Assignment 2

Davide Capacchione, Álvaro Esteban Muñoz and **Luca Trambaiollo**
Master's Degree in Artificial Intelligence, University of Bologna
{ davide.capacchione, alvaro.estebanmunoz, luca.trambaiollo }@studio.unibo.it

Abstract

This assignment focuses on addressing the Human Value Detection task, which involves a multi-label classification challenge within natural language processing. We'll begin by restructuring the initial dataset, We'll then define 3 variants of BERT Models trained on increasingly complex textual input. We'll compare these model variants and assess their performance in comparison to simpler baseline models.

1 Introduction

The Human Value Detection task aims to identify human values from textual arguments presented in Conclusion, Stance, and Premise fields. The task uses a collection of 20 value categories, drawn from social science and literature, which are grouped in super-level categories (Kiesel et al., 2022). Known approaches to this problem involve utilizing pre-trained large language models, such as BERT, due to their ability to capture contextual information effectively. Our approach also involves using a large language model due to its versatility and ability to capture contextual information. In this experiment, we modified the initial dataset to focus on 4 super-level categories. We constructed three distinct datasets, each utilizing a different input concatenation approach. Subsequently, we used these datasets to train individual BERT models implemented using the PyTorch library, all sharing a similar structure. Following training, a classification process was conducted. The obtained results from the various BERT models were then compared both amongst themselves and against baseline models using F1 score. The experiments revealed that increased input complexity, represented by the inclusion of more fields in the dataset concatenation, resulted in improved model performance.

2 System description

The initial dataset contains binary information about 20 level 2 categories. We created a new dataset by merging annotations of level 2 categories into four specific level3 categories: 'Openness to change,' 'Self-enhancement,' 'Conservation,' and 'Self-transcendence.'

For model input preparation, three datasets were formed using different text concatenation approaches:

- DatasetC: using solely the Conclusion field.
- Dataset_CP: formed by concatenating the Conclusion and Premise fields.
- Dataset_CPS: formed by concatenating the Conclusion, Premise, and Stance fields.

The input was then encoded by tokenizing textual information using the bert-base-uncased tokenizer. In the last dataset, stance was encoded as a binary variable and added separately to the embedding matrix.

Initially, two baseline models, a random uniform classifier, and a majority classifier were implemented. They each consisted of individual binary classifiers for each of the four categories. The BERT models—BERT w/C, BERT w/CP, BERT w/CPS share the same implementation. Each of them takes the respective dataset as input. They utilize the bert-base-uncased pre-trained model for embedding input sentences, incorporating a dropout layer to prevent overfitting (see Figure 1).

Stance information, if included, is added to the pooled output before passing through class-specific classifiers (see Figure 2) (**OTC_cls**, **SEnh_cls**, **Cons_cls**, **STra_cls**). CrossEntropyLoss criterion is applied to compute the loss against the true labels.

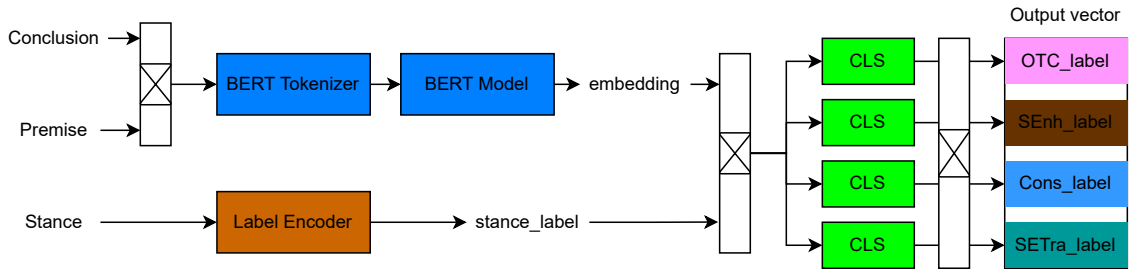

Figure 1: Architecture of the custom BERT models (stance included)

