Assignment 2

Alessandro Folloni, Gabriele Fossi, Daniele Napolitano and **Marco Solimé**
Master's Degree in Artificial Intelligence, University of Bologna
{ alessandro.folloni2, gabriele.fossi, daniele.napolitano4, marco.solime }@studio.unibo.it

Abstract

This paper tackles the Human Value Detection challenge as a multi-label classification problem. We devised and evaluated five different models: two simple baselines and three based on the BERT architecture. The latter surpassed the baselines, while exhibiting higher performance for the predominant classes in the dataset.

1 Introduction

This work deals with the Human Value Detection challenge (Habernal et al., 2023), the process of identifying the human values expressed or implied by a textual argument, which consists of a premise, a conclusion, and its stance.
This challenge was approached as a multi-label classification problem, where each textual argument can be assigned to one or more of the four value classes of the third level of Shwartz's Value continuum (Schwartz et al., 2012). This paper evaluates five different models: two baselines that use random uniform and majority classifiers, and three models that use BERT (Devlin et al., 2018) with different input features. The first only uses the premise of the argument, the second one uses both the premise and the conclusion, and the third one also incorporates the stance of the argument. We measure the performance of each model using the binary F1 score for each class.
Alternative approaches not tried in our experiments, based on SVM, have empirically been proven to be worse than BERT ones (Kiesel et al., 2022). As expected, the BERT models outperformed the baselines, and the ones with more inputs (CP and CPS) performed better than the one that only relies on Conlusion.

2 System description

The architecture and tokenization process play pivotal roles in shaping the BERT models' performance. The tokenization mechanism segments the input text into tokens, each representing a specific subword or word piece. This process holds significance in capturing nuanced value expressions or complex argument structures. A detailed breakdown of the tokenization process, augmented with visual representations showcasing the model's architecture from input tokens to classification output, would aid in comprehending the information flow and understanding the model's functioning.

The two baseline models, implemented with Sklearn's DummyClassifier class, provide simple yet essential comparisons. The BERT models were defined using HuggingFace's Transformer library, thus the amount of code to define the architectures was minimal. By setting the problem_type parameter to *multi_label_classification*, the model outputs four binary prediction values. The definition of the three models is essentially the same, since the only thing that change is the input: for the C model, only the Context tokens are fed as input, then for the CP one, Context and Premise are concatenated, separated by the **[SEP]** special token. Finally, the same applies for the CPS model, where Stance is concatenated as text and not in its numerical format.
Both BERT models and the Tokenizer were loaded from *bert-base-uncased*'s model card, which has 110M parameters, case-insensitive, and only trained on English corpora.
Tokens are padded with a 128 max length for the C model, and 256 for CP and CPS.

3 Experimental setup and results

HuggingFace's Trainer API library served as the cornerstone for training our three distinct BERT models. Leveraging this robust API, we fine-tuned

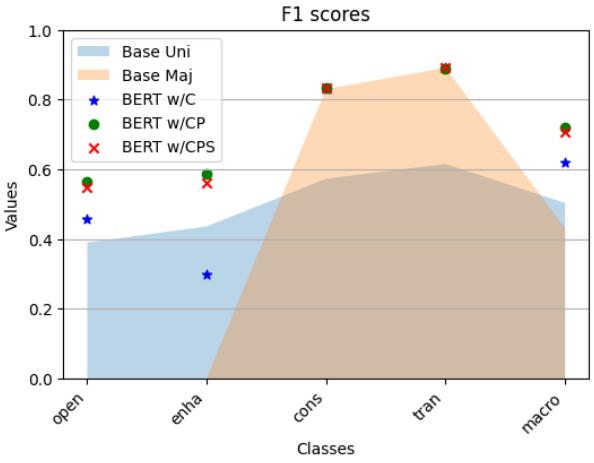

Figure 1: F1 score graph for all the models, for each class. The last bar represents macro-F1.

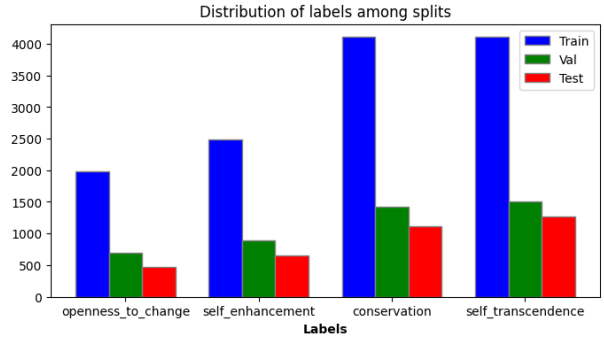

Figure 2: Graph of the dataset distribution of each category, for all the splits.

the models' hyperparameters using the versatile TrainingArgument class. The aim was to tailor the base BERT model specifically to the intricacies of the challenge's corpus, ensuring optimal performance in detecting nuanced human values within textual arguments.

To optimize the models' learning process, a learning rate of $2e^{-5}$ was employed, accompanied by a weight decay set at 0.01. These chosen hyperparameters were selected based on empirical observations and prior fine-tuning experiments, aiming to strike a balance between model convergence and generalization capacity.

The training regimen was designed, encompassing a 3-epoch training cycle with a conservative batch size of 8. This cautious approach aimed to ensure model stability and prevent overfitting while allowing the model to grasp intricate patterns within the textual arguments.

The resultant F1 scores for each model across individual quality classes are depicted in Figure 1. Notably, the bar graph captures the nuanced performance of each model, elucidating their strengths and weaknesses in discerning specific value expressions within textual arguments.

4 Discussion

Regarding CP and CPS BERT models, we do not observe a clear difference as the model performances are very close to each other.
Regarding the Baselines, all the models outperform the Uniform reference when it comes to the macro F1 score. A similar behaviour is observed in each individual category. A particularly strong Majority Baseline is evident especially in the conservation and self_transcendence categories, as all the BERT models equal such performances. The random uniform classifier performs worse than all the others, except for *self_enhancement*, where BERT w/C has a lower score, also much lower than CP and CPS. To show a wrongly classified example, an argument *against* the abolition of COVID digital pass says:

> "You owe the fact that you can drink your coffee in public to the people who got vaccinated and thus made the virus a little more manageable."

In this case, the BERT w/CPS model predicted the *openness_to_change* value as true, while it was not for the true label. The argument in fact is *against* the abolition of the green pass, so it's in favour of not changing it as it is.

5 Conclusion

In this work we trained, validated and compared five different models to solve the Human Value Detection challenge. At the end of this process we noticed the following things: the information contribute of Stance is negligible, as the CP and CPS models behave almost identically. Still, the F1 of CPS is always greater or equal than the one of CP. Imbalances in the dataset distribution (as seen in Figure 2) are clearly reflected in the results: the first two categories (*openness_to_change* and *self_enhancement*) have a lower support compared to the others; this fact may suggest the models struggle to learn such concepts due to the lack of samples in the train dataset.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. Cite arxiv:1810.04805Comment: 13 pages.

Ivan Habernal, Henning Wachsmuth, Martin Potthast, and Benno Stein. 2023. Human value detection challenge. https://touche.webis.de/semeval23/touche23-web/.

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.

Shalom Schwartz, Jan Cieciuch, Michele Vecchione, Eldad Davidov, Ronald Fischer, Constanze Beierlein, Alice Ramos, Markku Verkasalo, Jan-Erik Lönnqvist, Kursad Demirutku, Ozlem dirilen gumus, and Mark Konty. 2012. Refining the theory of basic individual values. *Journal of Personality and Social Psychology*, 103:663–88.