Assignment 2

Paolo De Angelis, Giacomo Piergentili, Francesco Pigliapoco and **Lorenzo Scaioli**
Master's Degree in Artificial Intelligence, University of Bologna
{ paolo.deangelis7, giacomo.pergentili2, francesco.pigliapoco, lorenzo.scaioli }@studio.unibo.it

Abstract

In this experiment we try to solve the Human Value identification problem on the level 3 categories by training three different BERT-based models and two baselines. To create each model we insert certain information in the input data in order to discover their influence on the results. We analyze the errors made by the three models and guess the causal factors. A great source of surprise was the fact that the real improvement between the models was due to the ability to categorize certain classes. Lastly, we propose possible solutions to the problems that we faced.

1 Introduction

This report presents an approach to tackle the task of Human Value identification (Kiesel et al., 2022). The state-of-the-art language model BERT has shown impressive results on various language understanding tasks, and we show here one possible fine-tuning implementation of a specific pretrained model. This was done by using a multi-label classification approach and by evaluating each level 3 category separately. BERT is a powerful model that can understand and use complex language patterns to identify different ways values are shown in arguments. However, it might not work as well if there is not enough data, if the data is not balanced, or if the data has cultural biases. We proposes some possible solutions to address this challenge and improve the performance of BERT on this task. Our experiments show how the more information are provided to the `Trainer` the better the results on the metrics are. The key difference we found in the performance of the trained models lies in their ability to correctly predict the two least frequent categories.

2 System description

In order to solve multi-label classification of human values on the dataset we had to define transformer-based models, and modify the rough data in order to fit the pretrained models requirements.
Firstly we ordered the dataset with reference to the four categories of human values, then, we use a specific tokenizer to represent every word in a suitable format for the models defined. In particular, we chose the BERT model `prajjwal1/bert-tiny` therefore the tokenizer was defined through `AutoTokenizer` in relation to the selected model. At the end, every sentence is padded so that each element of the dataset is composed by a `premise`, a `conclusion` and a binary value representing the `stance`.
For this experiment we define three BERT-based models and two baseline:

- **c model** is a BERT model that takes as input just the `conclusion`.
- **cp model** is a BERT model that takes as input the `conclusion` and the `premise`.
- **cps model** is a BERT model that takes as input the `conclusion`, the `premise` and the `stance`.
- **Majority classifier** is an implementation of a `DummyClassifier` model for each category, with strategy `most_frequent`.
- **Random uniform classifiers** is an implementation of a `DummyClassifier` model for each category, with strategy `uniform`.

3 Experimental setup and results

For this project we utilized the Hugging Face `Transformers` library to take advantage of a class called `AutoModel ForSequenceClassification`, designed to select automatically the appropriate model

architecture based on the name or path of the pre-trained model supplied. In our case we chose to use `prajjwal1/bert-tiny` which is one of the smallest variant of BERT (Bhargava et al., 2021) (Turc et al., 2019). While creating the model we also specified the problem type as `multi_label_classification` which will set the loss function to `BCEWithLogitsLoss`. The process of training the model is then executed by using the `Trainer` class (also provided by Hugging Face) in PyTorch. We customized the training process by specifying the `TrainingArguments` like the learning rate (1e-3), the bacth size (8), the optimizer (`ASGD`) and the scheduler (`ReduceLROnPlateau`). In particular, the optimizer is a component that updates the parameters of a model in order to minimize the loss function and the scheduler reduces the learning rate when a metric has stopped improving for a specific number of epochs.

Validation set	f1	precision	recall
c model	0.6009	0.6094	0.6316
cp model	0.7158	0.7357	0.7245
cps model	0.7324	0.7296	0.7555
Majority Baseline	0.4360	0.3866	0.5
Uniform Baseline	0.5279	0.5952	0.4963

Table 1: evaluation of the macro metrics of the models on validation set.

4 Discussion

As we can see from the results in Table 1, the model that achieved the best performance scores is **cps model**. It is interesting to see that the biggest improvement in performance is present when the premises are added to the conclusions, in fact the f1 goes from 0.6009 to 0.7158, while adding the stances just slightly increases the scores.

```
Conclusion:       We should fight urbanization
Stance:           1
Premise:          urbanization creates too many problems with the infrastructure
True labels:      [0. 1. 1. 1.]
Pred labels CPS:  [0. 1. 1. 1.]
Pred labels CP:   [0. 1. 1. 1.]
Pred labels C:    [0. 0. 1. 1.]
```

Figure 1: Example of how the premise helps the model to categorize correctly

During the error analysis we also noticed that the improvement of the f1 mostly depends on the least frequent categories: this means that, as shown in the table 2, the f1-score, and consequently all the other metrics, greatly improve on the categories with less samples (*Openness to change* and *Self-enhancement*), while they mostly remain the same on the categories with a much bigger sample size (*Conservation* and *Self-transcendence*).

f1	OTC	SE	C	ST
c model	0.2188	0.4411	0.8585	0.8854
cp model	0.4978	0.6243	0.8586	0.8825
cps model	0.5596	0.6236	0.8652	0.8810
Majority BL	0.0	0.0	0.8585	0.8854
Uniform BL	0.4438	0.5016	0.6073	0.6141

Table 2: evaluation of the f1 metric of every category for each model.

