Prodicus at SemEval-2023 Task 4: Enhancing Human Value Detection with Data Augmentation and Fine-Tuned Language Models

Erfan Moosavi Monazzah, Sauleh Eetemadi
Iran University of Science and Technology
moosavi_m@comp.iust.ac.ir
sauleh@iust.ac.ir

Abstract

This paper introduces a data augmentation technique for the task of detecting human values. Our approach involves generating additional examples using metadata that describes the labels in the datasets. We evaluated the effectiveness of our method by fine-tuning BERT and RoBERTa models on our augmented dataset and comparing their F_1-scores to those of the non-augmented dataset. We obtained competitive results on both the Main test set and the Nahj al-Balagha test set, ranking 14th and 7th respectively among the participants. We also demonstrate that by incorporating our augmentation technique, the classification performance of BERT and RoBERTa is improved, resulting in an increase of up to 10.1% in their F_1-score.

1 Introduction

When it comes to arguments, different people sharing the same values may have different opinions about whether one argument is persuasive or not. One cause of this is the difference between people ordering of values used to assess arguments. Within computational linguistics, human values can provide context to categorize, compare, and evaluate argumentative statements, allowing for several applications: to inform social science research on values through large-scale data sets; to assess argumentation; generate or select arguments for a target audience; and to identify opposing and shared values on both sides of a controversial topic (Mirzakhmedova et al., 2023). Human Value Detection task (Kiesel et al., 2023) is concerned with automatic classification of textual arguments to determine whether an argument draws on a specific human value. Provided dataset of arguments (Mirzakhmedova et al., 2023) is collected from different resources within different countries and beliefs. All arguments are presented either in English or have been translated into it.

To address the problem of detecting values behind arguments, we can design a multi-label classifier to determine whether an argument draws on a specific value or not. To tackle the classification problem, we propose two multi-label classifiers based on two widely used language models: 1) RoBERTa (Liu et al., 2019) and 2) BERT (Devlin et al., 2018); We evaluated these models performances on multiple test sets in two scenarios: 1) fine-tuned on the Main training set (Mirzakhmedova et al., 2023), 2) fine-tuned on our Augmented dataset. We provide all the code necessary to replicate our work in the paper's GitHub repository[1].

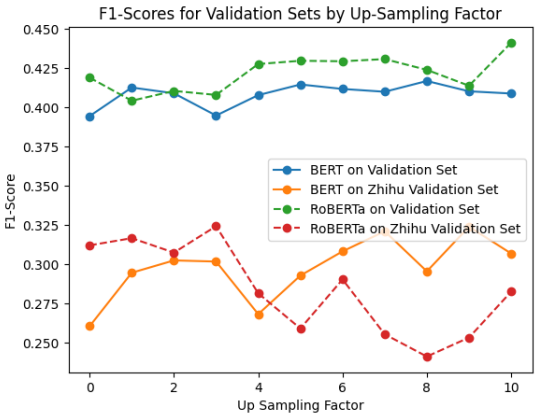

Figure 1: Comparing the effectiveness of augmentation among models and validation sets considering different up-sampling factors.

2 Background

The main challenge of identifying human values behind an argument is that in most cases the argument does not explicitly refer to the desired value. The first attempt to computationally extract human values behind an argument was done by (Kiesel et al., 2022). They first proposed a 4-level taxonomy of human values. After that they provided three baseline models to classify values in each level. For sake of this paper and this task (Kiesel et al., 2023)

[1] https://github.com/ErfanMoosaviMonazzah/SemEval2023-Task4-Human-Value-Detection

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 2033–2038

Conclusion	Stance	Premise	Label
We should ban fast food	in favor of	fast food should be banned because it is really bad for your health and is costly.	0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0
We should ban fast food	against	we all have the right to eat what we want, to prohibit a specific type of food is an abuse of power	1,1,1,1,0,0,0,0,0,0,0,0,0,1,0,0,1,0,1,0

Table 1: Example arguments from training set, Refer to (Mirzakhmedova et al., 2023) for more information regarding datasets.

