

Human Value Detection: Multi-Label classification bases in a pretrained Bert Model

Miguel Eduardo Correa Gonzalez   
ID: 987175

Contents

[1. Introduction 2](#_Toc158744283)

[2. Data and model 2](#_Toc158744284)

[2.1. Data 2](#_Toc158744285)

[2.2. Model 3](#_Toc158744286)

[2.3. Data preparation 3](#_Toc158744287)

[3. Methodology 4](#_Toc158744288)

[3.1. Data acquisition 4](#_Toc158744289)

[3.2. Model definition. 4](#_Toc158744290)

[3.3. Hyperparameter optimization 5](#_Toc158744291)

[3.4. Model Training 5](#_Toc158744292)

[4. Results 6](#_Toc158744293)

[5. Bibliography 8](#_Toc158744294)

1. Introduction

The quest to understand human values and beliefs through textual data is a formidable challenge in natural language processing (NLP). This endeavor, known as Human Value Detection, explores nuanced and often implicit contextual cues within text to uncover the complex tapestry of people's convictions and principles. The task's inherent complexity requires advanced computational strategies to navigate the subtleties of human language. Recognizing the importance and difficulty of this task, the NLP community has embraced SemEval Task 4, aimed at benchmarking and enhancing computational models' ability to identify a predefined set of 20 value categories within textual arguments. This project employs a pre-trained BERT model, specifically its uncased variant, as a cornerstone approach to address Human Value Detection. Chosen for its deep contextual understanding and proven effectiveness in various NLP tasks, our project leverages BERT's architecture to discern implicit meanings and sentiments in text, thus tackling the key challenges of SemEval Task 4. By fine-tuning this model on a task-specific dataset, we aim to accurately predict the presence of the 20 value categories, represented as unique combinations of premise, stance, and conclusion, thereby achieving high accuracy in multi-label classification and enriching our understanding of the patterns and correlations between human values and textual expressions.

1. Data and model  
   1. Data

The data used for this challenge was provided by the challenge “SemEval 2023 Task 4. ValueEval: Identification of Human Values behind Arguments”.[[1]](#footnote-2)

This task consists in classifying a given text between 20 categories of human values compiled from social science literature. The data is provided as tab-separated values files with one header line. For each dataset (training, validation and test) there are 2 files, the first of them contain one argument per line: its unique argument ID, the conclusion, the premise' stance towards the conclusion, and the premise itself. The second one also contains one argument per line: its unique argument ID and one column for each of the 20 value categories where in each of these columns, (1) meaning that the argument belongs to the category and (0) it does not.

* 1. Model

In selecting our model, I leaned towards the Pretrained Base-Uncased BERT [[2]](#footnote-3)Model offered by Hugging Face's Transformers. My decision was influenced by BERT's innovative architecture, which integrates 12 layers, 768 hidden units, and 12 self-attention heads, embodying a robust framework for understanding language subtleties through its extensive network of 110 million parameters. BERT's foundational training on vast datasets like Books Corpus and English Wikipedia employs Masked Language Modeling (MLM) and Next Sentence Prediction (NSP), methodologies pivotal for enhancing its linguistic comprehension. MLM challenges BERT to predict hidden words within sentences, fostering a contextual grasp superior to traditional models. NSP further sharpens its acuity by assessing sentence sequence logic, a crucial element for textual relation understanding.

This dual-pretraining regimen equips BERT with an adeptness in recognizing complex language patterns, establishing it as an ideal backbone for our multi-label classification endeavor. My approach aimed to utilize BERT's pre-acquired linguistic intelligence, fine-tuning it with minimal adjustments to efficiently discern the nuanced depiction of human values within text. This method epitomizes transfer learning, where BERT's richly learned representations are meticulously adapted to our specialized task, presenting an advanced instrument for unraveling the intricate dynamics between language and human values.

