Assignment 2

Pelinsu Acar, Rubin Carkaxhia, Calin Diaconu and **Ruud Johannes Wilhelmus Korsten**
Master's Degree in Artificial Intelligence, University of Bologna
{ pelinsu.acar, rubin.carkaxhia, calin.diaconu, ruudjohannes.korsten }@studio.unibo.it

Abstract

The current assignment has the purpose of creating and comparing solutions consisting of BERT-based (Devlin et al., 2018) models that determine the human value category in a textual argument. To achieve this, three different configurations were verified. The best performing one includes the conclusion and the premise as input. It achieves an F1-score of 0.65 and a mean accuracy of 0.70.

1 Introduction

This project is based on the Human Value Detection 2023 competition (Kiesel et al., 2022). The task was to detect a subset of the 20 value categories, that were compiled from social science literature, especially from the work of Schwartz et al. (Schwartz, 1994) (Schwartz et al., 2012). The arguments were given as premise and conclusion texts, and binary stance of the premise to the conclusion, with the 0-1 values being associated to the "against"-"in favor" meanings.

Here, all three types of arguments are used, but they are used to split them only into the level 3 categories, which means mapping the 20 level 2 classes to 4 level 3 ones. Some level 2 categories are associated to multiple level 3 instances.

A training dataset (Mirzakhmedova et al., 2023) was provided for the competition, containing the arguments and the categories. It was split in 5,393 arguments for training, 1,896 for validation, and 1,576 for testing. Other validation and test sets were also made available, extracted from the Zhihu website, the Nahj al-Balagha book, Coronavirus-related New York Times articles. However, all these former ones were not used in the development of the current solution.

The competition runs were evaluated using the F1-score, the Precision, and the Recall, both averaged, and per-category. The procedure was similar to the competition one, but within this assignment, the F1-score and the Accuracy were used. The best scoring submission achieved an F1-score of 0.56 on average, while the current solution reaches 0.65, with a 0.70 average Accuracy. However, it should be noted that the competition score is based on all 20 level 2 categories.

2 System description

For this task, we employed two baselines and three BERT-based classifiers:

- Baseline1: Random uniform classifier
- Baseline2: Majority classifier
- BERT C: Takes conclusion as input
- BERT CP: Includes argument premise as additional input to the previous one
- BERT CPS: Incorporates stance as additional input to the previous one

All BERT models are loaded through the Hugging Face Transformers library, using PyTorch. The corresponding tokenizer provided by Hugging Face was used for each model.

Level 2 annotations are merged into level 3 categories due to consideration of only these latter classes. We introduce a numerical stance column, and merge the different data frames based on Argument ID. Labels are concatenated to the conclusion, premise, and numerical stance columns in the data frames, to create the datasets. Subsequently, we tokenize the conclusion column, the conclusion, and premise columns together, and the combination of conclusion, premise, and numerical stance. This generates the input datasets for BERT C, BERT CP, and BERT CPS classifiers, respectively.

3 Experimental setup and results

The experiment involves training the two baseline models and the three BERT based models mentioned above, set up in a Python environment. Setting up the baseline models consists only in passing the labels, using a DummyClassifier from the scikit-learn software package (Pedregosa et al., 2011).

Training the BERT models is done using PyTorch (Paszke et al., 2019). All three BERT models use the same hyper-parameters: 20 epochs, 2e-5 learning rate, a batch size of 8, and 0.01 weight decay. They are all based on the bert-base-uncased pretrained model, since other compatible ones (like deberta-base-mnli (He et al., 2021), MiniLM-L12-H384-uncased (Wang et al., 2020), and bert-tiny (Bhargava et al., 2021) (Turc et al., 2019)) yielded inferior results. Three different random seeds (42, 123, and 999) are used to ensure invariability in the training process. For each seed, the random number generators for NumPy and PyTorch are seeded to maintain reproducibility.

After training, the model's predictions are evaluated on the test dataset. For each of the four classes, the Area Under the Receiver Operating Characteristic (AUC-ROC) score and optimal threshold are calculated. The test predictions are transformed using these obtained thresholds. The F1 scores and accuracy are also computed. They are reported in Table 1.

4 Discussion

4.1 Quantitative Analysis

Our evaluation focuses on the label-wise F1 score and its mean overall labels (macro-average), as well as the accuracy. According to the mean values over level 3 labels, BERT CP gives the best F1 score and accuracy. Comparing with the baseline1 model, we can observe that the performance is improved by 21.8%, 37.9% and 36.7% for the Bert based models respectively. Note that Baseline 2 is especially strong for the accuracy, since by definition it makes label-wise random guessing according to the label frequency.

4.2 Error Analysis

To get a deeper understanding, we can analyze the metrics for each level 3 classes. BERT C performs best for the 4th class. One of the reasons may be the label inconsistency on duplicated conclusions for each class. Since BERT C only takes the conclusion as input, and we have different human values based on the stance and premise for the same conclusions in our dataset, it can be difficult for the model to differentiate these samples.

