

Exploring Geothermal Energy with Large Language Models

Kamran Haddadian¹, Roman J. Shor², Shengnan Chen¹

¹Department of Chemical and Petroleum Engineering, University of Calgary

²Harold Vance Department of Petroleum Engineering, Texas A&M University

Summary

While classical machine learning methods have found extensive application in engineering disciplines, the adoption of NLP remains relatively limited within the field. This paper addresses this gap by proposing a methodology that leverages advanced NLP techniques, specifically focusing on the fine-tuning of large language models (LLMs), such as BERT (Bidirectional Encoder Representations from Transformers), for classifying and analyzing scientific papers in the domain of geothermal science. The methodology involves collecting a corpus of scientific data from reputable platforms, fine-tuning the BERT model for paper classification based on abstracts, and further analysis using other LLMs like PHI-2 or llama for question answering tasks. Preliminary results demonstrate the effectiveness of the fine-tuned BERT model in accurately classifying abstracts related to geothermal science.

Theory

In recent years, the emergence of Artificial Intelligence (AI) has revolutionized various fields, including energy engineering. AI technologies, such as machine learning and deep learning, have enabled significant advancements in process optimization, predictive modeling, and decision-making within the energy industry[1]. While classical machine learning has found widespread application across engineering discipline, there has been limited application of leveraging Natural Language Processing (NLP) as a burgeoning technology within the field. The adoption of NLP presents new opportunities for enhancing data analysis, knowledge extraction, and decision-making processes in energy engineering [2]. Thousands of articles are published each year in this field, each reviewed by top experts in the field. However, many important insights within these articles may eventually be overlooked. Natural Language Processing (NLP), with its capacity to comprehend and interpret human language, offers the potential to learn these insights and automate tasks such as literature review, information extraction from textual data, and question answering.

One important feature of large language models (LLMs) is their adaptability through fine-tuning. Fine-tuning a pre-existing language model involves adjusting its parameters to specialize in a specific task or domain. This process is highly advantageous as it requires significantly less time and resources compared to training a large language model from scratch. By leveraging a pre-trained model's existing knowledge and structure, fine-tuning allows for rapid adaptation to new tasks or domains, resulting in improved performance with minimal additional training data.

Fine-tuning language models can be challenging due to the significant amount of calculation required. These models often have billions of parameters, leading to memory-intensive operations during fine-tuning. However, techniques such as Low-Rank Adaptation (LORA) [3] and Quantized Low-Rank adaptation (QLORA) [4] have emerged as effective strategies to address this issue. LORA method involves fixing the weights of the pre-trained model and integrating trainable rank decomposition matrices into each layer of the Transformer architecture. This strategy

substantially decreases the number of trainable parameters for subsequent tasks, all while preserving performance. QLORA takes this a step further by incorporating quantization, which reduces the precision of weights and activations in the model, further decreasing memory requirements without significantly compromising performance. By employing these techniques, practitioners can train large language models more efficiently, making them more accessible and practical for various natural language processing tasks.

As of today, no large language model has been specifically trained on energy-related literature to address sector-specific questions. State-of-the-art (SOTA) language models, such as GPT-3.5 or GPT-4, with hundreds of billions of parameters, are capable of answering general questions as they have been primarily trained on general knowledge. However, when it comes to answering specific domain-related questions, fine-tuned large language models with significantly fewer parameters can outperform SOTA language models.

The initial phase of fine-tuning a large language model for a specific domain involves accessing high-quality data, often sourced from books and papers. However, with millions of words published daily, manually evaluating each document for relevance to the intended domain becomes impractical. Here, the role of natural language processing (NLP) becomes crucial in efficiently annotating these documents based on their content. The BERT language model has demonstrated exceptional proficiency in content annotation [5]. In this study, we aim to leverage the capabilities of the BERT language model as the first step in classifying text and selecting relevant papers for subsequent training of an expert large language model in the domain of geothermal energy.

Method

Scientific data in the field of geothermal energy is sourced from reputable platforms like ScienceDirect, while general data is typically obtained from sources such as Wikipedia. The articles from these platforms are accessed using the Elsevier and Wikipedia APIs. BERT (Bidirectional Encoder Representations from Transformers) is an advanced pre-trained natural language processing model developed by Google, leveraging a bidirectional transformer architecture to produce contextualized word embeddings [6]. As a pioneering and efficient language model, BERT has demonstrated robust performance in classifying text containing fewer than 512 tokens [5] (where each word or a part of the word can be taken as a token). Consequently, BERT has been fine-tuned for the classification of papers based on their abstracts to determine their relevance to geothermal science.

