Natural Language Processing

Post-project paper

Marco Kuhlmann
Department of Computer and Information Science

LINKÖPING UNIVERSITY

This session

- Project guidelines
- Grading criteria
- Concrete examples

Instructions

Project assignments

The project assignments are centred around a small research project. You can choose to work on a pre-defined project or propose one independently. The total time budget for the project is 80 hours of work per person, evenly divided between group work and individual work.

The group work will be done in teams of 4–6 students, and most of the joint effort will be concentrated during the two weeks before re-exams week. The individual work will mostly happen during re-exams week and exams week. Each part is assessed on the basis of several deliverables, which are explained below.

You can get help and feedback on your project from the examiner. We recommend scheduling at least one meeting to pitch your project idea, but you are welcome to book additional appointments as needed.

Grading criteria

Group deliverables

Project presentation

The main group deliverable is a 10-minute oral presentation of your project at the “course conference”, which will take place during re-exams week.

Your presentation should answer the standard questions also answered in a research article:

- What have you done in this project?
- What question did you want to answer with it?
- What was your method?
- What results did you obtain?
- What conclusions do you draw from these results?

Guidelines for the post-project paper

D

Describe

Describe your work with the project. Focus on things that let you illustrate what you have learned.

E

Examine

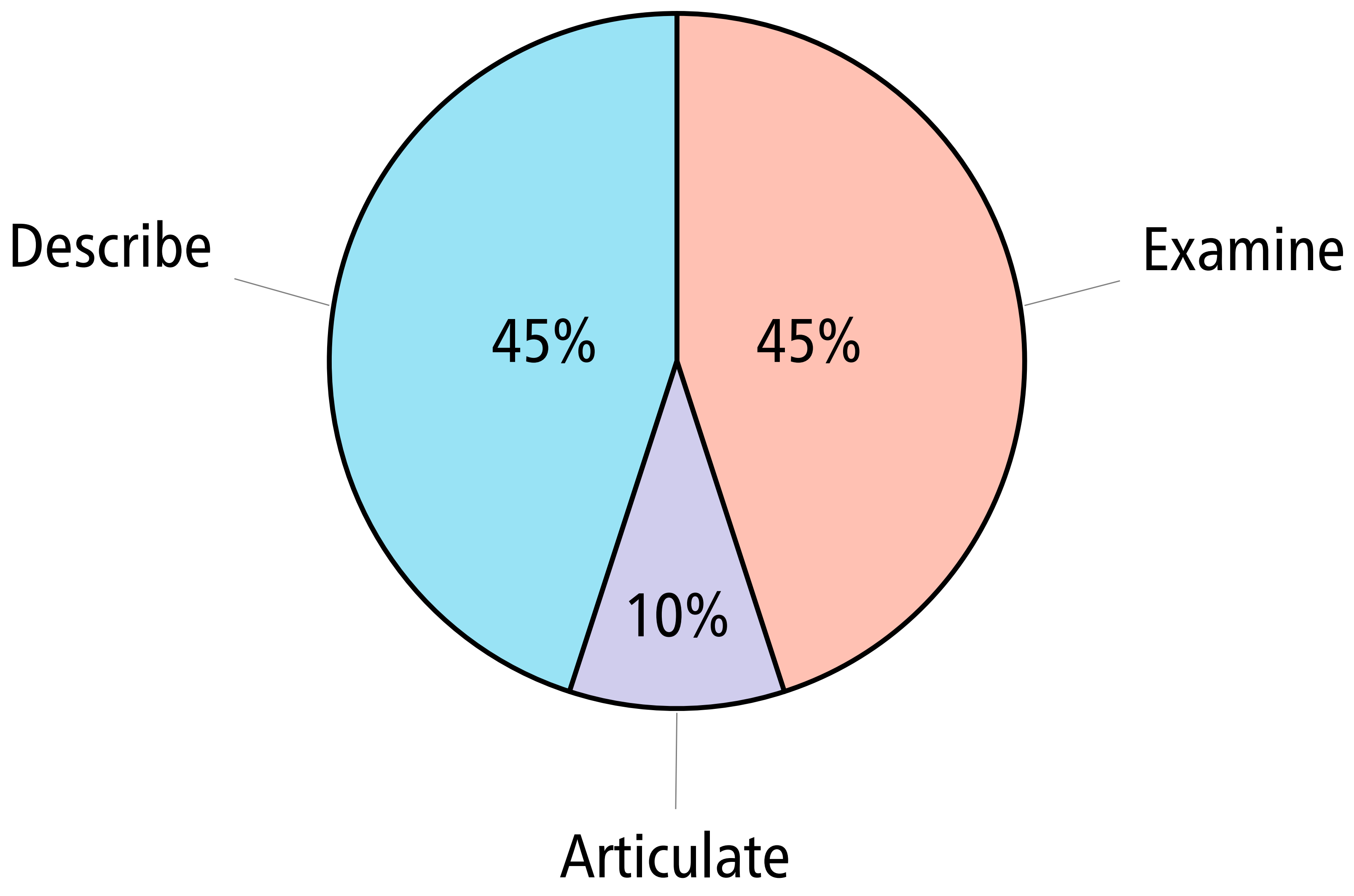
Examine your experience and link it to the relevant course content.