To verify this hypothesis, we check the percentage of majority labels on each class for these duplicated samples. The consistency is computed as an average of the majority class percentages for every class. For example, if for a conclusion, for class 1, 8 samples are marked as belonging to that class, and 2 are not, then the consistency of this situation will be 80%. These consistencies are averaged over the entire set.

The 4th class has 84.67% consistency, which is higher than the all other 3 classes. Both BERT CP, and BERT CPS give similar results, a higher performance on 'Openness to change' and 'Self-enhancement', and a relatively lower performance on 'Conservation' and 'Self-transcendence'. The confusion matrix shows that the high number of false negatives is responsible for this low performance on these two classes. Moreover, the Baseline 2 classifier performance indicates that we have imbalanced classes. Our accuracy is the same to our precision, meaning that it is dominating the overall accuracy measure – all the cases that belong to the third and fourth classes are classified as positive, since they are the majority class.

4.3 Future Developments

Considering that adding the stance to the input makes the performance worse, weighting it down may help the results.

The f1 score indicates a decent overall performance, but there is room for improvement, especially for classes with lower metrics. This could be done by addressing class imbalances, and further tuning the model could.

Further training could also be done on the other datasets that were provided on the competition website for validation and training.

5 Conclusion

In this assignment, we tested and compared different models that classify arguments into the level 3 categories of the Human Value Detection 2023 competition. We defined three BERT-based models that would take as input 1, 2, and 3 of the 3 available arguments: conclusion, premise, and stance. We discovered that the best class-average F1-score and Accuracy were yielded by the model taking as input the conclusion and the premise.

6 Links to external resources

The notebook can be found here: http://bit.ly/3vAeOxX .

	F1 Score					Accuracy				
Level 3 Class	**1**	**2**	**3**	**4**	**Average**	**1**	**2**	**3**	**4**	**Average**
Baseline 1	0,46	0,51	0,49	0,45	**0,48**	0,49	0,51	0,51	0,50	**0,50**
Baseline 2	0,41	0,37	0,42	0,45	**0,41**	0,70	0,59	0,71	0,80	**0,70**
BERT C	0,56	0,52	0,40	0,61	**0,52**	0,57	0,52	0,41	0,73	**0,56**
BERT CP	0,69	0,69	0,66	0,58	**0,66**	0,75	0,71	0,69	0,68	**0,71**
BERT CPS	0,69	0,69	0,63	0,59	**0,65**	0,75	0,69	0,70	0,68	**0,71**

Table 1: Per-class and average numerical results of all models, measuring the F1 score and the accuracy.

References

Prajjwal Bhargava, Aleksandr Drozd, and Anna Rogers. 2021. Generalization in nli: Ways (not) to go beyond simple heuristics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

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.

Nailia Mirzakhmedova, Johannes Kiesel, Milad Alshomary, Maximilian Heinrich, Nicolas Handke, Xiaoni Cai, Barriere Valentin, Doratossadat Dastgheib, Omid Ghahroodi, Mohammad Ali Sadraei, Ehsaneddin Asgari, Lea Kawaletz, Henning Wachsmuth, and Benno Stein. 2023. The touché23-valueeval dataset for identifying human values behind arguments.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

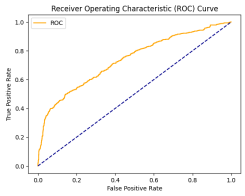
Shalom H Schwartz. 1994. Are there universal aspects in the structure and contents of human values? *Journal of social issues*, 50(4):19–45.

Shalom H Schwartz, Jan Cieciuch, Michele Vecchione, Eldad Davidov, Ronald Fischer, Constanze Beierlein, Alice Ramos, Markku Verkasalo, Jan-Erik Lönnqvist, Kursad Demirutku, et al. 2012. Refining the theory of basic individual values. *Journal of personality and social psychology*, 103(4):663.

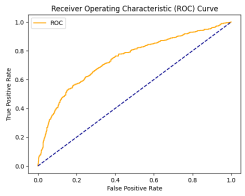
Iulia Turc, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Well-read students learn better: The impact of student initialization on knowledge distillation. *CoRR*, abs/1908.08962.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers.

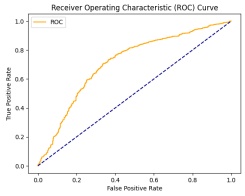
Appendix



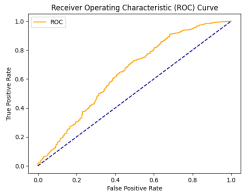
(a) Openness to change



(b) Self-enhancement



(c) Conservation



(d) Self-transcendence

Figure 1: The ROC curve for the best performing model, for the four level 3 categories.