* 1. Data preparation

For the preparation of the data, I took both files for each dataset then I combined it in a unique data frame using pandas dataframe from python. The columns with the premise, stated and conclusion were combined in just one column, then this new column is tokenized using the module BertTokenizer from the Transformers library. Then this tokenized array is put together with the labels into a tensor dataset which will be used to feed the model during the tuning phase.

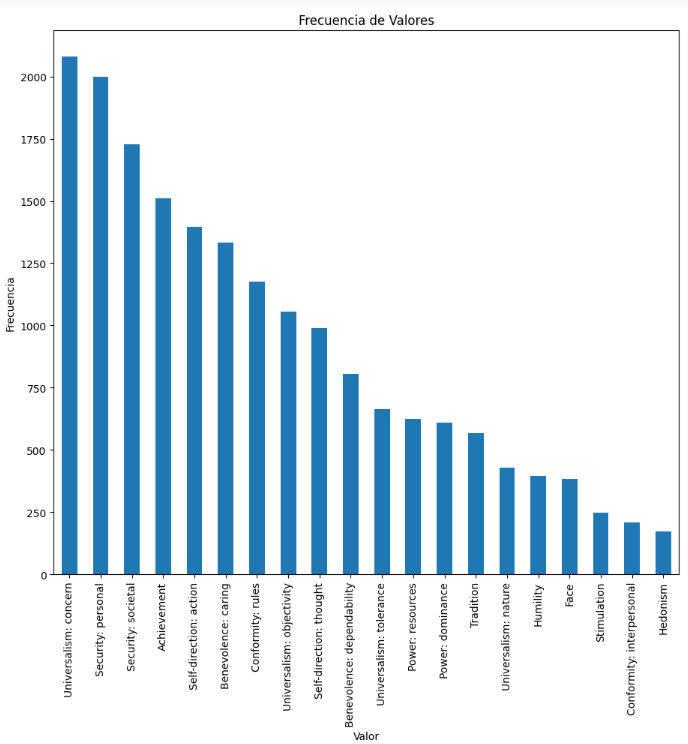
The data and the model have been created in Pytorch, so the dataset is created with the module tensor dataset and then the data loaders used to manage the data charged into the memory for each epoch of the optimization, training and test batch were created with the module Data loader.

1. Methodology

This project was developed following some steps to improve the chances to get the best possible result, with that purpose, I create the next methodology to optimize the getting result.

* 1. Data acquisition

Here, the data used in this project was downloaded from the original source and loaded into google drive to handle the connection with google Collab in an easy way because the code was written there. As soon as the data was updated in the Collab, I plot a graph of the frequencies for each label to analyze the distribution of the data, here I can see that the labels are not distributed uninformedly. However, they appear to be distributed in a similar way in the three different dataset where there are a clear dominance of some of the values and other less frequents.

A graph of blue and white lines

Description automatically generatedA graph of different colored lines

Description automatically generated with medium confidence

Figure 1: Frequency of label in the datasets.

* 1. Model definition.

The model implemented through the library Transformers allows to personalize the last linear layer of the model to adjust the output according to the specifications of the project, in this case I need a final linear output where for each text return a vector with 20 values associated each one to the human values, each value represents the probability of the presence of the value in the text.

The methodology is like a model of binary classification but in this case for each human value there is a binary classification independent of the other values even for the same text.

* 1. Hyperparameter optimization

In this step, I focused on optimizing the model's performance through a grid search, evaluating two crucial hyperparameters: batch size and learning rate. The batch size determines the number of samples processed simultaneously before the model's parameters are updated, while the learning rate specifies the magnitude of this update, guiding the model's learning direction during backpropagation. From this optimization, I identified optimal settings:

* Learning rate = 0.00015080099528917156
* Batch size = 32

Due to the limitations of Google Colab's GPU and to manage computational resources effectively, I reduced the dataset sizes for this optimization process—400 samples for training, 60 for validation, and 80 for testing. This reduction was necessary to facilitate multiple iterations without overloading the available computing resources, allowing for 30 trials to determine the best combination of hyperparameters without the constraints of the full dataset size.