Conclusion	Stance	Premise	Model Input
We should ban fast food	in favor of	fast food should be banned because it is really bad for your health and is costly.	We should ban fast food in favor of fast food should be banned because it is really bad for your health and is costly.
We should ban fast food	against	we all have the right to eat what we want, to prohibit a specific type of food is an abuse of power	We should ban fast food against we all have the right to eat what we want,to prohibit a specific type of food is an abuse of power

Table 2: Example inputs created by concatenation of Conclusion, Stance and Premise.

we only classified values from level 2 which consists of 20 value categories (Classes). Later on we compare our results to the result (Kiesel et al., 2022) obtained for level 2 value categories using these three baselines: 1-baseline, BERT-kiesel, as well as one of our own baselines, which is zero-shot classification of arguments.

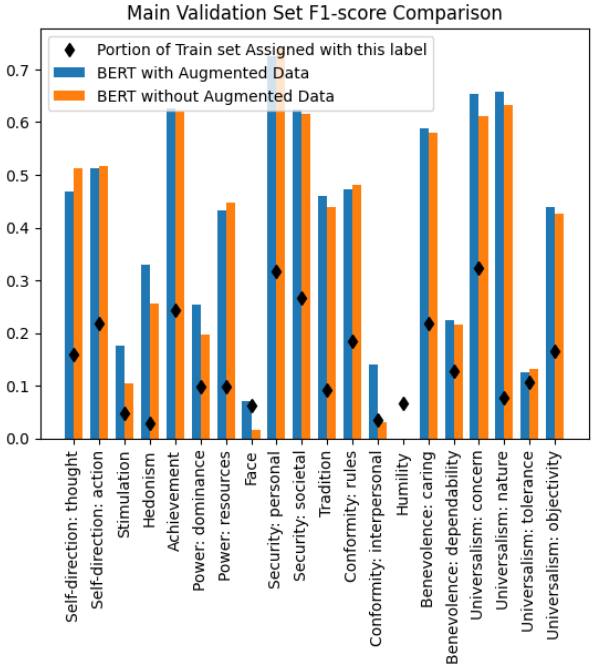

Figure 2: Demonstrating the augmentation effect on categories with small number of examples

3 Dataset

Task's Main dataset contains 9324 arguments on a variety of statements written in different styles, including religious texts (Nahj al-Balagha), political discussions (Group Discussion Ideas), free-text arguments (IBM-ArgQ-Rank-30kArgs), newspaper articles (The New York Times), community discussions (Zhihu), and democratic discourse (Conference on the Future of Europe) (Mirzakhmedova et al., 2023). All the arguments are either in English or translated into English. Each argument has a Premise, a Stance and a Conclusion and a list of binary values corresponding to value categories selected from all 20 value categories in level 2. (See Table 1) Arguments are then divided into 5 splits: Main Train, Main validation, Zhihu validation, Main test, Nahj al-Balagha test and New York Times test. They also accompany description data to describe the meaning of each value category by providing examples for them. The original training set consists of 5393 examples.

A challenging aspect of this task is that the models tend to struggle with predicting labels that have only a small number of examples in the training set (See Figure 2). To tackle this problem, we propose a data augmentation method which will create additional examples from meta data provided by task organizer. Including more examples can enhance the prediction of values that constitute a smaller fraction of the dataset. This, in turn, leads to an overall improvement in the macro-average F_1 score of models across all labels, by increasing the F_1 score of each label.

The main datasets consist of arguments, each of which comprises three parts: a premise, a stance, and a conclusion (Table 1). The conclusion takes either an in favor or against stance on its corresponding premise. To input these arguments into our models, we concatenate the three parts to form a single sequence (Table 2).

The value definition data is structured as a dictionary within a dictionary. The outer dictionary contains twenty key-value pairs, where each key represents a value category to be predicted. For each key (value category), the corresponding value

is an inner dictionary. These inner dictionaries consist of various key-value pairs, where each key corresponds to a level-1 label, and each value is the definition of that level-1 label ((Kiesel et al., 2022)).

