Tübingen at SemEval-2023 Task 4: What can Stance Tell? A Computational Study on Detecting Human Values behind Arguments

Fidan Can
Department of Linguistics, Eberhard Karls Universität Tübingen
fidan.can@student.uni-tuebingen.de

Abstract

This paper describes the performance of a system which uses stance as an output instead of taking it as an input to identify 20 human values behind given arguments, based on two datasets for SemEval-2023 Task 4. The rationale was to draw a conclusion on whether predicting stance would help predict the given human values better. For this setup—predicting 21 labels—a pre-trained language model, RoBERTa-Large was used. The system had an F_1-score of 0.50 for predicting these human values for the main test set while this score was 0.35 on the secondary test set, and through further analysis, this paper aims to give insight into the problem of human value identification.

1 Introduction

Arguing can take place in our everyday life to express and rationalize ideas, and the way arguments are delivered can vary in their style, language and goal (Boltužić and Šnajder, 2014). In parallel with this variation, it is stated that it is a challenge to detect values expressed in the arguments as they are not always opinionated explicitly (Kiesel et al., 2022). Since argumentation is a part of our everyday life—for instance, while making decisions—it is likely that our values play a role, although this may not always be overt. Therefore, a large amount of literature on human value studies is an understandable effort. With this study, I hope to contribute to computational approaches to identifying human values behind arguments.

Kiesel et al. (2023) describe this current human value identification task as classifying whether a given textual argument falls into a given human value or not.

Out of all the experiments, the versions with stance used as a separate label in addition to the existing 20 labels (resulting in 21 labels) outperformed the rest. The underlying motivation was to mark whether the model would withdraw any information from the stance of the premise towards the conclusion. After the data pre-processing, the pre-trained transformer model, RoBERTa-Large (Liu et al., 2019), was implemented.

This approach did not turn out to perform as well on the secondary test set, Nahj al-Balagha, as it did on the main test set when the F_1-scores are taken into account—corresponding scores and ranks will be discussed later in the paper in addition to other approaches. This can very well ring a bell to how daunting it can be to identify human values due to their covert usage in arguments (Kiesel et al., 2022), especially considering that the arguments in the Nahj al-Balagha test set are from Islamic religious texts with grandiloquent content that includes sermons (Mirzakhmedova et al., 2023).

The aim of this paper is to present the role of stance and a pre-trained language model to identify human values behind given arguments for SemEval-2023 Task 4 by Kiesel et al. (2023) by drawing attention to where the system shows no immaculateness, along with two other approaches. Stanley Grenz was the code name I used for the system submission. For those interested, the source code is publicly available on GitHub.[1]

2 Background

An argument is comprised of a premise (or premises) and a conclusion (Boltužić and Šnajder, 2014), and for this task, we have the additional information of stance: whether the premise is in favor of the conclusion or against it. While the object of the task still remained as human value identification, I included the stance information as a separate label and used the RoBERTa-Large model to predict the stance as well as the 20 human values, which in principle resulted in this becoming a stance detection task, too.

According to Al-Khatib et al. (2020), an argument generally includes a core claim with a few

[1] https://github.com/fidan-c/human-value-detection.git

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 1763–1768
July 13-14, 2023 ©2023 Association for Computational Linguistics

supporting evidence for that. As stance expresses information in terms of premise and conclusion relationship, predicting the stance could mean for the model to learn how an argument can be constructed with parallel or contrasting ideas.

In addition, stance classification or detection tries to find out the position of a text towards a certain topic that is generally more abstract and may not be explicated in the text (Kobbe et al., 2020), and considering that the secondary test set, Nahj al-Balagha, differs in the way arguments are expressed with implicit messages, as a result of which predicting human values can become extra challenging, stance prediction could be a possible help to alleviate this challenge.

I chose a pre-trained language model for this task. As Wang et al. (2022) states, pre-trained language models (PTMs) have made great achievements in the field of NLP, which shifted the way from supervised learning approaches to "pre-training and fine-tuning".

Since the task in question is in line with the Natural Language Understanding (NLU) problem (i.e. comprehending text sequences) *transformer-encoder-only* (e.g. BERT and RoBERTa) architecture was chosen over *transformer-decoder-only* and *transformer-encoder-decoder* architectures.

More specifically, the system used RoBERTa-Large. For one of the two alternative approaches described later, another transformer model, BERT-Large (uncased), was used so as to compare performances. While RoBERTa-Large model is a model pre-trained on English language with the Masked Language Modelling (MLM) objective (Liu et al., 2019), BERT is a pre-trained language model on English—with MLM and Next Sentence Prediction goals (NSP) (Devlin et al., 2018).

Liu et al. (2019) points out that BERT was significantly undertrained, and with certain modifications, such as training the model with bigger batches on more data and on longer sequences and removing the NSP objective, an optimized model named RoBERTa was built.

3 System Overview

As it was briefly mentioned earlier, the main system is based on using the stance information as an additional label rather than using it as a part of the input. For this reason, the goal turned into predicting 21 labels in total, 20 of which are the given human values. For the input, the conclusion and the premise columns in the dataset were concatenated together. While concatenating these two, I used the <*s*> token in-between. The input formation can be seen in Figure 1.

