

Multilabel Classification and Text Generation for Human Value Detection

Project Work

Michael Ghaly

Master's Degree in Artificial Intelligence, University of Bologna

michael.magdynasr@studio.unibo.it

Abstract

This report presents two approaches to address the challenges posed by the SemEval 2023 Task: "ValueEval: Identification of Human Values behind Arguments." The project emphasizes that normally, we only have access to the premise, making it the sole source of information for the predictions.

The first approach tackles the multi-label classification of human values based solely on the argument's premise by fine-tuning BERT, demonstrating the ability to extract nuanced information from argument premises and predict whether an argument draws upon one or more of the 20 predefined human value categories. The project explores the fine-tuning process and evaluation results, shedding light on the effectiveness of this approach in identifying human values within textual arguments using only the premise.

The second approach addresses the task of generating stances and conclusions based solely on the premise text. Here, FLAN-T5, a pre-trained language model, is fine-tuned to predict the correct stances (either "in favour of" or "against") and coherent conclusions of the premises. This project highlights the power of language models in producing contextually relevant responses, contributing to the generation of comprehensive and persuasive textual content while respecting the constraint of having only the premise as input.

This report delves into the methodology, data preprocessing, model fine-tuning, and evaluation metrics, discussing the challenges encountered and insights gained while working on these two distinct tasks, emphasizing the significance of the premise as the primary source of information. Additionally, an analysis is provided of the results achieved and their implications. The findings underscore the potential of advanced language models and pre-trained transformers in handling complex NLP tasks, such as argument analysis and human value identification.

1 Introduction

The surge in social media usage has ushered in an era of unprecedented user-generated content, providing a valuable window into individual thoughts, beliefs, and sentiments. However, the manual analysis and understanding of this vast reservoir of data have become increasingly formidable tasks. To address this challenge, the project centres on developing natural language processing techniques for the automatic comprehension and analysis of textual arguments.

The primary objective revolves around developing a value analysis system to discern the human values inherently tied to argument premises. This project leverages pre-trained transformers, which have exhibited exceptional performance across a spectrum of natural language processing tasks. Within the scope of this project, the BERT model has been fine-tuned for the multi-label classification task and tasked with assigning one or more of the 20 human values from the Schwartz value continuum to a given textual argument. The outcome of the fine-tuning process, as evident from model evaluations on the test set, yields an F1-micro score of 0.54 and an F1-macro score of 0.44.

Furthermore, a system for stance and conclusion generation has been developed. When presented with a textual argument, the objective of the system is to generate a corresponding stance statement, capturing whether the argument aligns with or opposes a specific conclusion extracted from the premise. This task holds profound implications across a diverse array of domains, encompassing politics, ethics, and the social sciences. It empowers policymakers to gauge public sentiment on pivotal issues while contributing to the development of AI-driven tools for content moderation and recommendation systems. To execute this task, sequence-to-sequence modelling, namely the Google FLAN-T5 architecture, has been employed.

The model has exhibited remarkable capabilities in natural language understanding and generation tasks. Through fine-tuning the FLAN-T5 model, the aim is to craft a robust stance generation system that excels in producing coherent and contextually apt responses. The model qualitative evaluation on the test set shows that language models can be used effectively for this particular task.

It is important to note that the models have been trained using only the premises provided in the dataset, mirroring the constraints of the real world where additional context is not guaranteed.

2 Background

2.1 Human Value Detection

In SemEval 2023 Task 4, also known as ValueEval, the primary objective revolves around the identification of human values conveyed within textual arguments. This task is centered on the classification of whether a given argument aligns with specific human value categories. To contextualize this task, it's essential to refer to the Schwartz Value Continuum. This framework was developed by Shalom H. Schwartz, a renowned social psychologist, with the aim of comprehending and categorizing human values. The Schwartz Value Continuum encompasses a broad spectrum of human values, offering valuable insights into the core principles that guide human behaviors, attitudes, and beliefs.

Within this continuum, various value categories represent distinct facets of human values. In the context of the "Human Value Detection" task, participants are presented with the challenge of classifying arguments that draw upon one or more of these value categories.

