NLP Assignment 2

Andrea Zecca, Samuele Marro and **Stefano Colamonaco**
Master's Degree in Artificial Intelligence, University of Bologna
{ andrea.zecca3, samuele.marro, stefano.colamonaco}@studio.unibo.it

Abstract

In this document, we detail our development and testing of a RoBERTa-based text classifier for a relaxed version of the TOUCHÉ Human Value Detection 2023 challenge. Using different combinations of BERT models and finetuned classification heads, we achieve a remarkable F1 score of 0.727, demonstrating that this version of the task is indeed simpler. We also study the effect of several techniques as class reweighting, single-headed classification and base model finetuning.

1 Introduction

Identifying the values upon which an argumentation draws represents a highly challenging task due to both the breadth of argumentation techniques and the required common-sense knowledge to understand the implied context. The TOUCHÉ Human Value Detection 2023 dataset represents a high-quality benchmark for such a skill. Previous work, which focused on using either SVMs or transformers to identify the fine-grained values that underpin an argumentation, have been found to be lacking, with macro F1 scores ranging from 0.5 to 0.6 (Schroter et al., 2023). We instead consider a relaxation of this problem, which focuses on identifying the general categories of the relevant values. This is possible thanks to the hierarchical structure of the labels in the dataset, since each so-called level-2 value (e.g. hedonism) is associated to one or more level-3 values (in the case of hedonism, "Openness to change" and "Self-enhancement"). This structure is described in depth in (Kiesel et al., 2022). By considering these categories, we can thus rely on a more high-level understanding of the context; additionally, reducing the number of labels from 20 to just 4 key categories greatly reduces the size of the classifiers needed to achieve a sufficient performance. From a practical point of view, we use the RoBERTa transformer to compute fixed-size vector embeddings of the arguments, which are then passed to four binary classification heads (one for each label). Several approaches have been tested and compared.

2 System description

Our tests were run on two baseline classifiers (namely Uniform and Majority) and three BERT-based architectures (BERTC, BERTCP and BERTCPS). Uniform is a simple classifier that outputs a uniform probability between 0 and 1 for each label. Similarly, Majority always outputs, for each label, the most common value in the training set (i.e. either 0 or 1). On the other hand, the BERT models use a BERT-based transformer (in our case RoBERTa) to compute an embedding of parts of the argument. Specifically, BERTC computes the embedding of the conclusion (e.g. "We should abolish trade tariffs"). BERTCP is an extension of BERTC that also computes the embedding of the premise (e.g. "Trade is the fundamental driver of growth...") and concatenates it to the embedding of the conclusion. Finally, BERTCPS also concatenates a one-hot encoded value that represents whether the premise is in support of or against the conclusion. The overall embedding is then fed to four different heads (one for each label) composed of two dense layers. Since RoBERTa outputs a vector of size [text length, embedding size], in order to convert it to a fixed-size embedding we compute the average vector over the text length. Despite the simplicity of this approach, we find that this technique leads to a highly informative embedding of the text. The output of the embedding layer is then fed to a classification head, composed of a FC layer, a Dropout layer and another FC one.

3 Experimental setup and results

The pipeline of our experiments is as follows. First, we perform a preprocessing step on the dataset,

	Validation F1	Test F1
Random	0.528	0.516
Majority	0.436	0.431
BERTC	0.659	0.598
BERTCP	**0.747**	0.725
BERTCPS	**0.747**	**0.727**

Table 1: Macro F1 scores on the validation and test sets.

merging level-2 annotations into level-3 values, as well as converting the stance label ("in favor of the conclusion" / "against the conclusion") into a one-hot encoding. Then, we initialize both the baselines and the BERT-based models. For BERTC, the heads have an input size of 784, while for BERTCP it is $784 * 2$ (to take into account the premise) and for BERTCPS $784 * 2 + 2$ (to take into account both the premise and the stance). After this phase, we train the models with three different seeds using the following hyperparameters: `initial_rl` = 10^{-2}, `weight_decay` = 10^{-2}, `lr_decay_factor` = 10^{-2}, `lr_decay_patience` = 2, `hidden_size` = 100 and `dropout` = 0.2. For each run, we store the model with the best validation loss and use it to compute the metrics. Finally, we average the results across all runs and compute the precision-recall curves and confusion matrices.