AL

Articulate Learning

Articulate your learning. What did you learn? How, exactly, did you learn that? Why does this learning matter?

Length distribution



Grading criteria

Grading criteria

https://www.overleaf.com/read/fvdjqsfjktjs#1dd21b

Description: Details and examples

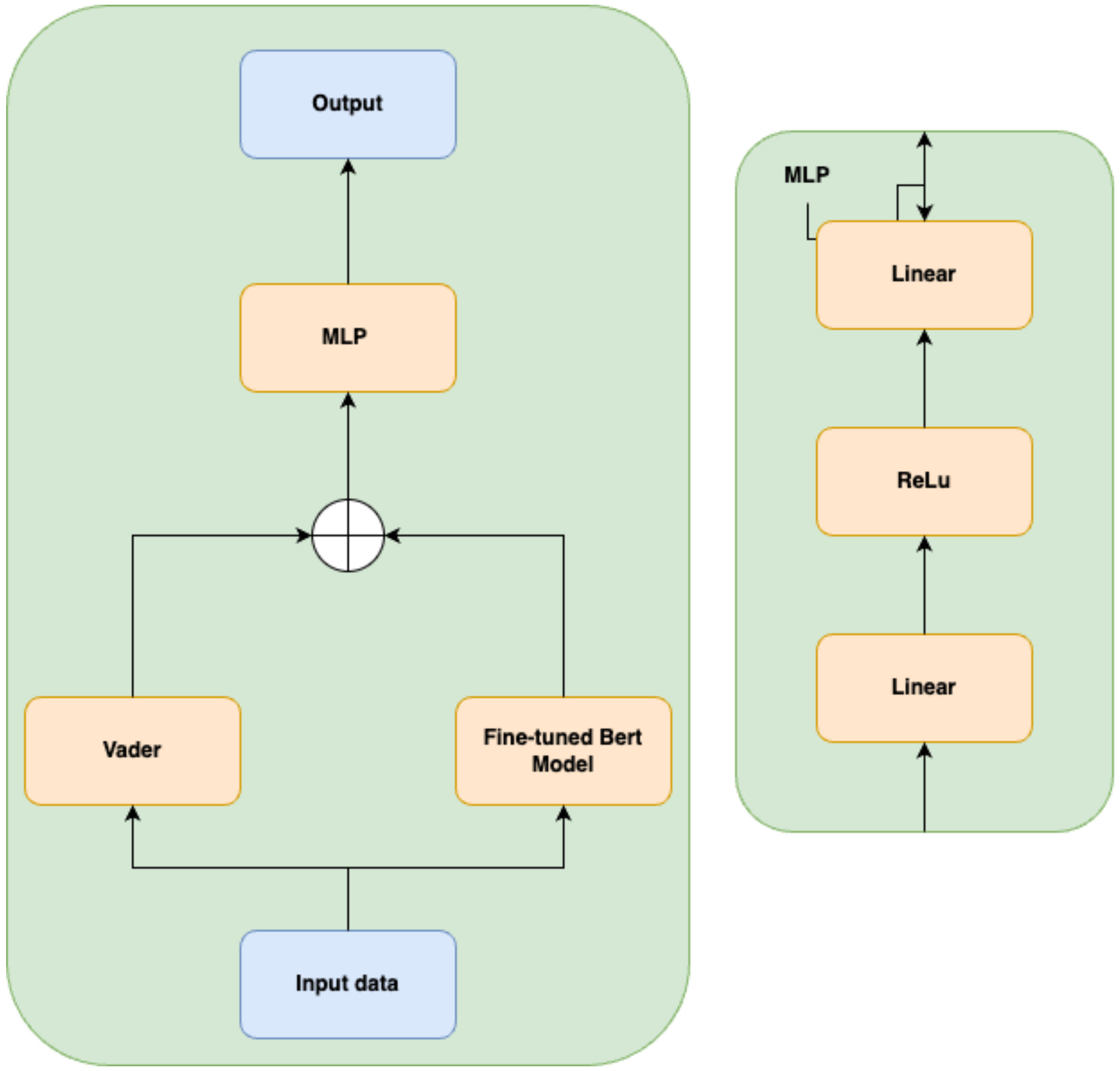


Figure 1: Method 1 MLP

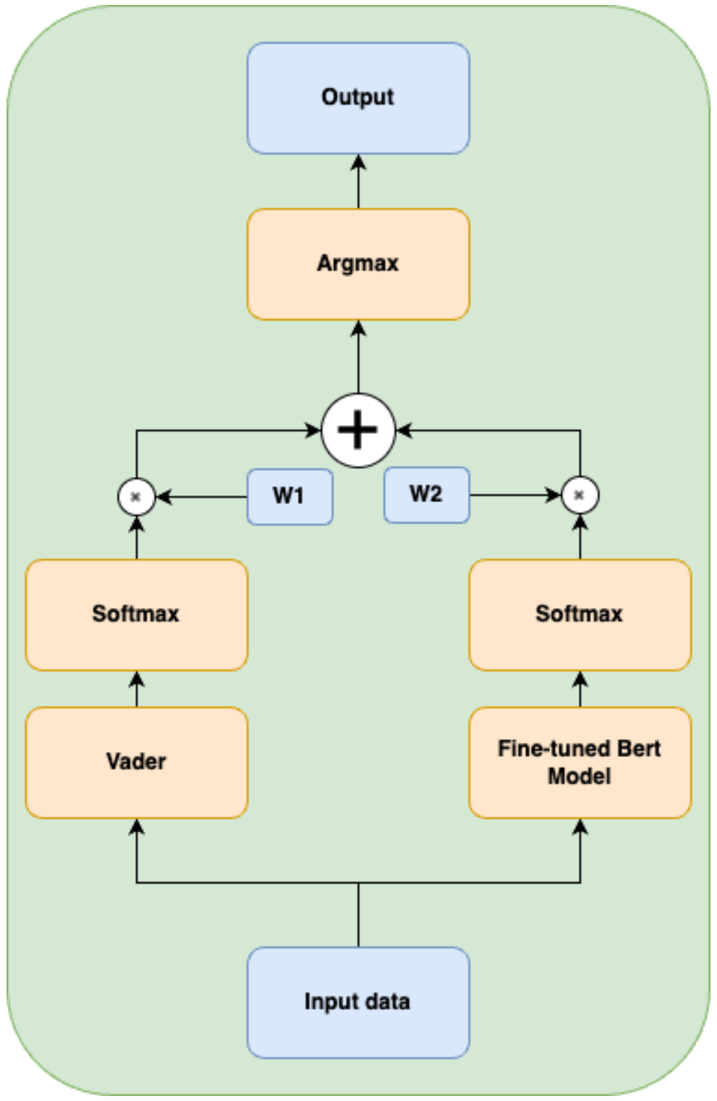
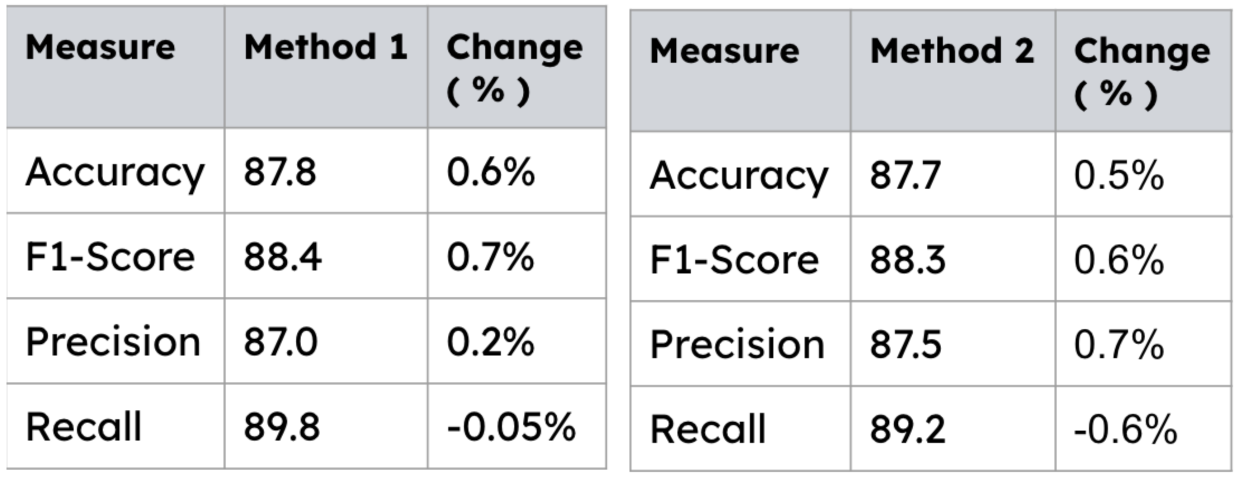