We propose a template for generating new examples from the definition data. The template involves placing the definition before its corresponding level-1 value and inserting the phrase "is an example of" in between (See Table 3). Although the newly generated sentences have a different structure compared to the arguments in the dataset, we hypothesize that using these sentences can aid in predicting categories by expanding the vocabulary size of categories.

Through the application of this method on description data, we generated a new dataset that consists of 218 examples, each containing a single label corresponding to the value category which it defines. We use an up-sampling factor in order to control the number of augmented examples in dataset. For up-sampling factor k, the augmented sentences repeated k times in the main training set. Our method was evaluated for $k \in [0, 10]$.

Level-2 Value	Security: societal
Level-1 Value	Have a safe country
Definition	resulting in a stronger state
Created Input	resulting in a stronger state is an example of having a safe country
Created Label	One-hot vector $\in R^{20}$ where there is an 1 corresponding to Level-2 Value position in values list

Table 3: An example of creating data and label from value definitions.

Model	Set Name	Dataset	F1-Score
BERT	Nahj al-Balagha	Original	26.1
		Augmented	**29.1**
	New York Times	Original	19.4
		Augmented	**29.5**
	Main Test Set	Original	42.6
		Augmented	**43.3**
RoBERTa	Nahj al-Balagha	Original	24.7
		Augmented	**26.6**
	New York Times	Original	21
		Augmented	**27.5**
	Main Test Set	Original	43
		Augmented	**45.2**

Table 4: Comparison of our results against the original training dataset.

4 System Overview

Our proposed classifiers are heavily relied on transformer based pre-trained language models. Language modeling is the task of predicting the next word given a sequence of input words.

$$\mathrm{P}(\mathrm{w}_i|w_1, w_2, \ldots, w_{i-1}) = \frac{P(w_1,w_2,\ldots,w_{i-1},w_i)}{P(w_1,w_2,\ldots,w_{i-1})}$$

Currently, state-of-the-art results in language modeling are achieved by training transformer models that use an attention mechanism to extract a representation for an input sequence (encoding) and generate the next word from the extracted representation in a separate decoding module (Vaswani et al., 2017). Although language modeling differs from our classification task, we can transfer the learned knowledge about language structures from these models to our task by changing the model's head, which is the output layer of the decoder part of the transformer. Instead of predicting the next word using the output embeddings of the decoders, one can classify these embeddings using a fully connected layer. It has the benefit of only training the classifier head instead of the entire model. To accomplish this task, we fine-tuned two commonly used language models, BERT and RoBERTa. We provided arguments as input and generated a binary vector as output. Each digit in the binary vector corresponds to whether the input argument draws on that particular value category or not (refer to Table 2). To obtain zero-shot classification results, we employed bart-large-mnli ((Lewis et al., 2019)) which is a NLI model. (Yin et al., 2019) presented a technique for utilizing pre-trained NLI models as off-the-shelf zero-shot sequence classifiers. This method involves formulating the sequence to be classified as the NLI premise, and constructing a hypothesis for each candidate label. For instance, to determine whether a sequence belongs to the "politics" class, a hypothesis could be constructed as "This text is about politics." The probabilities for entailment and contradiction are subsequently converted into label probabilities.