In the numerous tests that we carried out, we always trained the three models using the same setup (i.e. configuration, preprocessing pipeline, tokenizer and base model), with the goal of having an objective comparison of the results. For each of these tests, we also tried replacing the four independent classification heads with a single multi-output head. This experiment yielded comparable results.

4 Discussion

We observed that the performance of the model grows with the amount of information provided to it, with BERTCPS having the highest F1 score of 0.727 (see Table 1). Predictably, the scores on the test set are slightly below those on the validation set for all models.

Furthermore, we empirically found that training BERT together with the classification heads is very computationally intensive and does not particularly result in performance improvements. For this reason, we decided to only focus on training classification heads.

After the training process, for each architecture we chose the weights with the best macro F1 scores and performed an error analysis. During this phase, we noticed that all the BERT-based models performed well on the classes *Conservation* and *Self-Transcendence*, while the other two had a worse performance, potentially due to the fact that there is an imbalance between positive and negative samples. We tried to solve this problem by using a reweighted loss (with the weights derived from the label distribution in the training and validation sets), but this actually led to a slight reduction in performance.

To further support this theory, we plotted a comparison between the frequency of positive values and the F1-score obtained by the models (see Figure 1).

We also generated a detailed breakdown of precision, recall and F1-score on each level-3 category for every model, as well as computing the confusion matrices for every combination of classes and models. Finally, we plotted precision-recall curves for the most frequently misclassified class, the least frequently misclassified one and for all the predictions.

5 Conclusion

In this study, we developed and tested a RoBERTa-based text classifier for the TOUCHÉ Human Value Detection 2023 challenge, reducing the number of labels from 20 to just 4 key categories thanks to the hierarchical structure of the dataset. This enabled us to focus on a more high-level understanding of the context and to use a smaller classifier. We compared our RoBERTa-based models (BERTC, BERTCP, and BERTCPS) against two baseline classifiers (Uniform and Majority), showing that BERT-based models are effective for this task, with the best performing model being BERTCPS. We also compared the results with the ones obtained in Kiesel et al. (2022) showing that our approach using the RoBERTa architecture and independent classification heads led to a slightly higher F1 score. Despite some limitations due to the unbalanced distribution of classes, our models performed well on the Conservation and Self-Transcendence categories. Future work could explore further refinements to the model, such as incorporating additional context or addressing class imbalance issues, with the goal of enhancing performance and applicability across a broader range of NLP tasks.

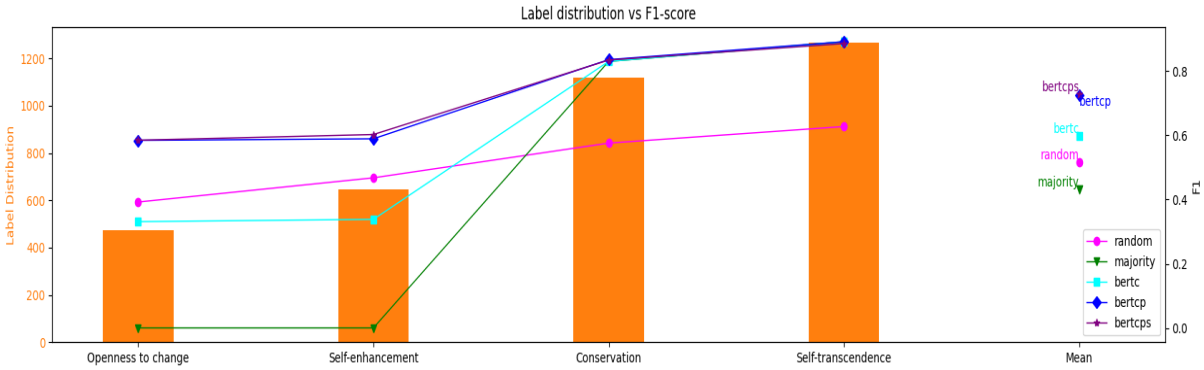

Figure 1: Label count distribution vs F1 score.

References

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.

Daniel Schroter, Daryna Dementieva, and Georg Groh. 2023. Adam-smith at semeval-2023 task 4: Discovering human values in arguments with ensembles of transformer-based models. *arXiv preprint arXiv:2305.08625*.