ML4LG - assignment 1

Jaka Kokošar

1 Introduction

The increasing number of published randomized controlled trials (RCTs) poses a challenge for medical investigators in efficiently discovering the required information [1]. As researchers often start their literature review by skimming through abstracts. Structured abstracts with semantic headings such as objective, method, result, and conclusion can significantly aid this process. However, most abstracts are not structured in this way, making it difficult to access relevant information quickly. Automatic classifying of abstract sentences could enable researchers to find relevant information quicker, particularly in fields such as medicine, where abstracts can be lengthy. This motivation highlights the importance of accurate algorithms for sequential short-text classification and emphasizes the potential benefits it could bring to medical research.

2 Data

Dernoncourt and Lee [1] have created a dataset named PubMed 200k RCT [1], comprising about 200,000 abstracts of randomized controlled trials containing 2.3 million sentences. Each abstract sentence is labeled with one of the following categories: *BACKGROUND*, *OBJECTIVE*, *METHOD*, *RESULT*, or *CONCLUSION*, indicating its role in the abstract. In our work, we used a subset of 20k abstracts. The training dataset consists of 180040 sentences, the validation dataset consists of 30212 sentences, and the test dataset includes 30135 sentences. Class and document length distributions are summed up by Figure 1.

3 Methodology and Results

We used the Hugging Face Transformers library for this assignment to develop a solution for sequence classification tasks. Specifically, we fine-tuned two state-of-the-art language models: BERT and GPT. We aimed to evaluate these models' performance on the PubMed 20k RCT dataset.

We used the tokenizer corresponding to each language model to preprocess the data. This involved tokenizing the sentences in the abstracts and converting them into the input format required by the models. We also applied a truncation step to limit the maximum length of the input sequence. The truncation threshold was set at the 99th

[1] https://github.com/Franck-Dernoncourt/pubmed-rct

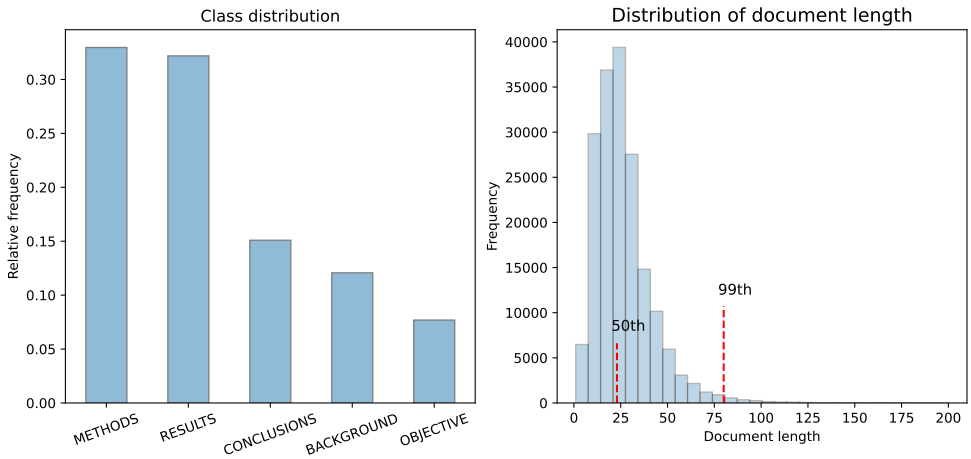

Fig. 1 Left image: class distribution in the training dataset. Right image: document length distribution with indicator for median and 99th percentile of the distribution.

percentile of the length distribution of the sentences in the training set. Additionally, we applied padding to the input sequences to ensure they were of equal length.

It should be noted that we did not perform any extensive hyperparameter tuning for this task. Instead, we primarily relied on the default parameters the Hugging Face Transformers library provided. However, we did refer to the BERT paper [2] for recommended fine-tuning parameters. Specifically, we used a batch size of *32*, a learning rate of *5e-5*, and we settled on *four* epochs. Additionally, to reduce training time, we utilized the Transformers API for early stopping during training. The best model was selected based on the *f1* metric. The results are summed up in table 1

Table 1 Computed metrics on validation nad test dataset for best models.

	Validation			Test		
Model	precision	recall	f1	precision	recall	f1
BERT	0.8738	0.8725	0.8717	0.8628	0.8646	0.8628
GPT2	0.8441	0.8441	0.8418	0.8646	0.8628	0.8620

Note: computed metrics are average weighted. Training time for BERT was around 62min (8500 steps) and for GPT was around 142min (17500 steps).

Our experimental results showed that the fine-tuned BERT and GPT models achieved performance metrics that were comparable to the baselines reported in the original paper (best achieved *f1-score* is 90.0). However, our results did not significantly outperform the baselines, suggesting that additional steps may be necessary to improve the performance of the models for this task. Nonetheless, our findings demonstrate the power of newly created language models and the open-source community enabling their rapid use for a wide range of natural language processing tasks. With reasonably low code, researchers and practitioners can fine-tune state-of-the-art language models for specific tasks, which helps accelerate research progress in NLP.

References

[1] Dernoncourt, F., Lee, J.Y.: Pubmed 200k rct: a dataset for sequential sentence classification in medical abstracts. arXiv preprint arXiv:1710.06071 (2017)

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