The behaviour we just described has to be expected because the dataset is highly unbalanced: almost all of the sentences belong the the *Conservation* and *Self-transcendence* categories, therefore even a simple model, like the one based on **Majority Classifiers** is able to correctly classify most of the sentences that belong to those 2 categories.
After the training we plotted various metrics and noticed that the loss of the training kept decreasing for every model, while the loss of the evaluation remained steady or just slightly decreased. This is probably not caused by not enough training, but from the inability of the model to recognize those less frequent categories given the imbalance in the dataset. This imbalance was probably caused by merging level 2 annotations to level 3 category. Having a more balanced dataset would have probably lead to an improved ability of the model to detect the correct category. Another possible solution to achieve better results could be using a bigger pre-trained model like `roberta-base`.

5 Conclusion

In this experiment we trained three different models to solve the Human Value identification problem and analyzed the results of the best model. Our study has demonstrated promising results. We have successfully developed a model that can identify and categorize key human values with a degree of accuracy depending on the frequency of a category in the dataset. We were surprised to see that by adding the premises to the conclusions there is a big improvement in the ability of the model to detect the right category, while, adding the stance, produces just a slightly better result.

6 Links to external resources

Here's the link to our GitHub repository with the code and all the trained models: https://github.com/LorenzoScaioli/NLP_multi-label-text-classification-with-transformers.

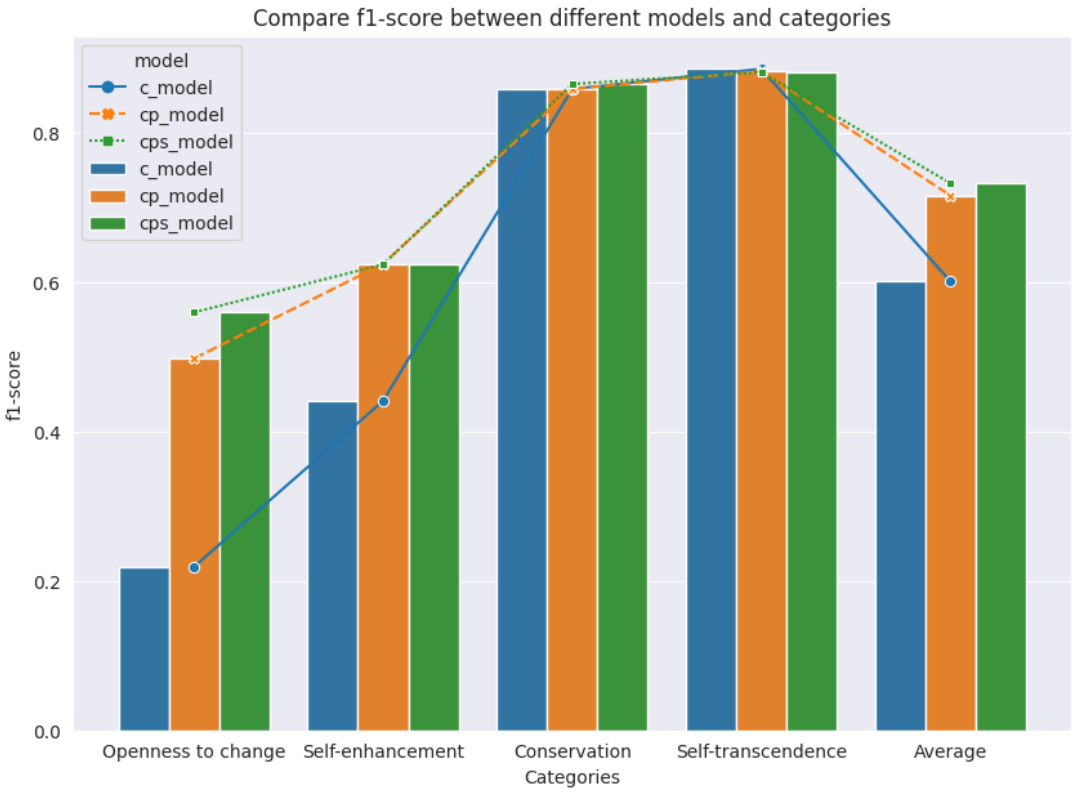
References

Prajjwal Bhargava, Aleksandr Drozd, and Anna Rogers. 2021. Generalization in nli: Ways (not) to go beyond simple heuristics.

Johannes Kiesel, Milad Alshomary, Nicolas Handke, Xiaoni Cai, Henning Wachsmuth, and Benno Stein. 2022. Identifying the human values behind arguments. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4459–4471, Dublin, Ireland. Association for Computational Linguistics.

Iulia Turc, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Well-read students learn better: The impact of student initialization on knowledge distillation. *CoRR*, abs/1908.08962.

Appendices


Compare f1-score between different models and categories
model
c_model
cp_model
cps_model
c_model
cp_model
cps_model
f1-score
0.0
0.2
0.4
0.6
0.8
Openness to change
Self-enhancement
Conservation
Self-transcendence
Average
Categories