

5.4 Results

The results of the experiments are summarized in the table above. As mentioned, we utilize four models in this research: one baseline model without finetuning, one finetuned on yelp, one finetuned

on imdb, and one finetuned on both. Each model's accuracy is evaluated on the Yelp, IMDb, and combined test sets, providing insights into their performance across different domains.

Table 1: Model Performance on Different Test Sets

Model	Yelp Test Set Accuracy	IMDb Test Set Accuracy	Combined Test Set Accuracy
Yelp Fine-tuned	86.8%	80.7%	84.6%
IMDb Fine-tuned	83.0%	86.6%	85.3%
Combined Fine-tuned	86.4%	86.6%	88.7%
Baseline Model	41.1%	51.2%	46.4%

As we can see from the table, the highest performing models for each test set are the models fine-tuned on solely the original dataset. This is quite a surprising result as we initially hypothesized that the fine-tuning process might benefit from knowledge outside of the original domain and enable the model to classify sentiments.

We notice that on a given test set, the best performing model is always the one fine-tuned on itself, then the combined model, and the model fine-tuned on the other dataset performs worst. We also notice that the accuracies don't deviate much from each other compared to the baseline model. This signals that despite being trained on a dataset from a different domain, the model still learns enough to perform well in sentiment analysis on another domain.

The Yelp fine-tuned model achieved an accuracy of 86.8% on the Yelp test set, 80.7% on the IMDb test set, and 84.6% on the combined test set. This model shows strong performance on its domain-specific dataset (Yelp) but lower performance on the IMDb dataset, indicating some limitations in cross-domain generalization.

The IMDb fine-tuned model achieved an accuracy of 83.0% on the Yelp test set, 86.6% on the IMDb test set, and 85.3% on the combined test set. Similar to the Yelp fine-tuned model, it performs best on its specific domain (IMDb) and shows slightly reduced performance on the Yelp dataset. Interesting to note, the IMDb trained model outperforms the Yelp trained model on testing against the other dataset as well as in testing against the combined dataset. This indicates that the IMDb dataset could contain a broader domain of learning compared to Yelp and thus result in a better general performance.

The combined fine-tuned model, which was trained on both Yelp and IMDb datasets, achieved an accuracy of 86.4% on the Yelp test set, 86.6% on the IMDb test set, and 88.7% on the combined test set. This model demonstrates robust performance across both domains and achieves the highest accuracy on the combined test set, suggesting that training on multiple datasets can enhance generalization.

The baseline model, which was not fine-tuned on the specific datasets, performed significantly worse, with an accuracy of 41.1% on the Yelp test set, 51.2% on the IMDb test set, and 46.4% on the combined test set. These results underscore the importance of fine-tuning for improving model performance on specific tasks.

Overall, our results indicate that our initial hypothesis is incorrect that a combined training dataset would produce more superior single domain performance. However, our data introduces a very interesting thought that our model performs well in cross-domain learning and doesn't require a single instance of a particular domain in training to still perform well in testing.

6 Analysis

The results of our experiments reveal several insightful aspects regarding the performance and generalization capabilities of fine-tuned BERT models across different domains.

Cross-Domain Generalization One of the key observations from our study is the robust cross-domain generalization exhibited by the fine-tuned models. Despite being trained on datasets from different domains, the models still achieved relatively high accuracies on test sets from the other domain. For instance, the Yelp fine-tuned model achieved 80.7% accuracy on the IMDb test set,

and the IMDb fine-tuned model achieved 83.0% accuracy on the Yelp test set. This suggests that BERT's pre-trained knowledge, when combined with domain-specific fine-tuning, enables the model to generalize effectively across different types of sentiment data.

Combined Dataset Performance Contrary to our initial hypothesis, the combined fine-tuned model did not outperform the domain-specific models on their respective test sets. However, it showed competitive performance across all test sets, achieving the highest accuracy on the combined test set (88.7%). This indicates that while the combined dataset does not provide superior performance in individual domains, it enhances the model's ability to handle mixed-domain inputs effectively. This finding suggests a trade-off between domain-specific optimization and broad generalization capabilities. This also brings forth potential concerns about overfitting to the training dataset when utilizing only knowledge from a singular domain.

