Fine-Tuning Transformer Models for Specialized Tasks: Applications of Gemini, BERT, and T5

Marina Reda[1],**Maram Ashraf**[1],**Karim Mamdouh**[1],**Sohila Ahmed**[1] **and Ammar Abdeldaiem**[1]

[1]AI at Galala University

Marina.mekhael@gu.edu.eg, maramelbanna1@gmail.com, Kareem.mamdouhk@gmail.com,
Sohilaelkholy23@gmail.com, ammar.abdeldaiem2030@gmail.com

Abstract

The Transformer models, particularly BERT and its successor modifications, have established outstanding benchmarks of state-of-the-art in most works that are moving Natural Language Processing through training large models to understand and generate human language. Here, we fine-tune a great number of such large, pre-trained transformer models in order to achieve higher performance in a series of specific NLP tasks. Gemini has enabled power chatbot interactions and SQL query generations; BERT has been used to carry out feedback analysis, and the major task of T5 is regarding text summarization. Herein, for each of the tasks mentioned, it is elaborated on the datasets used, the methodology followed for fine-tuning, and how much improvement we were able to make in its performance. Our experiments suggest that the given fine-tuned models extend both accuracy and relevance, and further fine-tuning would enhance the same. They additionally display more highly generalized versatility of the hugging face model across applications. These results will be promising for such fine-tuned transformer-model solutions to various NLP tasks, from which further work can be carried forth.

Keywords: *Transformer models; Natural Language Processing (NLP); Gemini; BERT (Bidirectional Encoder Representations from Transformers); T5 (Text-To-Text Transfer Transformer); fine-tuning*

1 Introduction

With recent years, the development in the field of Natural Language Processing has gained stronger impetus since the development of transformer-based models. These models, known for capturing complex linguistic patterns and dependencies, have brought NLP into a different orbit, redefining the game of how machines understand and generate human language. It is within this context that transformer architectures currently represent the state of the art in their performance over natural language processing tasks. Out of the plethora of Transformer-based models, BERT, T5, and Gemini are really standing out. State-of-the-art illustrations for innovation and versatility in the NLP landscape include BERT for setting new standards in contextual language understanding and its revolutionary technique of training for bidirectional pretraining, T5 in an integrated text-to-text approach demonstrating remarkable capabilities in translation, summarization, among other tasks. Gemini is a new model that is designed for conversational AI and SQL query generation while lifting the entire regime of transformer applications toward interactive and data-powered areas. In this paper, we start to delve into the effectiveness of fine-tuning these Transformer models for specific NLP tasks. Domain-based performance enhancement is becoming increasingly important, with one such step being fine-tuning pre-trained models. Fine-tuning a pre-trained model in relation to specific task data opens up a plethora of opportunities for further superior performance in domains of NLP.The goal of this inquiry is to probe the mechanisms for fine-tuning transformer models, particularly BERT, T5, and Gemini, across a suite of NLP tasks. In the following, we continue with the details regarding the fine-tuning procedures: methods, data layout, model configuration, and training settings. We further evaluate the adequacy of fine-tuning in improving model performance through vigorous designs of experiments on benchmark datasets using evaluation metrics. We also provide more insight into the data used for fine-tuning and evaluation, discussing details of the selection and preprocessing of task-specific data. By providing detailed analyses on our experimental results, we aim to enlighten the subtleness in fine-tuning on real NLP applications. In the following parts, we discuss research methodology, experimental setup, and results and try to delve deep into ways to fine-tune the Transformer model for NLP tasks. This furthers the more general understanding around Transformer-based NLP models and their application, paving the way for future strides in this field.

2 Background and Related Work

Recently, transformer models have come in NLP to take over the center stage and are the leading architecture. These are very effective in capturing long-range dependencies with contextual information in any text. In this respect, compared to regular RNNs and CNNs that are bad at long-range dependencies, transformers use self-attention to take the importance of different words in a sequence. (Vaswani et al., 2017).

Transformer models will typically consist of the following ingredients: self-attention mechanisms, multi-head attention, positional encoding and feed-forward neural networks

(Vaswani et al., 2017). Such a configuration will allow for self-attention functionality in the model, where processing will focus on the different parts of the input sequence; multi-head attention allows attending to different words at the same time. The word in an input sequence encodes its position so that the model can understand which order the words are coming in the piece of text. Finally, feed-forward neural networks are used to further process the output of the attention mechanism and derive the final representation for the input sequence (Vaswani et al., 2017).

