Assignment 2

Balzani Lorenzo, Deana Alessia and **Guizzetti Thomas**
Master's Degree in Artificial Intelligence, University of Bologna
{ lorenzo.balzani, alessia.deana, thomas.guizzetti }@studio.unibo.it

Abstract

In this project, we tackled the Human Value Detection challenge, a complex task centered around multi-label classification. We utilized three BERT-based classifiers, each designed to take 1 to 3 additional arguments as input, to ascertain the most effective model. Our observations revealed a direct correlation between the number of arguments and performance enhancement; classifiers with more arguments tended to perform better, indicating that additional context boosts effectiveness. However, we also recognized that augmenting the number of arguments necessitates corresponding adjustments in the model's architecture to fully capitalize on the improved performance. Finally, it appears that class imbalance significantly impacts the results. Further studies should be conducted to address and mitigate this issue.

1 Introduction

This study delves into the complex domain of multi-label classification, focusing on categorizing instances into multiple, often overlapping labels across four advanced Level 3 categories: "Openness to Change," "Self-Enhancement," "Conservation," and "Self-Transcendence." For the purposes of our assignment, we are using BERT-based models and the core of our investigation is a detailed evaluation of these models, under varying conditions using different seed values during training and compared to two different baselines.

In terms of results, he BERT w/CPS model consistently demonstrated superior performance, achieving the highest macro F1-scores, thus indicating its robustness in generalizing across diverse categories. Our findings highlighted a consistent challenge in accurately categorizing instances under "Openness to Change," suggesting a potential avenue for further improvement and a potential vulnerability to class imbalance, since the category was the least features in the dataset. Conversely, the "Self-Transcendence" category was more effectively recognized by all models, indicating distinct, identifiable features within this category.

2 System description

In this task, several models have been implemented, each fundamentally designed to accept a textual input and return classifications into distinct classes. These models are primarily aimed at processing and categorizing textual data based on predefined criteria.

BERT w/ C: This model defines a BERT-based classifier that takes the conclusion of an argument as its input. It consists of BERT with a classification multi-head.

BERT w/ CP: This variant extends the previous model by adding the argument's premise as an additional input, which we concatenate with a separator to feed as input to the model. It consists of BERT with a classification multi-head.

BERT w/ CPS: Further building on the previous models, this version incorporates the stance from the premise to the conclusion as an extra input, which again we concatenate with a separator to feed as input to the model. It consists of BERT with a classification multi-head.

BERT w/ CPS 2: this version consists of two BERT model that take as input respectively the premise and the conclusion. It then concatenates the embedded premise and the embedded conclusion with the stance and feeds it into the classification multi-head.

Furthermore, we use **Random Uniform Classifier** (which makes predictions by assigning a class randomly to each instance) and the **Majority Classifier** (which always predicts the most frequently occurring class label in the training dataset) as baselines for comparison.

3 Experimental setup and results

All models were systematically trained over a span of five epochs, employing three distinct seed values: 42, 64 and 512. The main hyper parameters for the training process were set for the Adam optimizer and the batch size, with the learning rate configured at 2e-5 and the batch size established at 8.

We evaluated our models using per-category binary F1-score and the average binary F1-score over all categories (macro F1-score). The results for each model are listed respectively in Table 1 on page 2 and Table 2 on page 2.

Table 1: Summary of Model Performance

Model	Train. Loss	Val. Loss	F1
BERT w/C	0.5157	0.6876	0.6673 ± 0.0080
BERT w/CP	0.1984	0.8748	0.7437 ± 0.0065
BERT w/CPS	0.0558	1.4313	0.7414 ± 0.0017
BERT w/CPS 2	0.3615	0.5574	0.7514 ± 0.0092

Table 2: F1-Score Comparison of Different Models

Model	Op. to chan.	Self-en.	Conv.	Self-tra.	Macro-F1
RC	0.3639	0.4630	0.7392	0.7991	0.5200
MC	0.0000	0.0000	0.7521	0.7943	0.4400
BERT-C	0.4695	0.7176	0.7752	0.8125	0.5700
BERT-CP	0.5802	0.6538	0.8042	0.8204	0.7400
BERT-CPS	0.5959	0.6641	0.8032	0.8207	0.7400
BERT-CPS2	0.6121	0.6860	0.8617	0.8675	0.7600

4 Discussion

The performance metrics of various BERT-based models are showcased in the tables, focusing on training and validation loss and the F1 score for model evaluation. The "BERT w/CPS 2" model stands out with a superior F1 score, suggesting a commendable balance between overfitting and generalizing abilities, corroborated by its lower validation loss and higher F1 score.

In Table 2, "BERT-CPS2" demonstrates enhanced performance across multiple evaluative categories, indicating the efficacy of the CPS2 configuration in these specific contexts. However, the Majority Classifier (MC) shows a zero F1 score, which is a baseline indicator that merely predicts the most frequent class and fails to provide discrimination between classes.

A critical look at class distribution reveals that "Conversation" and "Self-transcendence" dominate the training and validation datasets, while "Openness to change" and "Self-enhancement" are less represented. The test dataset follows a similar pattern. This imbalance is significant as it can lead to inflated F1 scores for models that may be biased towards predicting these majority classes more accurately.

Although the "BERT w/CPS 2" model appears to outperform others, the underlying class distribution must be considered to ensure the model's true predictive strength is not overestimated due to dataset imbalances. The promise shown by "BERT w/CPS 2" in performance metrics necessitates a nuanced interpretation to confirm its robustness across varied class distributions.

5 Conclusion

In our analysis of BERT-based models across different seeds and configurations, we observed notable variations in performance. The BERT w/CPS2 model consistently outperformed others, achieving the highest macro F1-scores, suggesting that certain architectural modifications can enhance model generalization. In particular, the more arguments the models take as input the better they seem to perform based on our metrics.

Finally, it is noteworthy that there exists a significant correlation between a category's F1-score and its frequency within the dataset across all models. This highlights the potential impact of class imbalance on the results, emphasizing the need to interpret the findings with care.

6 Links to external resources

References