2.1.1 Multi-Label Classification

Multilabel classification is a task in machine learning where each input instance can be assigned to multiple labels or categories simultaneously. Unlike traditional classification, where one label is assigned per instance, multilabel classification handles scenarios where data can belong to multiple categories concurrently. This is useful in various applications like text classification, image recognition, and recommendation systems, where items can have overlapping attributes. Performance evaluation involves metrics like F1-score, Precision, and Recall.

2.1.2 Stance and Conclusion Generation

Stance generation involves determining whether a given premise is in favor of or against a particular issue or statement. This task requires models to capture the nuanced positions and beliefs expressed in the argument and to classify them accurately. In the context of SemEval 2023 Task, stances are binary, representing either support for or opposition to a particular idea or proposition.

Conclusion generation, on the other hand, requires models to synthesize a concise and coherent statement that encapsulates the main point or recommendation made in the argument's premise. It involves not only extracting key information but also rephrasing it in a clear manner.

2.2 BERT

BERT, standing for "Bidirectional Encoder Representations from Transformers," is a groundbreaking model in natural language processing (NLP) introduced by Google AI in (Devlin et al., 2019). BERT is celebrated for its contextual language understanding, setting a new standard in NLP. BERT's strength lies in its transformer architecture, which effectively captures word relationships in sentences. It undergoes two essential phases: pre-training on extensive text data to grasp general language understanding and fine-tuning for specific NLP tasks, making it versatile for various applications. BERT excels in producing contextually enriched word embeddings, allowing it to understand word meanings in diverse sentence contexts and handle complex language nuances. For this reason, the model's impact extends widely, setting new benchmarks in NLP tasks inspiring subsequent models.

2.3 Sequence-to-Sequence Architecture

Seq2Seq models are a class of deep learning models designed for natural language processing tasks. These architectures consist of two main components: an encoder and a decoder. The encoder encodes the input sequence into a fixed-length representation, often called a "context vector." The decoder then generates an output sequence based on this context vector. This architecture is particularly well-suited for tasks like machine translation, text summarization, and conditional content generation as they excel in capturing the underlying structure and relationships within sequences, allowing them to handle complex language tasks effectively. A variant of seq2seq models, known

as transformer-based architectures, has gained significant attention in recent years. Transformers leverage self-attention mechanisms to process input sequences in parallel and capture long-range dependencies. Pre-trained transformer models, such as BERT, GPT, and T5, have demonstrated remarkable capabilities across various natural language processing tasks due to their capacity to understand context and generate coherent text.

2.4 FLAN-T5

FLAN-T5 stands for "Fine-tuned LAnguage Net" and "Text-To-Text Transfer Transformer," respectively. The origins of FLAN-T5 can be traced back to Google's work in (Raffel et al., 2020) when they introduced the original T5 architecture. T5 demonstrated its prowess in various tasks, particularly in translation and summarization. Building upon this foundation, Google further refined the model in 2022 through (Chung et al., 2022). This evolution resulted in FLAN-T5, a language model that achieved strong few-shot performance even compared to much larger models. At its core, FLAN-T5 is an encoder-decoder model. During its training, the model has been exposed to a diverse range of tasks phrased as instructions rather than being specialized in a single task, which enables models to respond better to instructions and reduces the need for few-shot exemplars. The paper found that this technique, called instruction finetuning, coupled with scaling the number of tasks and scaling the model size dramatically improves performance on a variety of model classes (PaLM, T5, U-PaLM), prompting setups (zero-shot, few-shot, CoT), and evaluation benchmarks.

3 System description

3.1 Task 1: Multi-Label Classification

The dataset undergoes tokenization to convert the premise texts into input tensors suitable for the encoder model. This process results in a tokenized dataset with features for each dataset split (train, validation, test).

The system employs the BERT (Bidirectional Encoder Representations from Transformers) architecture, specifically using the "bert-base-uncased" pre-trained model for sequence classification. The model is configured for multi-label classification, as each argument may belong to multiple value categories.

The system defines custom evaluation metrics,

with a primary focus on the F1 score, computed as a micro-average. The *multi_label_metrics* function applies sigmoid activation to model predictions, converts them into integer predictions using a specified threshold, and computes the micro-average F1 score. The *compute_metrics* function integrates these metrics into the evaluation process for the Trainer.