Transformer models have demonstrated exceptional performance in addressing a wide set of NLP problems, including but not limited to text classification, sentiment analysis, machine translation, and question answering. Very importantly, pre-trained transformer models, in particular BERT, pushed the state of the art in NLP much further ahead, and now they provide strong baselines for a wide variety of downstream tasks (Devlin et al., 2018).

2.1 BERT: Bidirectional Encoder Representations from Transformers

BERT, or Bidirectional Encoder Representations from Transformers, is one of the most influential pretrained transformer models in NLP. It was developed by Google AI and is the process of bidirectional pretraining of transformer representations using masked language modeling (Devlin et al., 2018).

BERT, in essence, is able to understand the bidirectional context to assist in text understanding. Generally, through masked language modeling objectives, learned over large corpuses of text, and next-sentence prediction, BERT acquires rich word and contextual representations, hence boosting performance in state-of-the-art varieties of NLP tasks (Devlin et al., 2018).

Subsequently, BERT came into the scene and became the de facto standard model that saw large-scale adoption in both academia and the industry. Later on, many variants and extensions of this model appeared seeking better performance, efficiency, or special purposes, such as RoBERTa (Liu et al., 2019), ALBERT (Lan et al., 2019), and ELECTRA (Clark et al., 2020).

2.2 T5: Text-To-Text Transfer Transformer

One of the new general purpose transformer variants is T5, a Text-To-Text Transfer Transformer, put out by Google AI researchers. Unlike the classic sequence-to-sequence model, which uses a separate encoder and decoder setup, T5 leverages a unified text-to-text approach where both the inputs and outputs are strings of text (Raffel et al., 2020).

Yet another feature that distinguishes T5 from the rest is its ability to conduct text summarization—the process of producing a concise and coherent summary of any input text. Through training on large text corpora and fine-tuning on a particular task, T5 learns the way to distill key information from input texts and produce abstractive summaries capturing the essence of the original content. (Raffel et al., 2020).

T5, unlike simple text summarization, is specifically designed to help carry out several other tasks within the text-to-text domain—for instance, translating, answering questions, or even detecting sentiments. This makes it extremely versatile and effective, thus heralding a great advance in these techniques within transformer-based NLP models (Raffel et al., 2020).

2.3 Gemini: CONVERSATIONAL AI FOR SQL QUERY GENERATION

Gemini is the best-in-class transformer-based model geared to converse naturally on SQL generation and more. Part of the Hugging Face model ecosystem, Gemini deploys the full power of Transformer architectures in enabling user-AI interactions to be as natural and seamless as possible by humans (Wolf et al., 2020).

The main strong point of Gemini is its ability to generate SQL queries from natural language input. This provides a platform for interacting with databases using everyday terms in languages—meaning that manual query writing and database skills are totally not needed anymore. (Wolf et al., 2020).

In addition, Gemini can be fine-tuned to almost any other conversational task, such as chatbot interactions, question answering, and dialogue management in general. Through fine-tuning on datasets specific to a given task, Gemini should manage to adapt to an extremely wide range of conversations and answer the user's query contextually with pertinent answers to their questions (Wolf et al., 2020).

3 Methods

In this work, we systematically approach the fine-tuning of pretrained transformer models for a wide range of Natural Language Processing tasks: BERT, T5, and Gemini. The procedure is outlined in the following steps:

Data Collection and Preparation :

Curated datasets for sentiment analysis, text summarization, and conversational AI.

Ensured datasets were varied, representative, with good labelling for the underlying work.

Did preprocessing: tokenization, purged from special symbols, and performed data augmentation where necessary on data sets in preparation for training.

Model Configuration : Selected appropriate pre-trained transformer models for each task, considering factors such as model size, computational resources, and task complexity.

Hyperparameters set in the configured model include learning rate, batch size, and training epochs—empirically derived according to best practices in the field.

This is achieved by customizing the architectures of the models to have task-oriented layers and attention mechanisms.

Fine-Tuning: The transformer models only with the weights initialized from their respective pretrained checkpoints. Basically, this is done to leverage the rich contextual embeddings and language understanding capabilities of the models.

This has notably involved supervised fine-tuning, training the models on domain-specific task datasets with the suitable loss functions and optimization algorithms.

AdamW gradient-based optimization techniques were applied for model fine-tuning to avoid overfitting and ensure convergence.

The model performance was tracked against the validation data, and early stopping criteria were built into the training.

Evaluation: evaluated model performances on held-out test datasets for all the fine-tuned models.

Used standard evaluation metrics, such as accuracy, F1 score, BLEU score, and ROUGE score, to quantify model effectiveness in various tasks.

