BERT-based Multi-label Human Value classification
NLP Course Assignment

Edoardo Fusa, Alberto Luise, and **Angelo Quarta**
Master's Degree in Artificial Intelligence, University of Bologna
{ edoardo.fusa, alberto.luise, angelo.quarta }@studio.unibo.it

Abstract

Human values detection has been a very important task regarding the field of debating AIs, enhanced by the organizers of scientific events called *Touché*.

The objective of this task is the classification of values reported by an argument in a multi-label fashion. Such problem is approachable in several ways but we adopted the most common one in literature consisting of a BERT network as encoder followed by a fully connected classifier. Although the chosen baselines achieved a challenging score, said approach led to promising results and highlighted how the given dataset affects the performance of the designed model.

1 Introduction

The inspiration of this assignment was drawn by the a specific task proposed by *Touché* community about human value classification given some arguments.

Being this challenge a public competition, the reports of the best achievements are available online alongside the solutions adopted. Looking at them, the main approach seem to be a BERT model connected to a standard classifier. However, each solution has its own peculiarities defined by a custom loss function (Ma et al., 2023), a modified BERT model (Honda and Wilharm, 2023) or an ensemble approach (Schroter et al., 2023).

The problem this assignment proposes is different because of the granularity of the considered labels. Furthermore, we decided to consider a computationally-light architecture and optimize the hyper-parameters, reporting results above the best performing baseline.

2 System description

Being the assignment's task part of the one proposed by *Touché* communuity, the dataset was already split into training, validation and test set. Each of them contains three features called: *Conclusion*, *Stance* and *Premise*. While *Stance* becomes a binary feature after the pre-processing, the remaining two are modified by applying the *Hugging Face* tokenizer to each string in that columns.

In order to test the degree of difficulty of the task, the baselines are increasingly complex: a *Random Uniform Classifier* that chooses the categories independently for each class. Then, a *Majority Classifier* taking decisions based on which class is predominant considering each class separately. Finally, a *One Classifier* that classifies all the samples in input as belonging to all classes.

On the other hand, the developed models (called *Classifier C*, *Classifier CP*, *Classifier CPS*) are meant to explore which feature in the dataset has a major impact on the classification outcomes. The first model takes just the *Conclusion* column of the dataset, the second takes the *Premise* column alongside *Conclusion* while the third one gets the entire dataset. Other than this specificity, the overall structure of the models in terms of architecture is preserved.

Each single model has a pre-trained *RoBERTa* backbone retrieved on *Hugging Face* (Hug). Then, an average self-attention pooling applied to the inputs (Chen et al., 2023) stabilizes the output size of the backbone and its results are concatenated together. Considering that a numerical feature appears in the dataset, it does not pass through the embedder.

To predict the classes, the classifier heads take the concatenated tensor and connect it to a fully connected layer having as many neurons as the number of classes. At the end of that last layer, a normalized $\tanh$ activation function fits the last layer's outputs into the $[0, 1]$ interval.

3 Experimental setup and results

Since the architecture is designed to be computationally inexpensive, the training loop is fairly standard. During said loop, the model gets evalu-

Model	Validation	Test
Random Baseline	0.533	0.505
Majority Baseline	0.432	0.427
One Baseline	0.726	0.688
Classifier C	0.679	0.643
Classifier CP	0.733	0.726
Classifier CPS	0.736	0.720

Table 1: Validation and Testing results in terms of F1 score

ated by considering the validation F1 score that it produces. Then that procedure has been run over several seeds for robust estimation and the best performing model has been saved according to F1 score.

For the sake of comparison, each model went through the same training, validation and testing process using the same hyper-parameters. Since the task requires the classification of the arguments, the most natural loss function is the *Binary Cross-Entropy*. For the optimization process, inspired by (Ma et al., 2023), the chosen optimizer is *AdamW* attached to a scheduler on the learning rate starting from 0.01 and linearly decaying through the 10 epochs.

Since the task requires crisp predictions, we set a threshold of 0.5 to the network outputs. The said procedure gave some interesting results stored in the Table 1.

4 Discussion

By analysing Table 1 emerges that the best performing baseline reports a *F1 score* below the value related to the *Classifier C*. Furthermore, the similarity in score between the *Classifier CP* and the *Classifier CPS* is noteworthy. The reason behind that similarity is that the *Stance* column does not carry information for the classification purpose.

Classifier CP and the *Classifier CPS* perform better because their backbones extract more meaningful embeddings from the substantially higher amount of data. In that matter, the lack of correlation among the labels does not bias the models during the learning process.

Diving deep into the results, it emerges that those results present a huge recall but the performance deteriorates if we consider the precision. The cause of that behaviour is linked to the highly unbalanced nature of the dataset (as displayed in figure 2). So, the models are more prone to misclassifying low represented classes as over-represented ones. Indeed, that is supported by the figure 1 that evidently portraits the lack of precision for the classes *Self-Enhancement* and *Openness to Change*.

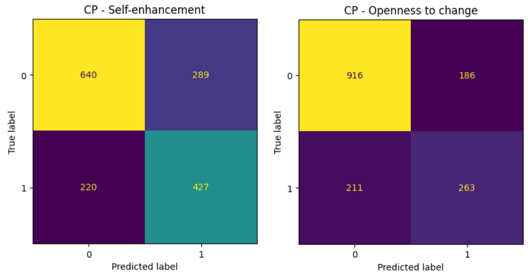

Figure 1: Confusion matrices of the least represented classes considering *Classifier CPS*

5 Conclusion

Concluding, the discussed solution balanced the computational inexpensiveness and the classification performance surpassing the baselines. However, there is room for improvements, even without modifying the networks' architectures.

Firstly, acting on the dataset would radically change the output scores. On the one hand, performing some data augmentation on the dataset in order to balance the number of represented classes. On the other hand, a better data labelization (Kiesel et al., 2022) could be heavily impactful over the solution for this task.

Finally, as a proven performance booster (Schroter et al., 2023), ensemble techniques can be employed: producing several models identical to each other trained over different splits of the dataset. That strategy efficiently tackles the imbalances carried by the dataset.

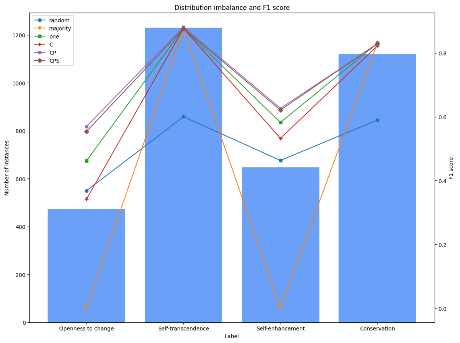

Figure 2: Distribution/F1 score relation

6 Links to external resources

- link to notebook: Notebook
- link to utils: Utils
- link to netwok weights: Networks Weights

References

Roberta-base. https://huggingface.co/roberta-base.

Fang Chen, Gourav Datta, Souvik Kundu, and Peter A Beerel. 2023. Self-attentive pooling for efficient deep learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3974–3983.

Sumire Honda and Sebastian Wilharm. 2023. Noam Chomsky at SemEval-2023 task 4: Hierarchical similarity-aware model for human value detection. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1359–1364, Toronto, Canada. Association for Computational Linguistics.

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.

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 The 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 256–261.

Daniel Schroter, Daryna Dementieva, and Georg Groh. 2023. Adam-smith at semeval-2023 task 4: Discovering human values in arguments with ensembles of transformer-based models. *arXiv preprint arXiv:2305.08625*.