5 Experimental Setup

The task organizer divided the data into three main sets (train, validation, test) and three supplementary sets (Zhihu validation, Nahj al-Balagha test, The New York Times test) (Mirzakhmedova et al., 2023). We generated eleven versions of the main training dataset, each one uses a different up-sampling factor k. We trained each model for 10 epochs and recorded its performance after each epoch. Both the BERT and RoBERTa models performed reasonably well after 10 epochs. We compared the effectiveness of augmentation with dif-

ferent up-sampling factors across different datasets and models (Figure 1). In terms of the main validation set, increasing the up-sampling factor usually resulted in an increase in the F_1-score. Both the BERT and RoBERTa models were trained on an NVIDIA GeForce GTX 1080 Ti for 10 epochs. For the task leaderboard, we submitted the results of RoBERTa fine-tuned on the augmented dataset with $k = 6$. We carefully tuned the batch size, learning rate and number of traning epochs of the model with respect to both Main and Zhihu validation sets. We repeated the training with different initial weights five times and submitted the best results out of the five attempts for the task. Our RoBERTa results placed 14th out of 41 on the leaderboard for the main test datasets and 7th among 20 submissions for Nahj al-Balagha test set (Table 5).

6 Results

We evaluated the effectiveness of our approach on the original test datasets using two models, BERT and RoBERTa, and the results are presented in Table 4. Our findings indicate that the proposed augmented dataset leads to an increase in F_1-score for the Nahj al-Balagha test set by 3% and the New York Times test set by 10.1% for BERT. For RoBERTa, the F_1-score increases are 1.9%, 6.5%, and 2.2% for the Nahj al-Balagha test set, the New York Times test set, and the main test set, respectively. Our carefully fine-tuned RoBERTa model achieves an F_1-score of 48% on the main test set, which is 6% higher than the results reported by (Kiesel et al., 2022) and 35% higher than the zero-shot baseline.

Regarding the Nahj al-Balagha test set, the RoBERTa model's F_1-score is 30%, which is 2% higher than (Kiesel et al., 2022) and 22% higher than the zero-shot baseline. A detailed summary of our results can be found in Table 5. We also observed that data augmentation had a positive impact on values with a smaller amount of data, as depicted in Figure 2. Augmenting the data for these values resulted in an increase in their F_1-score, as well as a slight increase in the F_1-scores of other values.

7 Conclusion

In this paper, we proposed two multi-label classification systems for the Human Value Detection task, using a fine-tuned version of BERT and RoBERTa language models. We also introduced a data augmentation method to improve the performance of our models. Our experimental results showed that our augmentation method can increase the F_1-score of the models. We also achieved competitive results on the Main test set and Nahj al-Balaghafor set, ranking 14th and 7th respectively among the participants. Our work demonstrates the effectiveness of using pre-trained language models for the Human Value Detection task and the potential benefits of data augmentation methods for improving model performance. Future work can explore additional augmentation methods and further improve the performance of the models by incorporating external knowledge or domain-specific information. Additionally, expanding the dataset to include arguments in other languages and from diverse cultures and perspectives can enhance the robustness of the models and better reflect the real-world scenarios.

8 Acknowledgments

I would like to express my gratitude to my parents and friends for their unwavering support throughout this project.