Limitations in Domain-Specific Optimization The results demonstrate that models fine-tuned on their respective original datasets perform best on their own domain-specific test sets. The Yelp fine-tuned model and the IMDb fine-tuned model achieved the highest accuracies on their respective test sets (86.8% and 86.6%, respectively). This highlights a limitation in leveraging cross-domain data for domain-specific optimization. The models appear to benefit more from intensive exposure to domain-specific data rather than a mix of different domains for achieving peak performance in a single domain. This could prove to be a valuable insight for dataset makeup considerations when fine-tuning other models as well.

Baseline Comparison The baseline model, which was not fine-tuned on any specific dataset, performed significantly worse compared to the fine-tuned models. This stark contrast underscores the critical importance of fine-tuning pre-trained models on domain-specific data to enhance performance. The baseline accuracies (41.1% on Yelp, 51.2% on IMDb, and 46.4% on the combined test set) serve as a clear indicator of the substantial performance gains achievable through targeted fine-tuning even without specific domain knowledge.

Overall, our analysis confirms that while domain-specific fine-tuning remains crucial for achieving peak performance in individual domains, there is significant potential for cross-domain learning. BERT's ability to generalize across different domains, as demonstrated in our experiments, underscores its versatility and robustness as a pre-trained language model.

7 Conclusion

In this study, we explored the ability of the BERT base model to transfer domain expertise through fine-tuning on various datasets, specifically focusing on Yelp and IMDb reviews. Our main findings indicate that while models fine-tuned on individual datasets perform best on their respective test sets, the combined fine-tuned model demonstrated competitive performance across all test sets. Surprisingly, combining datasets did not result in superior single-domain performance, contrary to our initial hypothesis. However, the combined model achieved the highest accuracy on the combined test set, highlighting its robust generalization capabilities.

Our analysis revealed that despite being trained on datasets from different domains, the models exhibited strong cross-domain generalization. This indicates that BERT's pre-trained knowledge, when fine-tuned with domain-specific data, enables effective sentiment analysis across various domains not constrained to the specific training domain. Additionally, the baseline model's significantly lower performance underscores the importance of general fine-tuning regardless of domain for enhancing model accuracy.

The primary limitations of our work include the potential for improved techniques in combining datasets to retain domain-specific benefits while enhancing generalization. Future research could explore different proportions of domain-specific data in the combined dataset and investigate more sophisticated methods for multi-domain training.

Overall, our study demonstrates BERT's versatility and robustness as a pre-trained language model, capable of adapting to different domains through fine-tuning, while also highlighting the complexities involved in transferring domain expertise.

Implications for Future Research The insights gained from our analysis open up several avenues for future research. One potential direction is to explore more sophisticated techniques for combining

datasets that might retain domain-specific benefits while enhancing generalization. Additionally, investigating the impact of different proportions of domain-specific data in the combined dataset could provide deeper insights into optimizing cross-domain learning.

8 Ethics Statement

This project involves the use of publicly available and anonymized datasets (Yelp and IMDb reviews), which mitigates privacy concerns associated with personal data. Ethical challenges related to our research relate to the broader use cases of sentiment analysis and model predictions.

One potential challenge is bias from one domain gets transferred into another originally bias-less domain through our techniques of cross-domain learning. This challenge relates to 'contaminating' an originally fair domain with bias from transferring learning from another domain.

This challenge can be mitigated by specifically looking at all domain-specific datasets used and ensuring that all specific datasets are bias free. However, there could be underlying biases that arise only when in the context of learning from multiple domains. The mitigation strategy in this case would be to implementing techniques to detect and mitigate biases in the training data and model predictions. This involves potentially re-balancing datasets, using fairness-aware algorithms, and conducting thorough bias audits on both models created by the individual datasets and combined datasets.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

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 L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

Sinno Jialin Pan and Qiang Yang. 2010. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2020. How to fine-tune bert for text classification?

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2016. Character-level convolutional networks for text classification.