Assignment 2

Giorgia Castelli, Alice Fratini and **Madalina Ionela Mone**
Master's Degree in Artificial Intelligence, University of Bologna
{ giorgia.castelli2, alice.fratini2, madalina.mone }@studio.unibo.it

Abstract

This report addresses the challenge of human value detection as a multi-label classification task within natural language arguments. We developed and assessed three BERT-based models, alongside two baseline classifiers: random uniform and majority classifiers. The BERT models were trained on varying combinations of text features. The results demonstrated that the BERT model utilizing all three features outperformed the baselines and other configurations, particularly in the Precision-Recall curves, while the configuration with only 2 features showed slightly better performances in macro F1 score and per-category evaluations, although displaying a high variability for the different seeds.

1 Introduction

Human Value Detection is a critical task in natural language processing, aiming to identify values embedded within textual arguments. This task is particularly important in fields like social science and ethics, where understanding human values is key to interpreting human behavior and decision-making. The introduction of deep learning, particularly transformer-based models like BERT (Bidirectional Encoder Representations from Transformers), has revolutionized this field by offering superior performance in understanding contextual relationships within text. BERT, pretrained on vast amounts of text, can capture subtle linguistic nuances and dependencies between labels, making it particularly effective for multi-label classification tasks such as human value detection. Our approach leverages the bert-base-uncased model, which is fine-tuned specifically for this task. The objective is to maximize both per-category and macro F1 scores, providing a robust framework for accurately detecting human values in text.

2 System description

2.1 Model Architecture

The architecture of our system is centered around the `bert-base-uncased` model, which contains 12 layers, 768 hidden units per layer, and a total of 110 million parameters. This model is known for its bidirectional context understanding, where each word is analyzed in relation to all others in a sentence, rather than just in sequence.

We experimented with three distinct model architectures:

- **Bert w/C:** Only the 'Conclusion' of the text was used as input.
- **Bert w/CP:** A combination of 'Conclusion' and 'Premise' was fed into the model.
- **Bert w/CPS:** The final model included 'Conclusion', 'Premise', and the stance encoded as numerical data.

These architectures were chosen to assess the impact of different input features on the model's performance.

2.2 Tokenization and Preprocessing

The tokenization process was handled by the `AutoTokenizer` from the HuggingFace Transformers library. Given the variability in text length, we padded sequences to a maximum length of 100 tokens, ensuring uniform input size for the model. This choice balances between maintaining sufficient context and managing computational efficiency.

2.3 Hyperparameter Tuning

The performance of BERT models is highly sensitive to hyperparameter choices:

- **Learning Rate:** We experimented with 1e-5 and 1e-7, settling on 1e-5 for its optimal balance between convergence speed and accuracy.

- **Batch Size:** A batch size of 16 was chosen after testing 16, 24, and 32, based on memory constraints and stability.
- **Epochs:** The model was trained for 5 epochs, a choice informed by experiments showing that more epochs led to overfitting.
- **Optimizer:** The Adam optimizer was employed due to its adaptability to different learning rates, crucial for handling the varying gradients in a multi-label classification task.
- **Loss Function:** BCEWithLogitsLoss was used, combining a sigmoid layer with binary cross-entropy loss, ideal for multi-label tasks.

3 Experimental setup and results

3.1 Setup

Our experiments focused on assessing the effectiveness of different model architectures in predicting human values. We evaluated the models using standard multi-label classification metrics, including per label F1 score and Macro avg F1 score, precision and recall.

3.2 Results

For what concerns the F1 score metrics, the CP configuration demonstrates a slightly better performance compared to the CPS, as shown in the following table:

Model | Seed	22	157	2024
Bert w/C	0.61	0.60	0.41
Bert w/CP	**0.71**	0.661	0.677
Bert w/CPS	0.662	0.675	0.672

Table 1: Performance comparison of models.

On the other hand, it is to be noted that F1 score performances for CP configuration show to be more variable, depending on the choice of seeds, than those for the CPS. Furthermore, when we examine the Precision-Recall Curves, illustrated in the accompanying figure, it becomes evident that the CPS configuration outperforms CP in terms of precision and recall dynamics.

4 Discussion

The experiments underscore the effectiveness of Model 3, where the integration of additional features such as Premise' and Stance' significantly

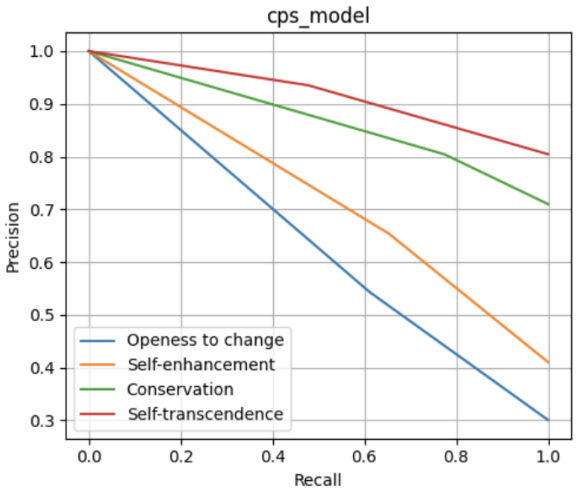

Figure 1: Precision-recall curve for Bert w/CPS.

enhanced the model's ability to detect human values. While the CPS configuration outperformed the baselines and other setups, particularly in the Precision-Recall curves, the CP configuration exhibited slightly better performance in terms of macro F1 score and per-category evaluations, although displaying a high variability for the different seeds. This suggests that although CPS provides superior precision and recall dynamics, as well as being a more stable configuration with regards to randomness, CP offers a slight edge in overall classification accuracy as measured by F1 score.

The choice of hyperparameters also played a critical role in the performance of the models: a learning rate of 1e-5 allowed the model to converge more effectively, while the use of the Adam optimizer ensured efficient handling of gradients.

5 Conclusion

In this work, we tackled human value detection using three BERT-based models and two baselines. The BERT model with all three input features excelled in precision, recall and stability, while the two-feature model slightly outperformed it in macro F1 scores, revealing a trade-off between metrics. While these findings mostly aligned with expectations, the lower F1 score in the CPS model might be due to how stance information was integrated, suggesting the model may not process stance effectively. Future work could focus on refining how stance content is represented and exploring additional features to enhance performance.