Human Value Identification

Marco Panarelli, Jacopo Meglioraldi, Mihail Stamenov
Master's Degree in Artificial Intelligence, University of Bologna
{ marco.panarelli, jacopo.meglioraldi, mihail.stamenov }@studio.unibo.it

Abstract

This report is a summary of our work on the human value classification problem, according to the assignment request. Our objective was to implement a BERT based model to solve a multi-label classification task about detecting human values behind textual arguments. The dataset used is the one proposed by the "Human Value Detection 2023" challenge, in which a set of triplet (argument, premise and stance) were manually labelled to categories from social science literature. The model presented achieved good results, scoring 0.746 of average-macro-F1-score for our best model. From results we see that the by including the premise in our model, the performances increase significantly.

1 Introduction

Human value detection inside text is a modern open problem, although in principle is a simple text classification. What makes it difficult is that human values are based on semantic information while lexical and structural information have less discriminatory power. Different approaches are described in literature to perform text classification. The most widely known are: rule based models, statistical models, Bayesan models and deep learning models. Deep network models are the nowadays standards for NLP application with the developing of large language models. They are able to learn to represent text in a more meaningful space and extract features from it. In particular Transformers have been proved effective into learning long distance dependency of words inside a text. The model proposed is based on BERT (2018), an "encoder-only" transformer architecture, and its variant. The model has been fine-tuned over the training set proposed by the "Human Value Detection 2023" challenge (2023) with the addition of a linear classifier aiming to classify the human values behind textual arguments. From the data, higher order values (second level) have been extracted according to the literature presented in the challenge (2023) and used as target for our multi-label classification task. Three models have been implemented by considering the three different part of each argument: conclusion, premise and stance. Model C is the model considering only the conclusion part, Model CP considers both conclusion and premise, Model CPS take in account all three of them. The three model were compared between each other and against a baseline, nominally a uniform and a majority classifier. For consistency check, the models have been trained three times with different seeds and the best model was chosen among them. The model is able to correctly classify the human value behind arguments, scoring an average of 0.746 macro F1-score for our best model. From results we noticed that Model CPS perform worse than Model CP, in a counter-intuitive way, probably due to distribution of data. "Self-transcendence" has been found to be the most easiest class to classify for all three models, while "Openness to change" and "Self-enhancement" are more easily misclassified.

2 System description

We proposed three model with slightly different architecture in order to account differently the three part of each argument, as the assignment required.

- Model C takes in account only the conclusion part of the argument, feeding it trough a BERT encoder and using its pooled output to classify the text through a linear classifier.
- For Model CP, both the premise and the conclusion were encoded through the BERT network and the concatenation of the two pooled output was fed to the classifier.
- Model CPS architecture is the same of Model CP, with the addition of 0-1 variable for stance in the concatenation step before the classification step.

A dropout layer is present between the BERT encoder and the classifier to regularize the training. The models are built using the HuggingFace's tranformers module, expanded when necessary for custom behaviour, for example in the CPS model. The three models have been tested over the dataset proposed by the challenge which contains the conclusion, premise and stance for each argument, with the manually annotated labels of first level human values. A first step of preprocessing was done by aggregating the first level human values into higher order values according to the assignment request. A uniform classifier and a majority classifier are used as a baseline to compare our models.

3 Experimental setup and results

The dataset used is the one proposed by the "Human Value Detection 2023" challenge (2023) which consist of 5394 entries as training set, 1897 entries for validation and 1577 for testing, with the distribution presented in Figure 1. The BERT model used is RoBERTa-base (2019) model in the transformers library and the full architectures were expressed with the AutoModel class family from the library. RoBERTa pretrained weights were fine-tuned over the task during the training. The training scheduler used is "cosine with restart" with $lr = 2e-5$ and $weightdecay = 0.01$. The train was performed by the Trainer class, as suggested in the documentation, with batch-size of 16 for 10 epochs. The dropout probability of the default model is 0.1, experiments with values 0.2 and 0.3 were performed to try to enhance the generalization capabilities of the models. The loss used is BCE with logits loss and the performances of the model were track with accuracy and macro F1-score metrics, the latter is the one used to choose the best model. Class weights have been added in order to compensate infrequent positive distribution of classes, each weight is the ratio between negative and positive examples. For consistency check, the three models were trained three times with different seeds ([42, 55, 666]) and tested to be similar, as presented in the Table 1. The best one of each model was then used for evaluation over the test set.