Figure 2: Method 2 Weighted Sum

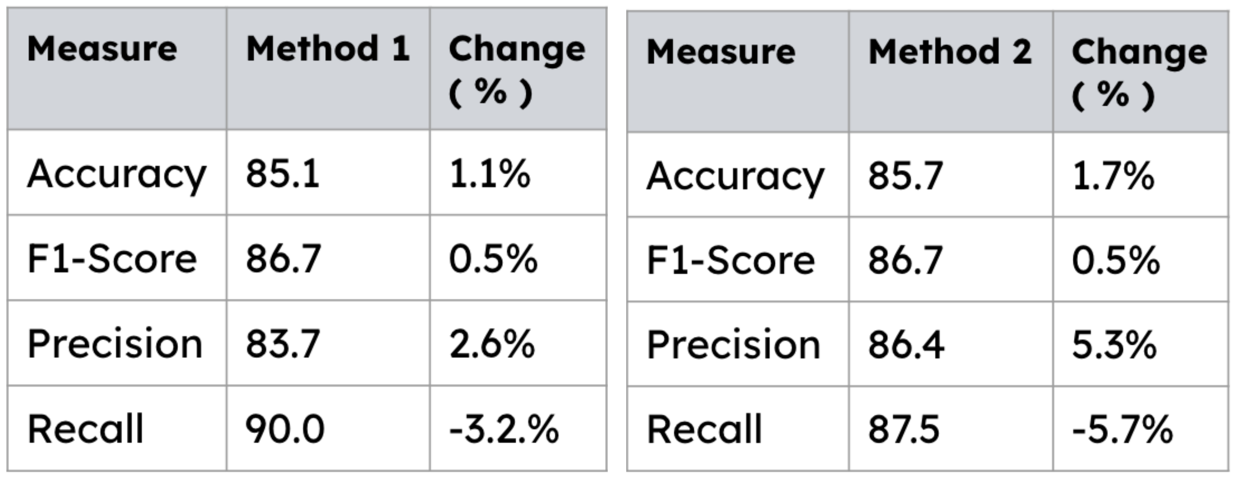
Description: Details and examples



Measure	Method 1	Change (%)
Accuracy	87.8	0.6%
F1-Score	88.4	0.7%
Precision	87.0	0.2%
Recall	89.8	-0.05%