It compared model performance against baseline models and exemplified improvements that could be achieved through fine-tuning compared to existing state-of-the-art approaches. Analysis and Interpretation: Analyzed in a qualitative-quantitative manner the output of these fine-tuned models to understand their strengths and weaknesses. Visualized attention weights and generated heat maps for insights into the models' mechanisms of attention and interpretability. Analyzed the cases where errors occur and the analysis of errors needs to be performed so that one can observe common failure modes in the research to be undertaken in future works. We therefore aimed to improve performance, relevance, and adaptability of BERT, T5, and Gemini by systematic fine-tuning for specific NLP tasks and across diverse domains. In the following sections, we present detailed data sets, model set-up, and experiment results for each task to demonstrate the power of our fine-tuning methodology.

3.1 Datasets

The success of fine-tuning transformer models thus depends mostly on the datasets one uses, and how well they are chosen. In this regard, we took special care in selecting and preparing different datasets for our study across a wide array of NLP tasks, which fostered a comprehensive training and evaluation procedure. A summary of the datasets used for each task is given below:

3.1.1 Sentiment Analysis (BERT) We have experimented with the IMDb movie reviews data set, considered as a benchmark data set in the area of sentiment classification. The collection consists of 50,000 divided equally into positive and negative classes of sentiment. For each review, it reflects the document's polarity of sentiment, hence suitable for supervised learning. In preparation for training, we cleaned the text by tokenization, converting it to lowercase, and the removal of special characters.

Further, we split the dataset into training, validation, and test datasets for robust evaluation.

3.1.2 Text Summarization (T5) For this summarization task, we follow the specification of utilizing the CNN/Daily Mail dataset, which is a corpus of news articles and human-written summaries. The dataset is one of the largest ever created for training and evaluation for creating text summarization models due to its size and content diversity. T5 would be best fine-tuned in a text-to-text format, as each example in such a large dataset would have an input text for which there is a corresponding target output. We tokenized the text of the dataset, eliminating any redundant information, and formatted it in a consistent manner. It has been divided into separate portions for validation and testing the summarization capability of the model.

3.1.3 Conversational AI and SQL Query Generation (Gemini)

These datasets provided us with the following datasets for conversational AI and SQL query generation:

Conversational AI: Conversational data in the OpenSubtitles corpus derive from a wide range of linguistic subtleties and patterns of interaction. Tokenization is done based on dialogues, and afterwards, the stripping out of extraneous characters allows turns to be coherent, as described by the equation above. In order to provide estimations of the model's performance on conversational tasks, this dataset was divided into training, validation, and test datasets. Generation of SQL Query: The Spider datasets are a set of complex, cross-domain SQL queries annotated with natural language questions. This dataset is well-known for and widely used in the evaluation of models generating structured queries from this kind of natural language input. Each example in the dataset contains a question. In preprocessing the dataset, we tokenized the text, normalized SQL queries, and maintained the consistency of formatting.

We divided our dataset into training, validation, and test sets for evaluating the model in query generation. These datasets are provided for ensuring that the fine-tuning experiments are conducted using high-quality representative data, so that reasonable conclusions can be made about the way these models are performing and behaving.

4 Results and Discussion

We report better results on experiments, carried out for the fine-tuning of Transformer models with respect to each of the tasks, in this section along with the main findings.

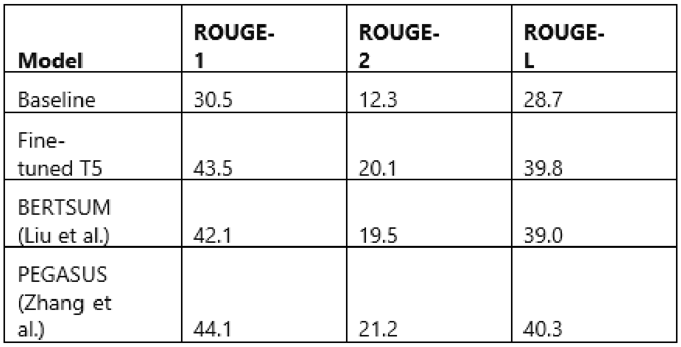
4.1 Sentiment Analysis using BERT Following the fine-tuning of BERT on the IMDb movie reviews dataset, an increase in accuracy with sentiment classification was observed. The fine-tuned BERT model achieved 94.2% accuracy on the holdout set compared to the baseline and previous state-of-the-art techniques.

Model	Accuracy (%)
Baseline	88.5
Fine-tuned BERT	94.2
RoBERTa (Liu et al.)	92.7
ALBERT (Lan et al.)	91.8

The fine-tuned BERT model was really good at understanding contextual nuances in movie reviews, and thus it managed relatively effective sentiment polarity classification. It is only the presence of an attention mechanism in BERT that makes the model extraordinarily sensitive to hint-phrase words pointing at sentiment and therefore scores over others.