4 Discussion

The results show that the best performing model is Model CPS, scoring 0.735 of F1-score on validation set, safely above our baselines scoring 0.432 and 0.526. From table 2 we can see an important skew in how our models perform over each class, in particular Model C classify almost all entries to Conservation and Self-transcendence, as a majority classifier would do. Our best model, Model CPS, instead can correctly classify a larger number of entries with more balance between the classes. Setting dropout probability to 0.3 in conjunction with the cosine with restarts scheduler resulted in the best performances over all the models, confirming that greater generalization helps in tasks where the class distributions are very skewed, such in this case.

5 Conclusion

Human value detection from text is task that is still an open problem. Our model perform well over the dataset, scoring 0.735 F1-score with more resilient performances over less frequent classes. BERT and other transformers models have been proved to work quite well in extracting semantic meaning from text. One limitation found is that our models are prone to over-fitting even with the dropout regularization.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Nailia Mirzakhmedova, Johannes Kiesel, Milad Alshomary, Maximilian Heinrich, Nicolas Handke, Xiaoni Cai, Barriere Valentin, Doratossadat Dastgheib, Omid Ghahroodi, Mohammad Ali Sadraei, et al. 2023. The touch\'e23-valueeval dataset for identifying human values behind arguments. *arXiv preprint arXiv:2301.13771*.

Touché. 2023. Semeval 2023 task 4. valueeval: Identification of human values behind arguments.

Model	score	Seed			Avg
		42	55	666	
Majority	**Acc.**		0.6736		0.6736
	F1		0.4322		0.4322
Uniform	**Acc.**	0.5034	0.503	0.5019	0.5027
	F1	0.5267	0.5267	0.5256	0.5266
Model C	**Acc.**	**0.6235**	0.6005	0.6143	0.6127
	F1	0.6943	**0.7224**	0.7135	0.7100
Model CP	**Acc.**	**0.670**	0.5953	0.6648	0.6433
	F1	0.7294	0.7296	**0.7328**	0.7306
Model CPS	**Acc.**	**0.704**	0.6863	0.6946	0.6949
	F1	0.7295	**0.7351**	0.7328	**0.7324**

Table 1: Grid search training results on the validation set

		Predicted values [F T]					
		Model C		Model CP		Model CPS	
True values [F T]	**Openness to change**	0.067	0.564	0.322	0.309	0.348	0.283
		0.05	0.318	0.107	0.260	0.102	0.265
	Self enhancement	0.004	0.528	0.122	0.410	0.251	0.282
		0.001	0.465	0.061	0.405	0.128	0.338
	Conservation	0.0	0.247	0.006	0.241	0.010	0.237
		0.0	0.752	0.004	0.747	0.011	0.741
	Self transcendence	0.001	0.204	0.032	0.164	0.002	0.203
		0.	0.794	0.001	0.793	0.005	0.789

Table 2: Normalized Confusion Matrix for best models on the validation set. For each nested matrix elements in position (0,0) are the count of true negatives, (0,1) are false positives, (1,0) are false negatives and (1,1) are true positives.

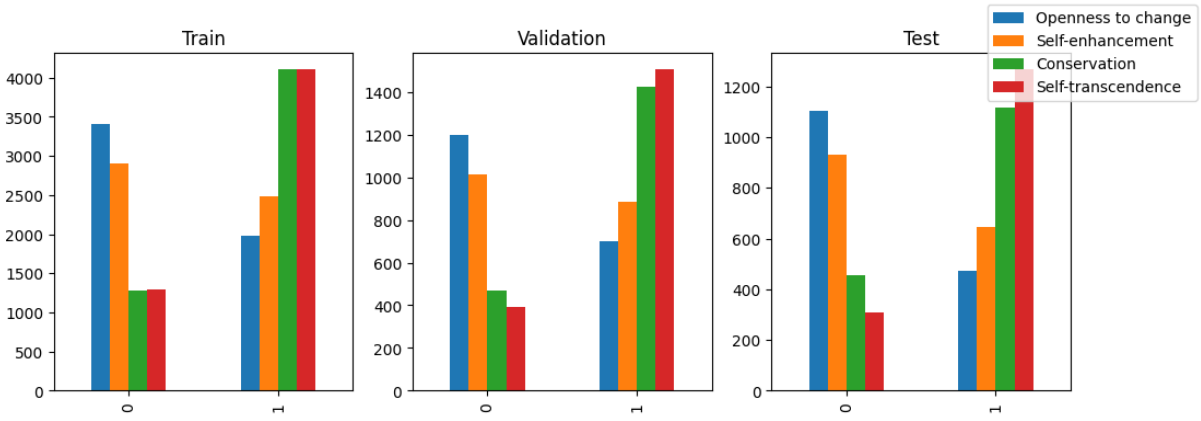

Figure 1: Distribution of class in the dataset