Measure	Method 2	Change (%)
Accuracy	87.7	0.5%
F1-Score	88.3	0.6%
Precision	87.5	0.7%
Recall	89.2	-0.6%

Figure 3: Results larger dataset(120 words)



Measure	Method 1	Change (%)
Accuracy	85.1	1.1%
F1-Score	86.7	0.5%
Precision	83.7	2.6%
Recall	90.0	-3.2.%

Measure	Method 2	Change (%)
Accuracy	85.7	1.7%
F1-Score	86.7	0.5%
Precision	86.4	5.3%
Recall	87.5	-5.7%

Figure 4: Results smaller dataset (80 words)

Do not forget to also include a summary in text form!
Make it clear what I should look at, what trends you want me to see.

Description: Don’ts

- The Description should resemble a research article. Do not use the first-person perspective unless you absolutely have to.

- I do not need to know that you were responsible for scheduling your meetings, maintaining your repo, buying fika …

- This paper is not the right place to reflect on the quality of your team work or to apologise for your choice of topic.

Discussion points

The first grading criterion is about quality. The descriptors for grade 3 and grade 5 target difference readers.

- What background can you expect from each audience?
- What would require explanation?

Examine your experience

Part 2: Examine your experience

In this part, you should critically examine your personal project experience and connect it to the course content and any additional reading.

Prompts Respond to the following prompt:[2]

> What specific ideas and skills from the first part of the course were relevant to your project? How exactly did you use them in the project? Focus on those aspects that were particularly important for your own individual learning and growth.

For a higher grade, additionally address one of these more advanced prompts:

- What similarities, what differences were there between your prior understanding of the course content and the way in which it emerged in the project?
- Based on your experience and your analysis, was your understanding of the course content and any additional reading adequate? If not, what exactly was lacking?
- How has the project enhanced your understanding of the course content? Given what you know now, how would you do the project today?

[2] The prompts were adapted from Duke Service-Learning (2018).

Project Work

A wide variety of regular project work skills were naturally essential, such as communication and planning. However, as no issues arose within this field and it does not directly relate to TDDE09 it's not discussed below.

Due to the importance of TF-IDF and S-BERT for the project, understanding these was vital. The project group learned these methods via reading the corresponding research literature and as such the ability to read and understand such work was also important. Furthermore, as the field was NLP, a good grasp of the foundations and terminology therein was needed.

The entire project being written in Python required a good understanding of the language. Personally, I didn't implement anything using the TF-IDF or S-BERT theory though others in the group implemented TF-IDF and Window S-BERT, both of which did require said theory.

During the Natural Language Processing (NLP) course, I was introduced to a type of neural network architecture called transformers. Transformers utilize a self-attention mechanism that provides direct access to all elements in a sequence, regardless of its length [1]. In our project, as outlined in section 1, we utilized a BERT model as it is a large pre-trained language model that employs the transformer architecture as its backbone.

From the course material, I learned that the BERT model comprises a considerable number of free parameters, making fine-tuning the model very computationally expensive. Therefore, we decided to explore different approaches to reduce the training time while maintaining accuracy.

In the NLP course, I also learned about generative pre-trained transformers (GPT), which can be fine-tuned for a classification task by freezing the model and only fine-tuning a classification layer [2]. This got me thinking that the same approach could be applicable to the BERT model. The article by Peters et al. [4] confirmed this notion. As illustrated in Table 2, the authors achieved nearly the same accuracy by fine-tuning and freezing the model and only fine-tuning the classification layer. However, our results for both techniques were much lower, possibly due to the fact that *SICK-E* dataset, used in the study, only contains three labels, while the *Riksdagen* dataset contains eight labels, and the latter contains texts up to 512 tokens.

Authors	Dataset	HEAD	FULL	Adapters
Peters et al.	SICK-E	84.8%	85.8%	-
Pfiffer et al.	SICK	76.30%	87.30%	86.20%
Our team	Riksdagen	34.22%	31.34%	59.35%

Table 2: Results obtained from various implementations conducted by our team, as well as by Pfiffer et al. and Peters et al. HEAD refers to freezing the model and only fine-tuning the classification layer. FULL referrers to fine-tuning model and classification layer. Adapters referrers to freezing the model and fine-tuning the adapters.