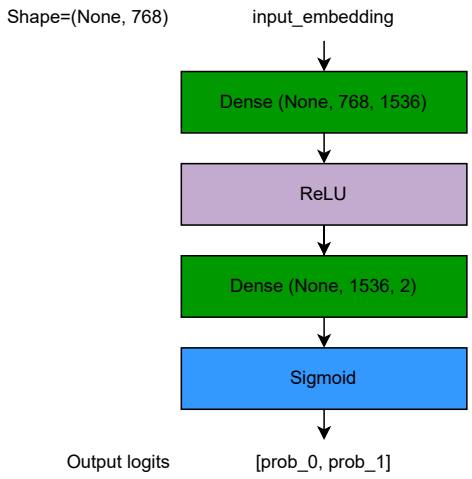

Figure 2: Class-specific Classifier

3 Experimental setup and results

We performed three different experiments using all of our previously defined classifiers. Each experiment has been run with a different random seed for reproducibility (23, 42 and 87). To avoid the confusion with the numbers we show the results obtained with just one of the seeds (23) after 5 epochs:

Classifier	OTC F1	SEnh F1	Cons F1	STrans F1	macro-F1
Rand Unif	0.3692	0.4442	0.5481	0.6337	0.4988
Majority	0.0000	0.0000	0.8304	0.8917	0.4305
w/C	0.3521	0.5898	0.8585	0.8853	0.6714
w/CP	**0.5774**	0.6654	**0.8649**	**0.8868**	**0.7486**
w/CPS	0.5571	**0.6822**	0.8620	0.8858	0.7468

Table 1: Results on the validation set

4 Discussion

From the results we can see that those models which include both conclusion and premise perform quite better than the rest of the models due to the fact that the baselines don't consider the input. The results with the other seeds do not allow us to draw a conclusion about which one of the two is better. This is because including stance changes has minimal impact on the results, likely due to the possibility that including stance information might introduce redundant or irrelevant details to the model.

The most remarkable things from the gotten scores can be summarized in the followings:

- Conservation and self transcendence seem labels easy to be classified, most likely due to its frequency with one of the values. We can justify taking a look at the majority classifier that always picking the majority label will give us a really high score.
- Openness to change and self enhancement are, on the other hand, labels difficult to be classified. This could be due to the incapability of the classifier to get a clear idea of the pattern with the amount of data.

We can confirm our suppositions by taking a look at the confusion matrices for each class-specific classifier (see Figure 3)

5 Conclusion

We addressed the Human Value Detection task exploiting the potentialities of large language models (BERT). We compared the results obtained by training our models using the different text fields of the dataset in various ways. The experiments showed that including more fields in the dataset concatenation leads to better results. Unexpectedly, the inclusion of stance field in the CPS model, did not affect significantly the metric.

Some improvements, as referenced in (Daniel Schroter, 2023) and (Ma et al., 2023), explore different dataset preprocessing methods, ensemble models or a more balanced dataset, demonstrating potential enhancements in performance.

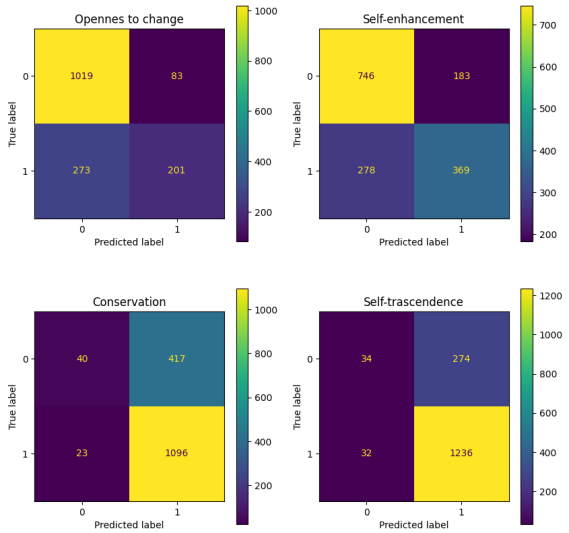

Figure 3: CP model confusion matrix for each label

References

Georg Groh Daniel Schroter, Daryna Dementieva. 2023. Adam-smith at semeval-2023 task 4: Discovering human values in arguments with ensembles of transformer-based models.

Johannes Kiesel, Milad Alshomary, Nicolas Handke, Xiaoni Cai, Henning Wachsmuth, and Benno Stein. 2022. Identifying the Human Values behind Arguments. In *60th Annual Meeting of the Association for Computational Linguistics (ACL 2022)*, pages 4459–4471. Association for Computational Linguistics.

Long Ma, Zeye Sun, Jiawei Jiang, and Xuan Li. 2023. PAI at SemEval-2023 task 4: A general multi-label classification system with class-balanced loss function and ensemble module. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 256–261, Toronto, Canada. Association for Computational Linguistics.