NLU project exercise lab: 11

Matteo Minardi (238789)

University of Trento

matteo.minardi@studenti.unitn.it

1. Introduction

In the first part, the goal was to create a pipeline model for Subjectivity and Polarity detection tasks. The pipeline has to be composed of two different models:

- The first model predicts if a sentence is subjective or objective;
- The second model performs the polarity detection of a document after removing the objective sentences predicted by the first model;

After this pipeline, the second exercise about Aspect Based Sentiment Analysis has not been implemented due to lack of time. All the necessary information to implement anything was taken from the papers cited in the references.

2. Implementation details

For the first part, I took inspiration from the previous lab and decided to fine-tune a BERT model for the classification of subjective/objective sentences. The implementation just needed a supplemental linear layer with 2 output neurons to classify the output class.
This architecture is so general that it was also used for the second part of the task about Polarity detection, since even that can be seen as a two-classes classification problem.
The problem on the padding in the collate function has been solved in the same way as the previous lab: I decided to pad all sequences of the dataset with the same length, so that there wouldn't be any differences between batches. The selected length is the maximum length of a sentence found in the corpus. This turned out to be a good idea because the maximum length of a sentence between the two different datasets of the tasks do not match, so it can be dynamically adapted to any case, without having to cut any sentence.

The model of the first part was used to remove the objective sentences for the second task, and was able to generalize to a completely different dataset in a very good way.

For the second task, I performed it in three different ways:

- removing only stopwords and punctuation
- removing objective sentences, stopwords and punctuation
- removing only objective sentences and stopwords.

I decided to remove the stopwords because they are not as meaningful and to increase the computational capabilities. Obviously, the objective sentences have been removed according to the predictions of the first model.
I did not remove the sentences of length greater than the maximum of the one in the first dataset to resemble a real life scenario for the removal of objective sentences, although removing those would probably increase the final performance.

3. Results

Subjectivity task Model	Accuracy
BERT	0.956

Table 1: *Best config of part 1*

Polarity task Model	Accuracy
BERT (no stopwords and punct)	0.522
BERT (no objective sents and stopwords)	0.541
BERT (no objective sents, stopwords and punct)	0.568

Table 2: *Best configs of part 2:*
Good to notice that all the tests performed an increase in the polarity task models, being able to guess more than half of the situations.

As a remark, the final accuracy could improve if the length of the padded sequences between the two datasets would match, because the initial model could make better predictions to filter the objective sents from the second set of sentences and the second model wouldn't need to predict some (possibily) objective sentences, but I kept it this way to resemble a real life scenario where a pretrained model is used on novel data. Also, the number of removed sentences would have been too much to produce statistically significant results.

4. References

[]