To fine-tune BERT, approximately 9400 paper abstracts were collected. This corpus includes nearly 4400 papers directly related to geothermal energy, 2,500 papers covering topics in energy and chemical engineering but not directly related to geothermal science, and approximately 2,500 papers on miscellaneous scientific topics unrelated to geothermal energy, and chemical engineering. The DOIs of relevant papers and books, as classified by BERT, were utilized as input into the code line to retrieve the full texts in XML format. Preprocessing of these XML documents was necessary before further processing.

Bidirectional Long Short-Term Memory (Bi-LSTM) networks were trained on the same dataset, chosen for their efficacy in text classification tasks[7]. Two distinct tokenization approaches were employed: initially, the Bert tokenizer was used, followed by a tokenizer trained specifically on the text data. Training was conducted using the TensorFlow library in Python. Table 1 delineates the architecture of the bidirectional neural network, comprising an embedding layer, two layers of Bi-LSTM, and two Dense layers with Relu and Sigmoid activation functions for the classification task.

Table 1 Bi-LSTM structure for text classification.

Layer (type)	Output Shape	Parameter number
Embedding(Embedding)	1024	First approach: 56865792 Second approach: 31255552
Bidirectional-1 (Bidirectional)	1024	6295552
Bidirectional-2 (Bidirectional)	1024	6295552
Dense-1 (Dense)	128	131200
Dense-2 (Dense)	1	129
Total params: First approach: 69588225 (265.46 MB)		
Total params: Second approach: 43977985 (167.76 MB)		

To evaluate the performance of BERT and LSTM against an untrained but state-of-the-art (SOTA) language model, GPT-3.5 Turbo was utilized for text classification. Abstracts were processed using the OpenAI library in Python, employing a suitable prompt to instruct GPT to classify the abstracts based on their content.

Retrieving all sections and paragraphs in the books posed a particular challenge due to the presence of numerous sections and subsections. The full text was downloaded based on these divisions, along with their respective paragraphs. Duplicated paragraphs were eliminated from both sections and subsections by labeling each paragraph with its unique ID tag in XML format. Additionally, each paragraph's content was labeled based on the title of the corresponding chapter, section, and subsection (if applicable). Furthermore, all possible reference styles were removed, and most XML flaws were addressed.

To ensure that each paragraph remained under 4000 characters (approximately 1024 tokens), the NLTK library in Python was utilized to segment larger paragraphs into sentences [8]. Subsequently, a series of Python code lines were employed to concatenate these sentences and generate larger paragraphs, each containing a maximum of 4000 characters.

The text data was processed using the PHI-2 language model, which consists of 2.7 billion parameters provided by Microsoft. The model was initialized using 8-bit integer format, followed by the application of low-rank adaptation techniques. Table 2 outlines the hyperparameters utilized in the model as well as the targeted layers. Subsequently, the model's performance was assessed through human evaluation after fine-tuning.

Table 2 LORA and training hyper parameters.

Hyper parameter	Value
Rank of the matrix	4
LORA alpha	16
LORA dropout	0.1
Target layer	All the layers
Learning rate	0.0001
Learning rate scheduler type	Cosine type

Results

Downloading the scientific text

of 0.98 for Bi-LSTM. Additionally, the F1-score for the test data for BERT is also higher than that of Bi-LSTM. Figure 4 shows the confusion matrix for the classification of the test datasets using BERT, Bi-LSTM, and GPT-3.5. The results indicate strong performance by the language models in classifying related abstracts accurately. The Bi-LSTM model, while achieving a misclassification rate of only 1.06% for unrelated abstracts, displayed lower performance compared to BERT and GPT-3.5. BERT and GPT-3.5 showed acceptable performance, with misclassification rates of 0.43% and 0.37% respectively for unrelated abstracts. The performance of the three models in classifying false negative abstracts varied, with GPT-3.5 exhibiting the weakest performance by misclassifying 10.22% of abstracts as related to geothermal energy. Conversely, BERT demonstrated strong performance in correctly classifying unrelated abstracts, almost entirely avoiding misclassifications.