Additional reading, particularly the paper by Pfeiffer et al. [5], proved useful in our project, as it introduced adapters and how these lightweight neural networks can be integrated into a pre-trained model for downstream tasks. Adapters enabled us to freeze the model parameters and only fine-tune the adapters, effectively doubling our accuracy while halving the training time compared to fine-tuning the model (*see Table 1*). When comparing our results with those obtained by Pfeiffer et al., I would say that our results are good, as we used longer sequences and had more labels.

The observation that fine-tuning only adapters could double the accuracy compared to fine-tuning the model might seem strange. However, Pfeiffer et al. suggestes that this could be due to the fact that adapters can have a regularization effect on certain datasets, resulting in better performance on average for specific tasks, even though only a smaller proportion of weights are trained.

One thing I found especially hard during the implementation process was verifying that the different implementations accurately updated only the desired weights. This includes verifying that the process of freezing the model and the implementation of the adapters were done correctly. This was a result of the fact that I found it difficult to fully understand the documentation provided by the Hugging Face Library. However, even though it was hard to verify that our implementation of adapters were correct, the fact that me and Linus collaborated helped. We made two significantly different implementations of

Discussion points

- What specific technical concepts and skills from the course are most relevant to *your* project?
- What could be an example of a situation where you found your knowledge from the course lacking?

Articulation

The specialised S-BERT model is something I didn't know of before the project. Reading the research paper and discussing it gave me a solid theoretical foundation which was further supplemented by an intuitive understanding when the evaluation and analysis were performed. Furthermore, having considered and discussed fine-tuning I could hopefully improve the model more on a particular domain. A company I'm working part time for with AI-applications is soon releasing an FAQ and as such, using S-BERT for question suggestions could be a way to naturally integrate this FAQ into their current chatbot.

The general BERT architecture was also something I didn't have much experience with beforehand. While discussed during the TDDE09 lectures, testing S-BERT and QA-BERT really showcased the flexibility of the model, and the general power of the architecture was noticeable when reading the theory. Therefore I now consider BERT to be an adaptable and strong tool at my disposal. Currently I'm planning for a project which would include multi-language sentiment analysis and am considering using varieties of BERT therein.

This project was also one of my first times doing a recorded video presentation. The largely positive response makes me believe it's something I could consider for other courses going forward. Of course, issues with this were also discussed but keeping these in mind should improve potential future recordings.

3 Articulation of Learning

Throughout the course of this project, I have come to understand that there are various approaches available for fine-tuning pre-trained language models, and that their efficacy can differ greatly depending on the task at hand. This was evidenced by a comparative analysis of the results generated by our implementation, and those presented by Peters et al. and Pfiffer et al. (*see Table 2*). As outlined in section 2, our findings indicated that adapter-based fine-tuning produced superior results, while fine-tuning the entire model and freezing the model while fine-tuning only the classification layer resulted in comparatively inferior performance. Conversely, Peters et al. and Pfiffer et al. achieved favorable outcomes by either fine-tuning the entire model or only fine-tuning the classification layer. The primary distinction between our task and theirs pertains to the input sequence size and number of labels, leading me to the conclude that label count is a crucial factor that has a significant impact on performance. However, as our task and the one by Pfiffer et al. achieved success using adapters despite vastly differing input sequence length and label counts, I will bear in mind that adapters appear to be a useful technique for a broad range of tasks in future work. Additionally, the realization that high accuracy can be attained without fine-tuning the entire model is valuable from an environmental perspective, as computationally intensive tasks consume a considerable amount of energy. This aspect will undoubtedly be critical in the future of IT.

Given the various methods available for fine-tuning pre-trained language models, as described above, I have gained insight into the significance of conducting a comprehensive literature review. To efficiently examine a substantial number of articles, I have come to appreciate the skill of quickly scanning through material to identify pertinent sources. Once such sources have been identified, a more thorough examination of the material enables me to further refine my search to locate the most relevant content. The ability to conduct effective literature reviews will be an invaluable tool throughout my professional career, particularly when tasked with familiarizing myself with new subjects.