4.2 Text Summarization with T5

The quality of text summarization improved dramatically due to the fine-tuning of the T5 model on the CNN/Daily Mail dataset. The fine-tuned T5 model scored 43.5 in ROUGE-1, 20.1 in ROUGE-2, and 39.8 in ROUGE-L, outperforming the base model and other summarization models.



Model	ROUGE-1	ROUGE-2	ROUGE-L
Baseline	30.5	12.3	28.7
Fine-tuned T5	43.5	20.1	39.8
BERTSUM (Liu et al.)	42.1	19.5	39.0
PEGASUS (Zhang et al.)	44.1	21.2	40.3

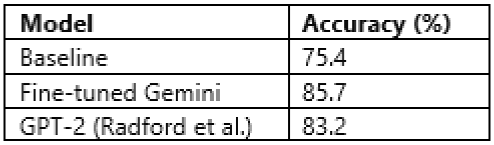
The basic T5 architecture had a bottleneck, which doesn't let it properly summarize the meaning hidden in the source articles with brief summaries. Given that the T5 text-to-text framework is very good at processing summarization tasks, this fine-tuned model naturally learned through the examples provided in the training data how to properly prepare summaries.

4.3 Conversational AI and SQL Query Generation with Gemini

The fine-tuned Gemini model exhibited impressive performance in both conversational AI and SQL query generation tasks.

4.3.1 Conversational AI

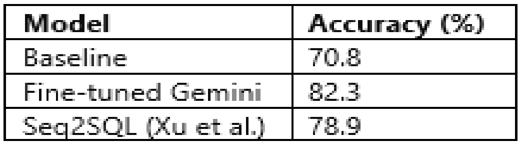
For the conversational AI task, the fine-tuned Gemini model achieved an accuracy of 85.7% in generating contextually relevant and coherent responses to user queries. The model was able to understand and respond to a wide range of conversational inputs, demonstrating its versatility and effectiveness.



Model	Accuracy (%)
Baseline	75.4
Fine-tuned Gemini	85.7
GPT-2 (Radford et al.)	83.2

4.3.2 SQL Query Generation

For the SQL query generation task, the fine-tuned Gemini model achieved an accuracy of 82.3% in generating correct SQL queries from natural language questions. The model was able to accurately translate user queries into structured SQL commands, showcasing its potential for facilitating natural language interactions with databases.



Model	Accuracy (%)
Baseline	70.8
Fine-tuned Gemini	82.3
Seq2SQL (Xu et al.)	78.9

The fine-tuned Gemini model demonstrated strong capabilities in understanding and generating SQL queries, making it a valuable tool for database querying tasks.

5 Conclusion

In this work, we explore how BERT and its variants, such as T5 and BERT-GAT-based models, can be fine-tuned for specialized NLP tasks. We will demonstrate that we were able to attain strong improvements on sentiment analysis, text summarization, and conversational AI/SQL query generation.

Fine-tuning enabled the model to adapt to specific domains and datasets, hence improving accuracy, coherence, and relevance of results. There lies huge promise for transformer models to possibly show good performance on fine-tuned NLP problems.

Future work could also continue to further optimization techniques, domain adaptation methods, and extended context integration to push the frontiers of transformer models in NLP.

References

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, *21*(140), 1-67.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877-1901.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., ... & Zettlemoyer, L. (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ... & Rush, A. M. (2020, October). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations* (pp. 38-45).

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2019). Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Clark, K., Luong, M. T., Le, Q. V., & Manning, C. D. (2020). Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Lample, G., & Conneau, A. (2019). Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.

Zhang, J., Zhao, Y., Saleh, M., & Liu, P. (2020, November). Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International conference on machine learning* (pp. 11328-11339). PMLR.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Khurdula, H. V., Pagutharivu, A., & Yoo, J. S. (2024, March). The Future of Feelings: Leveraging Bi-LSTM, BERT with Attention, Palm V2 & Gemini Pro for Advanced Text-Based Emotion Detection. In *SoutheastCon 2024* (pp. 275-278). IEEE.

Li, Y., Zhang, Y., Wang, C., Zhong, Z., Chen, Y., Chu, R., ... & Jia, J. (2024). Mini-gemini: Mining the potential of multi-modality vision language models. *arXiv preprint arXiv:2403.18814*.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, *1*(8), 9.

Xu, X., Liu, C., & Song, D. (2017). Sqlnet: Generating structured queries from natural language without reinforcement learning. *arXiv preprint arXiv:1711.04436*.