* 1. Model Training

Upon securing the optimal hyperparameter configuration, I advanced to the model training phase, opting for a slightly larger number of epochs than those recommended in "Identifying the Human Values behind Arguments" by J. Kiesel et al. Settling on 20 epochs, I incorporated routine evaluations on the validation dataset after each epoch to pinpoint the most effective model iteration throughout the training process. The criterion for model selection was the loss observed on the test dataset post-training. The best-performing model was archived in Google Drive for convenient access and subsequent analysis on the test dataset. It's critical to note that this training phase utilized the entirety of the available training data, with evaluations conducted using the complete test dataset to ensure comprehensive assessment and reliability of the results.

1. Results

In my analysis of the "SemEval 2023 Task 4: ValueEval" challenge results, I emphasize our model's achievement with a Micro F1-Score of 0.5515, demonstrating its competitiveness among the top models. This score reflects the robustness of our approach in multi-label classification of human values.

A screenshot of a computer

Description automatically generated

Figure 2: Results of Toronto Challenge[[3]](#footnote-4)

Further exploration through ROC analysis for each label, which treats the classification tasks individually, reveals our model's superiority over random choice in nearly all categories. This analysis, illustrated by the ROC graph, showcases the precision and recall balance achieved by our model.

A graph of different colored lines

Description automatically generated

Figure 3: Receiver Operating Characteristic for the multi-label classification.

Checking the F1-Score metric at the level of labels is possible to see a big difference between some of the classes and associate the performance of some of them with the number of samples available in the datasets. However, the model's performance on the "Humility" label, as indicated by an area under the curve (AUC) of less than 0.5, suggests a need for improvement. This specific finding points to a potential inadequacy in the dataset's representation for this value, implying that a more robust dataset could enhance training effectiveness.

The decision to utilize the Micro F1-Score for evaluation was strategic, allowing us to account for label frequency and assign appropriate importance to each label. This method contrasts with the Macro F1-Score, which averages the scores without considering label frequency, thus providing a more accurate reflection of the model's performance across varied datasets. This nuanced evaluation approach reveals the direct impact of sample size on the model's ability to learn and accurately classify human values.

A graph with blue lines and white squares

Description automatically generatedA graph of blue and white bars

Description automatically generated

Figure 4: F1-Scores by labels.

The analysis of the two graphs reveals a direct correlation between the performance of specific labels and their frequency within the dataset. This observation underscores the critical role of data quality during the model's tuning phase. High-quality, representative data is essential for ensuring the model's effectiveness and its applicability in real-world scenarios. This relationship between label frequency and model performance highlights the importance of thorough data preparation to enhance the model's ability to accurately classify and predict outcomes based on human values.

# Reference

Badr AlKhamissi, M. L. (2022). A Review on Language Models as Knowledge Bases.

Clément Delangue, J. C. (2016). *Hugging Face*. Retrieved from Hugging Face: https://huggingface.co/bert-base-uncased

Fabio Petroni, T. R. (2019). Language Models as Knowledge Bases? *Arxiv*.

Jacob Devlin, M.-W. C. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arxiv*.

Johannes Kiesel, M. A. (2022). Identifying the Human Values behind Arguments. *arXivLabs*.

Research, G. (2018). *Github*. Retrieved from Github: https://github.com/google-research/bert

Toutanova, J. D.-W. (n.d.). BERT: Pre-training of Deep Bidirectional Transformers for.

1. <https://touche.webis.de/semeval23/touche23-web/> [↑](#footnote-ref-2)
2. <https://huggingface.co/google-bert/bert-base-uncased> [↑](#footnote-ref-3)
3. <https://touche.webis.de/semeval23/touche23-web/> [↑](#footnote-ref-4)