Assignment 2

Antonio Lopez, Edoardo Merli, Matteo Vannucchi and **Alessandra Blasioli**
Master's Degree in Artificial Intelligence, University of Bologna
{ antonio.lopez2, edoardo.merli, matteo.vannucchi, alessandra.blasioli }@studio.unibo.it

Abstract

This study compares different model architectures for human values detection in textual arguments. Specifically, we focused on predicting higher-order values using BERT-based architectures with a classification head on top. Using all the components of textual arguments led to better results, reaching a macro F1 score of 0.74 on the validation set and 0.72 on the test set.

1 Introduction

The Human Value Detection 2023 challenge[1] presents the problem of identifying and classifying human values in text corpora. The organizers, along with the challenge, provided a dataset composed of a series of textual augmentations. These are tuples of "Premise", "Conclusion" and "Stance". Each instance is assigned to a hierarchical multi-label. In this report, instead of using the full labels, we focused only on level 3 categories.

The human value detection problem has been tackled by (Kiesel et al., 2022) employing a BERT model and an SVM classifier. They observed that the BERT-based model performed worse on level 3 categories compared to other categories.

In our approach, we adopt a standard procedure in NLP: A BERT-based model is used as a backbone to extract features from text argumentation; subsequently, these extracted features are fed into a multi-head classification MLP for classification. In this report, we propose three different architectures, progressively providing more information to the models. We will demonstrate that the model exhibits limited capability in classification when only provided with the "Conclusion". On the other hand, supplying the "Premise" significantly enhances the model's ability to classify effectively.

[1] Challenge Website

2 System description

We propose three different models, each comprising two main components:

- Backbone: a BERT-based model functioning as the embedding layer. We use the [CLS] token from the last layer as the encoding. The "Premise" and "Conclusion" are embedded separately, and are concatenated before the classification head. For our experiments, we tried with both BERT[2] (Devlin et al., 2019) and RoBERTa[3] (Liu et al., 2019) models.

- Classification head: given the multi-label nature of the problem, we utilized a single classification head with four outputs. This component is made of a feed-forward network with one hidden layer and a sigmoid activation in the middle. The input size depends on the model and the backbone chosen.

This architecture is common to the three models that we will compare:

- **BERT w/ C**: define a BERT-based classifier that receives an argument "Conclusion" (C) as input.

- **BERT w/ CP**: add argument "Premise" (P) as an additional input.

- **BERT w/ CPS**: also add "Stance" (S), represented by a single binary digit, as an additional input.

The model backbone is always frozen to ensure faster training. As baseline models, we used a random and a majority classifier.

[2] huggingface.co/bert-base-uncased
[3] huggingface.co/roberta-base

3 Experimental setup and results

All three models were trained using the same recipe and the same three seeds (6, 90, 157). The following hyperparameters were used:

- Batch size: 32
- Optimizer: Adam
- Learning rate: 10^{-3}, after trying several values this one was the best for having a good convergence speed.
- Scheduler: ReduceLROnPlateau, the addition of a scheduler improved the training and the results, allowing us to set a higher initial learning rate.
- Loss: multilabel binary cross entropy.
- Regularization: a dropout layer (0.1 dropout rate) in the CLF and early stopping on the validation loss.

At every epoch of training and validation the **per-category** and **macro** F1-score were calculated. We tried different combinations of hyperparameters and backbone: it turned out that *bert-base-uncased* achieved at least two points more than *roberta-base*, and hence the results displayed in table 1 use the first as the backbone.

Model	Validation F1	Test F1 macro
Random	0.53	0.50
Majority	0.43	0.43
BERT w/ C	0.58	0.55
BERT w/ CP	0.71	0.70
BERT w/ CPS	**0.74**	**0.72**

Table 1: Mean model F1 scores across the three seeds (6, 90, 157) in test and validation.

Plots of the loss and macro F1 metric during training can be found in the notebook, together with tables for validation and test set per-category F1 scores.

4 Discussion

The best model obtained is BERT w/ CPS, with an average F1 score of 0.72 on the test set. It is clear that only the "Conclusion" is not enough to achieve good results in classification, as the addition of the premise has drastically increased the performance (+0.13 validation F1, +0.15 test F1). It is also interesting to notice that BERT w/ C performs worse than the random classifier for the label "Openness to change".

The addition of the "Stance" has also improved the results, even though not as drastically as the "Conclusion" did, but it is noticeable how a single binary flag could lead to a significant improvement.

The backbone chosen was also important: surprisingly the standard version of BERT achieved the same results if not better than the improved RoBERTa.

An example where the model performs poorly is the sequent, where it predicts exactly the two opposite classes of the ground truth:
Premise: "There is no reason to force those who are fit for combat to leave the armed forces as long as they voluntarily want to be in them."
Conclusion: "We should prohibit women in combat."
Stance: against
Labels: 'Openness to change', 'Self-enhancement'
Predicted: 'Conservation', 'Self-transcendence'

5 Conclusion

In this work, we tackled the human-value detection from text-argumentation presenting three models. Each of them adds a piece of information starting from the "Conclusion" (C) to the "Premise" (P) and "Stance" (S). The best results are achieved by **BERT w/ CPS** with an average F1 score of **0.72**. According to our expectation, the models with more information are performing better. We can notice that only the "Conclusion" is not sufficient to solve the problem, since there is a big gap in the F1 score between BERT w/ C and the other models. We expected a relevant difference in performance between RoBerta and BERT as backbones in favor of the former, but surprisingly the performance was quite the same. It is also noticeable how a simple feed-forward classification head, without any fine-tuning of BERT, achieved good results.

We are also convinced that these results could be improved by:

- Fine-tuning BERT or using a more powerful backbone, for example, a sentence encoder like the one presented in (Cer et al., 2018).
- Using all the tokens provided by BERT without limiting to [CLS].

6 Links to external resources

- GitHub repository: repository
- Models weights: drive

References

Daniel Cer, Yinfei Yang, Sheng yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

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.

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.