# NLU course projects - Assignment 3 - SA

*Ettore Saggiorato (247178)*

University of Trento

ettore.saggiorato@studenti.unitn.it

## 1. Introduction

This assignment delves into the fine-tuning of BERT [1] for aspect term extraction in Aspect-Based Sentiment Analysis (ABSA) on the `laptop14` partition of the SemEval-2014 Task 4 dataset [2].

The evaluation of model performance is conducted using standard metrics: precision, recall, and F1 score, all of which are calculated based on the SemEval evaluation framework [3, 4].

The ultimate aim of this fine-tuning process is to refine BERT's ability to accurately extract only the aspect terms relevant to the task.

## 2. Implementation Details

The implementation builds upon the approach used in the second assignment, with key differences primarily in the data preprocessing and evaluation steps. The dataset is preprocessed by simplifying the labels to just two categories: ['T', 'O'], where T' corresponds to a token that represents an aspect term, and O' corresponds to other tokens. An example of the preprocessing is shown below:

```
Boot=T-POS time=T-POS is=O
super=O fast=O ,=O around=O
anywhere=O from=O 35=O seconds=O
to=O 1=O minute=O .=O
```

This snippet represents an original sentence from the dataset. After preprocessing, the transformed version of the same sentence would appear as:

```
Boot=T time=T is=O super=O
fast=O ,=O around=O anywhere=O
from=O 35=O seconds=O to=O 1=O
minute=O .=O
```

For consistency, the `bert-base-uncased` model and tokenizer are employed, consistent with prior assignments. The tokens are processed using BERT's tokenizer to ensure the input is correctly formatted for the model. Notably, only the first subtoken of a word receives the appropriate tag ID, with all subsequent subtokens being treated as padding (ID = 0). This strategy is essential for avoiding evaluation distortions caused by issues related to sub-tokenization.

### 2.1. Fine-Tuning

The fine-tuning process adheres to the guidelines provided in the original BERT paper [1]. Specifically, training is carried out over four epochs, with batch sizes of 16 or 32 tokens. To ensure the results are robust and reliable, each experiment is repeated five times, with the final outcomes averaged across the runs.

Several regularization techniques are employed to promote training stability:

- **Gradient Clipping:** This technique is utilized to prevent the occurrence of exploding gradients, ensuring smoother model training.
- **Learning Rate Scheduler:** A learning rate scheduler dynamically adjusts the learning rate during training, helping the model converge more smoothly. However, given that the task involves only four epochs, the benefits of this adjustment might be limited.

## 3. Results

The performance of the model is summarized in Table 1, where precision, recall, and F1 score serve as the primary evaluation metrics. The key insights from the experiment are as follows:

- **Gradient Clipping:** The inclusion of gradient clipping resulted in a modest improvement in the F1 score, especially when compared to the baseline model. This suggests that gradient clipping plays a subtle but valuable role in stabilizing training and slightly enhancing overall model performance.
- **Learning Rate Scheduler:** As expected, the learning rate scheduler did not exhibit any noticeable positive effect on model performance. Despite its potential for smoother convergence, it did not contribute significantly to improvements in precision, recall, or F1 score in this particular task.

Additionally, extending the training by two extra epochs led to a slight but noticeable enhancement in both F1 score and recall, further confirming the importance of prolonged training for better generalization and model accuracy.

## 4. References

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

[2] X. Li, "E2e-tbsa," https://github.com/lixin4ever/E2E-TBSA/tree/master/data.

[3] ——, "Evaluation script for semeval-2014 task 4," https://github.com/lixin4ever/E2E-TBSA/blob/master/evals.py.

[4] SemEval 2014 Organizers, "Semeval-2014 task 4: Aspect based sentiment analysis - data and tools," 2014. [Online]. Available: https://alt.qcri.org/semeval2014/task4/index.php?id=data-and-tools

| Part | Model name | Epochs | Batch size | Gradient clip. | Train. sch. | F1 (%) | Precision (%) | Recall (%) |
|---|---|---|---|---|---|---|---|---|
| 1.0 | simple_16_2e-5 | 4 | 16 | T | F | 69.25 | 68.90 | 69.62 |
| 1.0 | simple_16_3e-5 | 4 | 16 | T | F | 68.88 | **69.88** | 68.04 |
| 1.0 | simple_16_5e-5 | 4 | 16 | T | F | 69.05 | 68.23 | 69.97 |
| 1.0 | simple_32_2e-5 | 4 | 32 | T | F | **69.67** | 67.16 | **72.46** |
| 1.0 | simple_32_3e-5 | 4 | 32 | T | F | 69.19 | 68.01 | 70.47 |
| 1.0 | simple_32_5e-5 | 4 | 32 | T | F | 68.74 | 67.39 | 70.19 |
| 1.1 | grad_16_2e-5 | 4 | 16 | F | F | 68.95 | **69.15** | 68.83 |
| 1.1 | grad_16_3e-5 | 4 | 16 | F | F | 68.48 | 67.44 | 69.78 |
| 1.1 | grad_16_5e-5 | 4 | 16 | F | F | 70.13 | 69.08 | **71.42** |
| 1.1 | grad_32_2e-5 | 4 | 32 | F | F | 69.92 | 68.57 | 71.36 |
| 1.1 | grad_32_3e-5 | 4 | 32 | F | F | 68.82 | 67.60 | 70.22 |
| 1.1 | grad_32_5e-5 | 4 | 32 | F | F | **70.21** | 69.10 | 71.39 |
| 1.2 | sch_grad_16_2e-5 | 4 | 16 | T | T | 70.10 | 68.35 | 71.96 |
| 1.2 | sch_grad_16_3e-5 | 4 | 16 | T | T | 70.24 | 68.53 | 72.05 |
| 1.2 | sch_grad_16_5e-5 | 4 | 16 | T | T | **70.37** | **68.92** | 71.89 |
| 1.2 | sch_grad_32_2e-5 | 4 | 32 | T | T | 68.86 | 67.20 | 70.63 |
| 1.2 | sch_grad_32_3e-5 | 4 | 32 | T | T | 69.96 | 67.75 | **72.33** |
| 1.2 | sch_grad_32_5e-5 | 4 | 32 | T | T | 69.78 | 67.97 | 71.70 |
| 1.3 | longer_grad_16_3e-5 | 6 | 16 | F | T | 69.30 | **69.80** | 68.86 |
| 1.3 | longer_sch_grad_16_2e-5 | 6 | 16 | T | T | 70.91 | 69.47 | 72.43 |
| 1.3 | longer_sch_grad_16_5e-5 | 6 | 16 | T | T | **71.01** | 69.52 | **72.56** |

Table 1: *Experiment Results.* ***Bold*** *values represent the best scores for each part.*