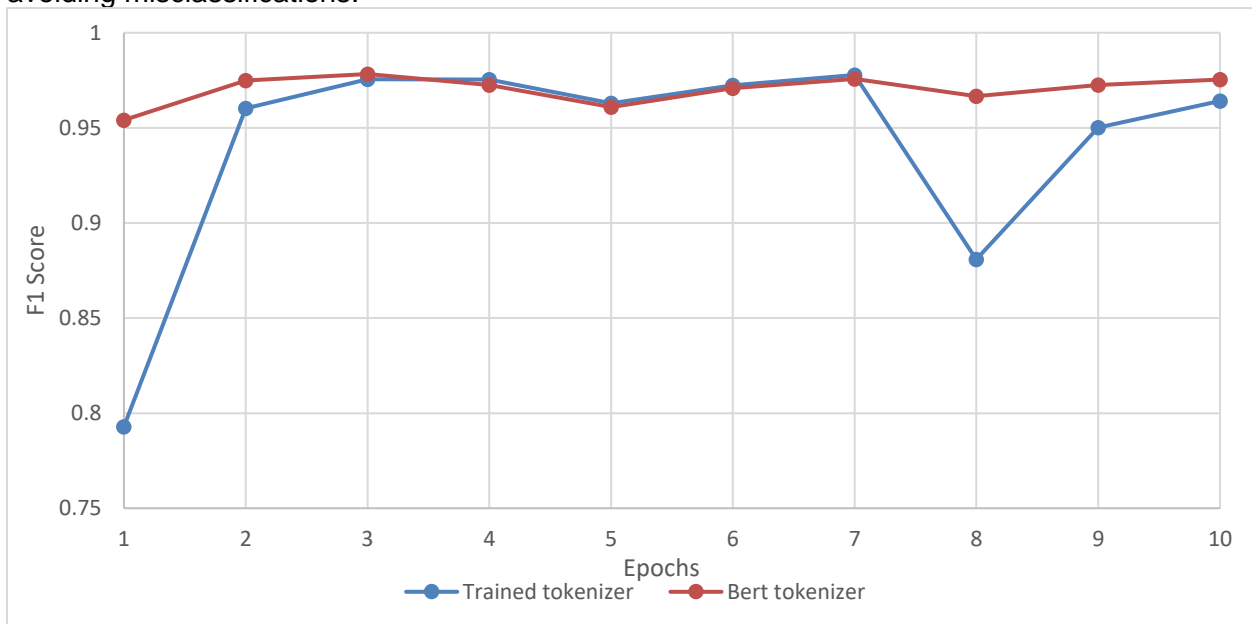


Figure 2. F1 scores for the Bi-LSTM neural network with different approaches.

After training BERT, 89,000 abstracts of papers were submitted to the model for classification. Following classification, 29,000 papers were annotated as related to geothermal science. From this subset, approximately 1,300 book chapters, identifiable by a specific DOI pattern, were processed to download the full text of each chapter. Of these, 413 book chapters were successfully downloaded, processed, and parsed into paragraphs, which were then stored in a data frame for further analysis.

The remaining papers were directed to another code segment to download the introduction section of each paper. The introduction sections were parsed into paragraphs and stored in a data frame format for subsequent processing.

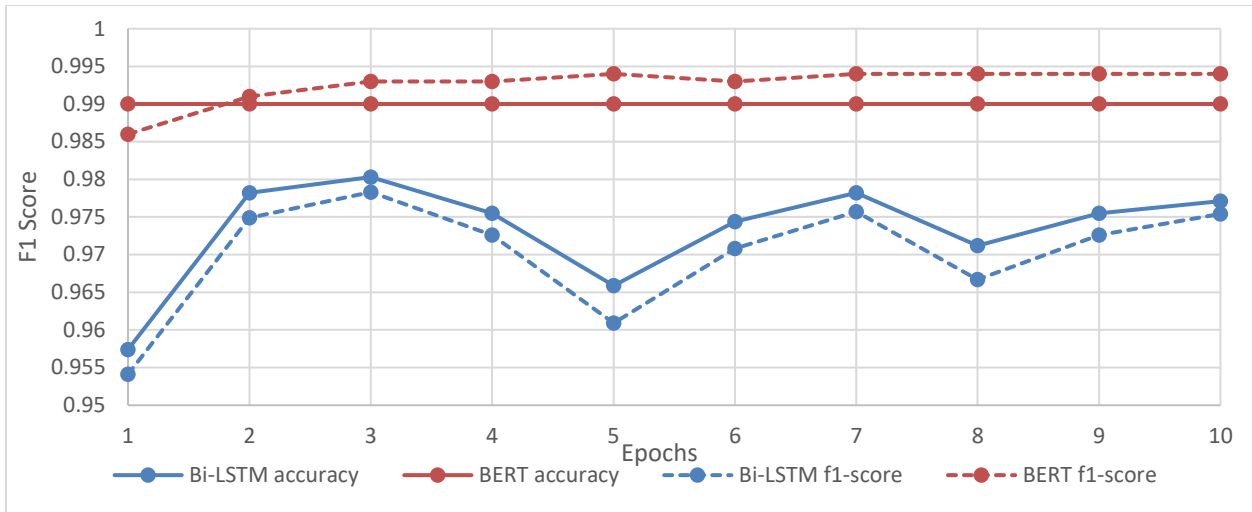


Figure 3. Comparison of Finetuned BERT with trained Bi-LSTM.

