SemEval-2023 Task 4: Identification of Human Values behind Arguments

Xingyu Chen **Kes Johnson**
University of Colorado at Boulder
Coursework for Natural Language Processing
{Xingyu.Chen, Kes.Johnson}@colorado.edu

Abstract

This paper is for shared task 4: Identification of Human Values behind Arguments. Given a textual argument and a human value category, classify whether the argument draws on that category. This task mainly focuses on a set of 20 value categories compiled from the social science literature. Arguments are given as premise text, conclusion text, and binary stance of the premise to the conclusion.

1 Introduction

Identifying human values in argumentative texts is the main purpose of this shared task. The data has been gathered from different cultures and resources to minimize bias. The data model is based on the existing model introduced in the white paper: *Identifying the Human Values behind Arguments.*

The motivation of this task is to improve current argument categorization, assessment, and generations. Human values often implicitly hide behind natural language arguments. Some applications use this concept both in real-world argumentation and theoretical argumentation frameworks such as semantic scene classification.

Human values are both studied in the social sciences and formal argumentation. Within computational linguistics, we can perform categorize, compare, and evaluate argumentative statements on this topic. However, due to the limitation of the data set and computation power, we need more data, time, and research to achieve the goal.

2 Data

We have introduced four different datasets in our research: WhitePaper, EarlyBird, Training, and Full. Each dataset was gathered from different resources. The dataset is split into two parts: labels and argument. Each argument consists of one premise, one conclusion, and a stance attribute indicating whether the premise is in favor or against the conclusion.

WhitePaper This dataset is being used in the white paper “Identifying the Human Values behind Arguments”. It has 5270 samples with an additional column for resources like ca country resources. We drop that column for this task.

EarlyBird This dataset is being released since the organizer of SemEval-2023 Task 4 has been announced. It has 5220 samples.

Training This dataset is being released on Dec 5. The completed dataset has been released along with the validation dataset and testing dataset. The testing dataset is not labeled, so we will use the model to output prediction and submit it for evaluation for the task. The training set has 5393 sample

Full This dataset is generated from combine two validation datasets and training datasets into one large dataset for training. The full set has 7389 samples.

The full dataset including with testing dataset has nearly 9000 arguments. The dataset is both available in Zenodo (where you can download everything except the test labels) and TIRA (where we directly submit our approaches as Docker images or upload our runs)

The full dataset is split into training, validation, and testing with a ratio of 60%/20%/20%. The validation set has two separate files because one of them was gathered from the Chinese Q&A site Zhihu.

The full dataset is roughly composed of 80% from the IBM argument quality dataset (95% in the original dataset), 15% from the Conference for the Future of Europe (new), and 5% from group discussion ideas (2% in the original dataset).

3 Data Model

The original F1 scores from the paper's level 2 model were 0.34 for BERT, 0.30 for SVM, and 0.28 for the baseline. The model we built using BERT achieved a 0.41 F1 score, which was slightly better than the paper.

The model was built by pulling in the small, uncased BERT model and adding layers on top of it for fine-tuning. Those layers were two cycles of a linear layer followed by a dropout of 0.2 followed by ReLU. The first cycle's node numbers went from the hidden side of the BERT model to 256 and the second went from 256 to 128.

The idea behind this was to give the model space to recognize more specific features before collapsing the layer to the 20-mode output. Dropout and ReLU were used because they had previous beneficial properties from previous models we had built. The learning rate was 0.001 and the epoch number was 4.

We found that adding more epochs didn't improve the outcome and took a long time to run. For the best model, we let BERT train along with my added layers, when we froze BERT the F1 score was 0.38. It ended up not being a huge difference.

	Precision	Recall	F1	Accuracy
1-Baseline	0.16	1	0.27	0.16
Bert	0.46	0.28	0.35	0.86
SVM	0.33	0.2	0.25	0.82

Table 1: Result for WhitePaper

	Precision	Recall	F1	Accuracy
1-Baseline	0.17	1	0.27	0.17
Bert	0.94	0.79	0.86	0.97
SVM	0.93	0.86	0.89	0.97

Table 2: Result for EarlyBird

	Precision	Recall	F1	Accuracy
1-Baseline	0.17	1	0.29	0.17
Bert	0.86	0.64	0.74	0.93
SVM	0.8	0.66	0.72	0.91

Table 3: Results from Training

We modify the parameter and configuration for the data model so the data model can accept the new dataset as input and perform prediction. We also modify the evaluation file to clean the output without level and country because they introduce the level and country features in the white paper

	Precision	Recall	F1	Accuracy
1-Baseline	0.17	1	0.29	0.17
Bert	0.85	0.62	0.71	0.92
SVM	0.78	0.62	0.69	0.91

Table 4: Result for Full

	Precision	Recall	F1	Accuracy
1-Baseline	0.17	1	0.29	0.17
Bert	0.23	0.04	0.07	0.84
SVM	0.98	0.98	0.98	0.99

Table 5: Result for Full of partial retrain

We use the original data model to evaluate the four datasets we introduced above. Due to retraining limitations. When we retrain the BERT model, it took 58 hours to finish the training on the Full dataset. Thus, we decide to predict all four datasets and test the dataset with the previous model.

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, Dublin, Ireland. Association for Computational Linguistics.

(PDF) Multi-Label Classification: An overview - researchgate. (n.d.). Retrieved December 11, 2022, from https://www.researchgate.net/publication/273859036_Multi-Label_Classification_An_Overview

(PDF) a review on multi-label learning algorithms - researchgate. (n.d.). Retrieved December 11, 2022, from https://www.researchgate.net/publication/263813673_A_Review_On_Multi-Label_Learning_Algorithms