test set / Approach	All	Self-direction: thought	Self-direction: action	Stimulation	Hedonism	Achievement	Power: dominance	Power: resources	Face	Security: personal	Security: societal	Tradition	Conformity: rules	Conformity: interpersonal	Humility	Benevolence: caring	Benevolence: dependability	Universalism: concern	Universalism: nature	Universalism: tolerance	Universalism: objectivity
Main																					
Best per category	.59	.61	.71	.39	.39	.66	.50	.57	.39	.80	.68	.65	.61	.69	.39	.60	.43	.78	.87	.46	.58
Best approach	.56	.57	.71	.32	.25	.66	.47	.53	.38	.76	.64	.63	.60	.65	.32	.57	.43	.73	.82	.46	.52
BERT-kiesel	.42	.44	.55	.05	.20	.56	.29	.44	.13	.74	.59	.43	.47	.23	.07	.46	.14	.67	.71	.32	.33
1-Baseline	.26	.17	.40	.09	.03	.41	.13	.12	.12	.51	.40	.19	.31	.07	.09	.35	.19	.54	.17	.22	.46
zero-shot Baseline*	.13	.15	.32	.06	.02	.25	.13	.18	.11	.19	.08	.04	.02	.0	.0	.02	.0	.0	.0	.0	.0
RoBERTa	.48	.53	.61	.07	.27	.54	.32	.41	.15	.73	.62	.54	.51	.35	.11	.53	.15	.73	.78	.37	.43
Nahj al-Balagha																					
Best per category	.48	.18	.49	.50	.67	.66	.29	.33	.62	.51	.37	.55	.36	.27	.33	.41	.38	.33	.67	.20	.44
Best approach	.40	.13	.49	.40	.50	.65	.25	.00	.58	.50	.30	.51	.28	.24	.29	.33	.38	.26	.67	.00	.36
BERT-kiesel	.28	.14	.09	.00	.67	.41	.00	.00	.28	.28	.23	.38	.18	.15	.17	.35	.22	.21	.00	.20	.35
1-Baseline	.13	.04	.09	.01	.03	.41	.04	.03	.23	.38	.06	.18	.13	.06	.13	.17	.12	.12	.01	.04	.14
zero-shot Baseline*	.08	.02	.09	.01	.03	.27	.08	.0	.18	.15	.14	.0	.08	.0	.0	.0	.0	.0	.0	.0	.0
RoBERTa	.30	.17	.33	.00	.40	.59	.00	.00	.37	.42	.27	.53	.26	.07	.00	.38	.35	.23	.00	.17	.41
New York Times																					
Best per category	.47	.50	.22	-	.03	.54	.40	-	.50	.59	.52	-	.33	1.0	.57	.33	.40	.62	1.0	.03	.46
Best approach	.34	.22	.22	-	.00	.48	.40	-	.00	.53	.44	-	.18	1.0	.20	.12	.29	.55	.33	.00	.36
BERT-kiesel	.24	.00	.00	-	.00	.29	.00	-	.00	.53	.43	-	.00	.00	.57	.26	.27	.36	.50	.00	.32
1-Baseline	.15	.05	.03	-	.03	.28	.03	-	.05	.51	.20	-	.07	.03	.12	.12	.26	.24	.03	.03	.33
zero-shot Baseline*	.05	.06	.03	-	.06	.17	.0	-	.0	.06	.0	-	.0	.0	.0	.0	.0	.0	.0	.0	.0
RoBERTa*	.32	.57	.00	-	.00	.45	.00	-	.00	.52	.40	-	.19	1.0	.22	.31	.38	.37	.29	.00	.36

Table 5: Achieved F_1-score of team prodicus per test dataset, from macro-precision and macro-recall (All) and for each of the 20 value categories. Approaches marked with * were not part of the official evaluation. Approaches in gray are shown for comparison: an ensemble using the best participant approach for each individual category; the best participant approach; and the organizer's BERT and 1-Baseline. The New York Times dataset contains no argument resorting to Stimulation, Power: resources, or Tradition.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.

Johannes Kiesel, Milad Alshomary, Nicolas Handke, Xiaoni Cai, Henning Wachsmuth, and Benno Stein. 2022. Identifying the Human Values behind Arguments. In *60th Annual Meeting of the Association for Computational Linguistics (ACL 2022)*, pages 4459–4471. Association for Computational Linguistics.

Johannes Kiesel, Milad Alshomary, Nailia Mirzakhmedova, Maximilian Heinrich, Nicolas Handke, Henning Wachsmuth, and Benno Stein. 2023. Semeval-2023 task 4: Valueeval: Identification of human values behind arguments. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, Toronto, Canada. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2019. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *CoRR*, abs/1910.13461.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Nailia Mirzakhmedova, Johannes Kiesel, Milad Alshomary, Maximilian Heinrich, Nicolas Handke, Xiaoni Cai, Barriere Valentin, Doratossadat Dastgheib, Omid Ghahroodi, Mohammad Ali Sadraei, Ehsaneddin Asgari, Lea Kawaletz, Henning Wachsmuth, and Benno Stein. 2023. The Touché23-ValueEval Dataset for Identifying Human Values behind Arguments. *CoRR*, abs/2301.13771.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach.