Text-based multitask model

This appendix of the work has no comparability with the other models that were developed, and thus it has been presented as auxiliary instead of attaching it to the main work.
The idea behind this section was to re-construct a model so as to fully train the BERT model for multi-task classification, given that the MultiOutputClassifier cannot be used too easily with the BERT model. Therefore, an algorithm based on different output heads was developed. However, as the training algorithm is slow and memory-costly, a base model, which was much more familiar and simple, was firstly constructed.

Initially, he BertForSequenceClassification model was used, as it supports multi-label classification[1]. The only difference with this and the regular BERT model is that an explicit request for multi-label classification has to be made. With this algorithm, different learning rates were tested, and although the final results were not so good (the algorithm is primarily trained for binary classification, so it does not work well with, for instance, arousal), the chosen learning rates were used for the second model. The obtained results are shown in Table 1.

	Validation	Test
Illness	0.76	0.75
2-labeled valence	0.71	0.49
Arousal	0.27	0.21

Table 1: Macro average results for a simple text-based multitask model with learning rate $1e-5$

The second configuration, with the more complex multi-headed algorithm[2], was then developed. With this new configuration, the test data would have given an error if sentences were filtered by length as it was done in the main work, as it would leave no negative instances for valence and a label mismatch would arise. Therefore, and seen that SMOTE could not be used either, the sentences were not discarded by length. This is expected to have a bit of a detrimental impact.
This algorithm uses BertModel as the body to which then a 3-headed model, each of them

[1]This idea was taken from https://discuss.huggingface.co/t/multilabel-text-classification-trainer-api/11508.

[2]This code was mostly based on github code https://gist.github.com/emillykkejensen/aa7535c29538a956d5b9c41e31f731a1

with classification purposes, is attached. The model used the learning rate and Adam epsilon optimized with the previous model, and the batch size and number of epochs were inherited from the base text model.

With this algorithm, the accuracies shown in Table 2 were achieved; only the test macro averages were computed, as the training algorithm automatically computes only the accuracy of the validation dataset. The final confusion matrices are also shown in Figures 1, 2 and 3. The results are worse than the ones computed with the pre-trained BERT model, but the principle of training the whole model for a more complex task is thereby shown. More fine-tuning could probably improve the results.

	Validation accuracy	Test accuracy
Illness	0.78	0.90
2-labeled valence	0.65	0.57
Arousal	0.60	0.82

Table 2: Test results for multitask classification based on text depending on initial learning rate

The test F_1-score macro averages were 0.80 for illness detection, 0.57 for valence classification, and 0.30 for arousal, with a weighted macro average of 0.76 for this last task. It is clear viewing the confusion matrices that the F_1-score depicts a more realistic efficiency than accuracy does.

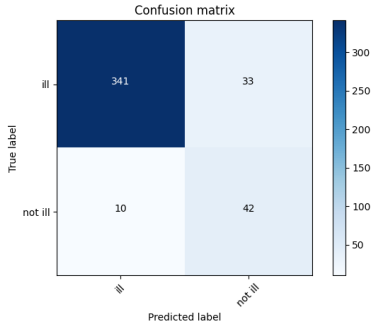

Figure 1: Text-based multilabel classification: illness detection (test)

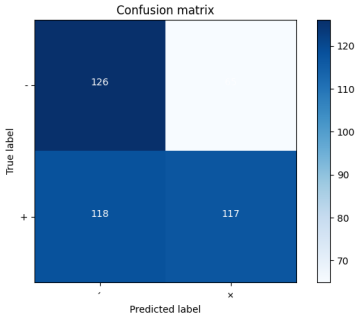

Figure 2: Text-based multilabel classification: valence classification (test)

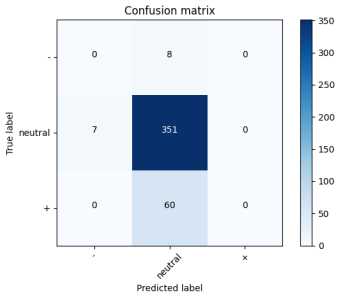

Figure 3: Text-based multilabel classification: arousal classification (test)