The model is fine-tuned on the tokenized dataset using the Hugging Face Transformers library. The validation set is used to configure the training parameters, including learning rate, batch size, number of epochs, weight decay, and evaluation strategy. Then, the Trainer concatenates the training and validation datasets for training the model with the best parameters and uses the test dataset for evaluation. The best model, determined by the F1 score, is pushed to the hugging face hub and can be loaded at any later time for inference.

3.2 Task 2: Stance and Conclusion Generation

The system's goal is to predict the stance and conclusion of a given argument premise. Several preprocessing steps prepare the dataset for model training. Initially, we load the dataset, which includes the argument premise along with its corresponding stance and conclusion. Next, a preprocessing function creates a new dataset, where the premise text serves as the input, and the stance and conclusion are concatenated into a single output text. Lastly, the dataset is tokenized, with an appropriate maximum token sequence length determined to ensure efficient model training and inference.

The system fine-tunes the FLAN-T5 model's base variant, a sequence-to-sequence model capable of handling text generation tasks. For model training, we utilize the Hugging Face Seq2SeqTrainer library, where we configure various parameters such as the output directory for saving the trained model and its associated artefacts, batch size, learning rate, and the number of training epochs. Throughout the training process, we log metrics and save the model at the conclusion of each epoch, ultimately retaining the two best models based on their evaluation performance.

To assess the model's performance, we employed two evaluation metrics: ROUGE and BERTScore. These metrics gauge the correctness of the generated text by comparing it to the ground truth. During training, the system showed a

slightly superior performance when using ROUGE as the evaluation metric. Once trained, the model is pushed to the Hugging Face Model Hub, making it accessible for later inference.

4 Data

The dataset used in this project is designed to detect human values behind textual arguments. It consists of a set of 20 distinct labels representing various human values. The dataset is structured as follows:

- **ArgumentID:** A unique identifier for each argument.
- **Conclusion:** A concise representation of the argument.
- **Stance:** The stance of the premise towards the conclusion ("in favour" or "against").
- **Premise:** The text of the argument.
- **Labels:** A list of 20 binary labels, one for each value.

It's divided as shown in the table below:

Dataset Split	Number of Samples
Train	5393
Validation	1896
Test	1576

This dataset is based on the original Webis-ArgValues-22 dataset, which accompanies the paper "Identifying the Human Values behind Arguments" (Kiesel et al., 2022) published at ACL'22. It incorporates 7368 arguments from the IBM-ArgQ-Rank-30kArgs corpus (Gretz et al., 2019), a set of 1098 arguments from the Conference on the Future of Europe portal, and 399 arguments from the GD IDEAS web page. Each row in the dataset corresponds to an argument, consisting of its unique identifier (Argument ID), and the argument's content, including its Premise, Stance, and Conclusion.

The Labels for each example are represented as an array of 1s (indicating that the argument resorts to a specific value) and 0s (indicating that the argument does not resort to a specific value). The order of the 20 value categories corresponds to the level 2 of the value taxonomy (Kiesel et al., 2022), consistent with the original dataset files.

4.1 Data Exploration

The dataset under consideration exhibits a notable imbalance, as illustrated in the accompanying graph. This imbalance is evident in the distribution of data points across various classes, where certain ones are disproportionately overrepresented or underrepresented compared to others. This imbalance in class distribution may result in overrepresented categories having much more support and thus, may negatively affect the model's performance, especially for the minority classes.

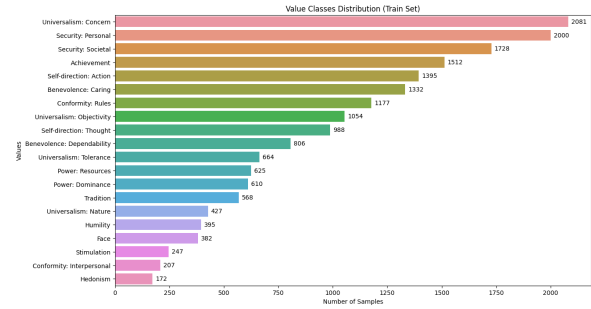


Figure 1: Class Distribution

Additionally, the distribution of the number of labels was analyzed for both the training and validation sets:

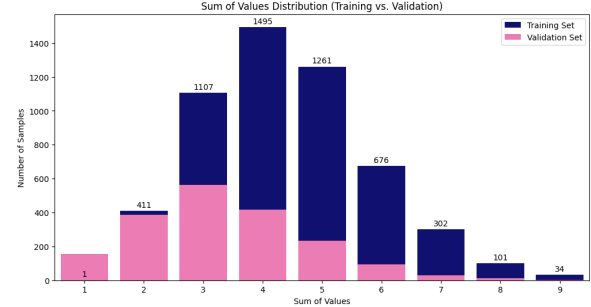


Figure 2: Number of Labels Distrubtion

Finally, the lengths of the premise texts were examined to ensure proper padding and truncation during the data tokenization, concluding that by setting the maximum input sequence length parameter to 64, just around 3% of the samples had to be truncated.

5 Results and Discussion

The performance of the multi-label classifier on the test dataset reveals a notable disparity in its ability to predict different human value labels. Precision, recall, and F1-scores were employed to gauge the classifier's efficacy for individual classes. Classes such as "Security: Personal" and "Univer-

salism: Concern" demonstrated high scores, indicating accurate predictions for these traits. Conversely, the "Humility" class displayed substantially lower results. This discrepancy in performance can largely be attributed to the imbalanced class distribution, where some traits are significantly underrepresented compared to others. The model's struggle to predict minority classes highlights the need for a more balanced dataset. Overall, the model achieved a micro-F1 score of 0.54 and a macro-F1 score of 0.44.

Meanwhile, the performance of the generation system was qualitatively evaluated by inspecting the generated predictions and comparing them to the ground truth provided in the test set. The model's results demonstrate promise in generating proper conclusions and classifying their stances.

Table 1: Comparison of Predictions and Ground Truths

Prediction: (in favor of) We should ban racial profiling Ground Truth: (in favor of) We should end racial profiling
Prediction: (in favor of) Guantanamo Bay should be closed Ground Truth: (in favor of) We should close Guantanamo Bay
Prediction: (in favor of) We should ban flag burning Ground Truth: (in favor of) We should prohibit flag burning
Prediction: (in favor of) We should ban women in combat Ground Truth: (in favor of) We should prohibit women in combat

6 Conclusion

This project undertakes the challenge of fine-tuning transformer models to automate the comprehension and analysis of textual arguments, with a focus on discerning human values associated with argument premises and generating corresponding stances and conclusions. The system fine-tunes the BERT model for multi-label classification of human values and employs the Google FLAN-T5 encoder-decoder architecture for stance and conclusion generation.

The evaluation of our multi-label classifier revealed varying levels of success across different human value labels. Classes such as "Security: Per-

sonal" and "Universalism: Concern" demonstrated high precision, recall, and F1-scores, indicating accurate predictions. However, the "Humility" class displayed lower performance, largely due to the imbalanced class distribution in the dataset. This underscores the importance of a more balanced dataset to improve model performance. Overall, our model achieved a micro-F1 score of 0.54 and a macro-F1 score of 0.44.

Meanwhile, our stance and conclusion generation system exhibited promise in generating coherent and contextually appropriate responses. Qualitative evaluation showed that the model was able to produce conclusions and stances that closely aligned with the ground truth in the test set.

While our models have demonstrated respectable performance, there is room for improvement, particularly in addressing class imbalance issues and employing bigger language model for more accurate generation.

The applications of this work extend across various domains, from politics to ethics and social sciences, enabling policymakers to gauge public sentiment and facilitating the development of AI-driven tools for content moderation and recommendation systems.

References

- Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. [Scaling instruction-finetuned language models](#).
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [Bert: Pre-training of deep bidirectional transformers for language understanding](#).
- Shai Gretz, Roni Friedman, Edo Cohen-Karlik, Asaf Toledo, Dan Lahav, Ranit Aharonov, and Noam Slonim. 2019. [A large-scale dataset for argument quality ranking: Construction and analysis](#).
- 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, Dublin, Ireland. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. [Exploring the limits of transfer learning with a unified text-to-text transformer](#).