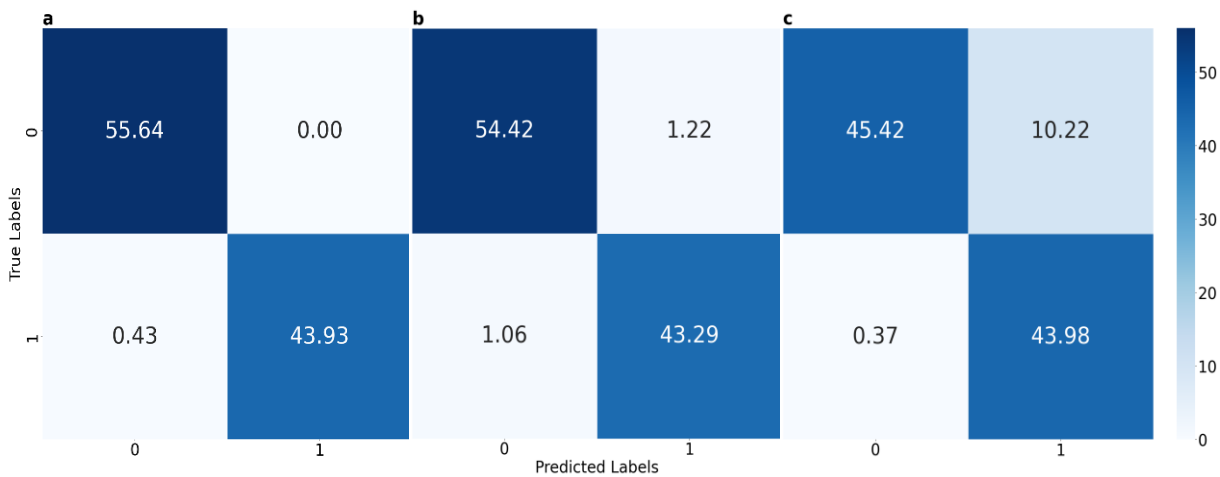


Figure 4 Confusion matrix for classifying test results a. fine tuned BERT b. trained Bi-LSTM c. using GPT-3.5 turbo

Training PHI-2 for question and answering.

Will be done in the future.

Conclusion

In conclusion, this study outlines a comprehensive methodology for leveraging advanced natural language processing (NLP) techniques, particularly focusing on the utilization of BERT (Bidirectional Encoder Representations from Transformers) for paper classification and other large language models such as PHI-2 for question and answering. Initial results demonstrate the exceptional performance of the fine-tuned BERT language model in classifying abstracts related to geothermal science, achieving high accuracy and F1 scores. Future work will involve training

PHI-2 for question answering tasks within the domain of geothermal science, promising further advancements in this area.

Overall, this study provides valuable insights into the potential applications of large language models in the field of geothermal science, paving the way for enhanced understanding, analysis, and automation of tasks within this domain.

References

- [1] D. Rangel-Martinez, K. D. P. Nigam, and L. A. Ricardez-Sandoval, "Machine learning on sustainable energy: A review and outlook on renewable energy systems, catalysis, smart grid and energy storage," *Chemical Engineering Research and Design*, vol. 174, pp. 414–441, Oct. 2021, doi: 10.1016/j.cherd.2021.08.013.
- [2] K. C. Leonard, F. Hasan, H. F. Sneddon, and F. You, "Can Artificial Intelligence and Machine Learning Be Used to Accelerate Sustainable Chemistry and Engineering?," *ACS Sustain Chem Eng*, vol. 9, no. 18, pp. 6126–6129, May 2021, doi: 10.1021/acssuschemeng.1c02741.
- [3] E. J. Hu *et al.*, "LoRA: Low-Rank Adaptation of Large Language Models," Jun. 2021, [Online]. Available: <http://arxiv.org/abs/2106.09685>
- [4] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "QLoRA: Efficient Finetuning of Quantized LLMs," May 2023, [Online]. Available: <http://arxiv.org/abs/2305.14314>
- [5] Y. Hu, J. Ding, Z. Dou, and H. Chang, "Short-Text Classification Detector: A Bert-Based Mental Approach," *Comput Intell Neurosci*, vol. 2022, 2022, doi: 10.1155/2022/8660828.
- [6] J. Devlin, M.-W. Chang, K. Lee, K. T. Google, and A. I. Language, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," 2018. [Online]. Available: <https://github.com/tensorflow/tensor2tensor>
- [7] K. R. J and D. T, "Accurate Short Text Classification for Improving Accuracy by using Bi-LSTM in comparison with LSTM," in *2023 Eighth International Conference on Science Technology Engineering and Mathematics (ICONSTEM)*, 2023, pp. 1–6. doi: 10.1109/ICONSTEM56934.2023.10142587.
- [8] E. Loper and S. Bird, "NLTK: The Natural Language Toolkit," 2002. [Online]. Available: <http://nltk.sf.net/>.