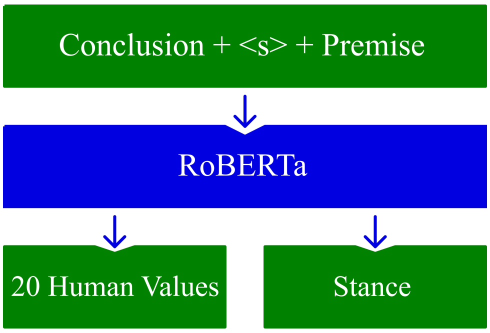

Figure 1: The main system for inference

The input then was fed into RoBERTa (more specifically RoBERTa-Large), and I fine-tuned this pre-trained language model, which included setting the number of iterations to 5 and the the batch size for training to 16 to predict 21 labels: predicting the 20 human values and the stance. The following sections will elaborate on the system setup and performance further.

4 Experimental Setup

4.1 Data

For this task, two different argument datasets were used, one being the main dataset and the other being a secondary test set, Nahj al-Balagha, for the purpose of checking the robustness of the model.

The main dataset was already split into training (61% with 5393 rows/arguments), validation (21% with 1896 arguments) and test (18% with 1576) sets (arguments-training/validation/test.tsv) each of which includes 4 columns: "Argument ID", "Conclusion", "Stance" and "Premise", respectively. The premise's stance towards the conclusion is described as *in favor of* and *against*. The labels for 20 human values were given as separate datasets with the corresponding splits (labels-training/validation.tsv).

This main argumentation dataset is compiled from IBM-ArgQ-Rank-30kArgs (2019-2020), Conf. on the Future of Europe (2021–22), where English, French, and German are the main languages among the other 23 languages— they were automatically translated into any EU24 languages— and from Group Discussion Ideas (2021–22). Therefore, it is safe to state that this dataset includes English as the original language as well as automatic translations into English.

As for the secondary test set, Nahj al-Balagha, it contains 279 arguments extracted from Islamic religious texts written originally in Arabic and includes advice and sayings where ideas are delivered implicitly due to its eloquent content. This dataset was manually translated to English. Further elaboration on these two datasets can be found in Mirzakhmedova et al. (2023).

4.2 Data Pre-processing

Having used the aforementioned datasets, building the system required data pre-processing and model fine-tuning. Following the random seed being set to 64, stance was used as an output, and conclusion and premise were concatenated together with the *</s>* token in-between and included as an input. After lower-casing this new input, RoBERTa-Large model was used through the Simple Transformers library (Rajapakse, 2019).

4.3 Hyperparameter Tuning

Hyperparameter tuning was solely comprised of specifying the number of epochs as 5 and the batch size for training as 16.

The other hyperparameters had the default values of the Simple Transformers library for the Multi-Label Classification model architecture: *4e-5* for the learnig rate, *AdamW* for the optimizer and *Binary Cross-Entropy* for the validation loss (Rajapakse, 2019).

4.4 External Libraries and Tools

The versions of the external tools and libraries used in the GoogleColab environment for the system are as follows:

- Python - 3.8.10.
- numpy - version: 1.21.6. [2]
- pandas - version: 1.3.5. [3]
- scikit-learn (Buitinck et al., 2013) - version: 1.0.2. [4]
- simpletransformers (Rajapakse, 2019) - version: 0.63.9. [5]
- transformers (Wolf et al., 2020) - version: 4.26.1. [6]

4.5 Experimental Setup - Other Approaches

Two more experiments were done to analyze input and the model difference; however, these were not submitted. I believe they still carry meaning that can be used to discuss the main approach.

First, for the sake of comparison on whether using stance as an additional label contributed to model performance or not, stance was included as a part of the input. In this setup, conclusion, premise and stance were concatenated together with the same data pre-processing that the main approach had (i.e. lower-casing and *</s>* token in-between), and everything else was kept the same (i.e. same seed number, using RoBERTa-Large with the same hyperparameters, etc.). This first alternative approach resulted in F_1 = 0.45 on the arguments-validation dataset.

The last approach concerned changing the model. Stance was again used as a label, but with a different pre-trained language model. Instead of RoBERTa-Large, BERT-Large (uncased) was used, and everything else was kept the same except for lower-casing and replacing the *</s>* token: *</s>* with *[SEP]*. This setup, when evaluated on the main validation set, resulted in F_1 = 0.39.

These two approaches were to determine which input design and which pre-trained model could contribute to the overall performance, and neither could outperform the main approach. The score details for the submitted/main approach will be given in the following section.

5 Results and Discussion

A glance at the published F_1-scores for the main argument test set and the Nahj al-Balagha test set can give a broad idea for the performance of the main system this paper is focused on. Table 1 in the Appendix is provided for the readers in case of a need for looking into further details.

It is noticeable that the model performed better on the main argument test set than the secondary test set: F_1 = 0.50—ranked ninth, with 0.56 being the highest score—compared to F_1 = 0.35—ranked second, with 0.40 being the highest.

There are 10 labels on the main test set where F_1-score is lower than 0.50, with *Stimulation* having the lowest score of 0.10, followed by *Humility* with

[2] https://www.numpy.org

[3] https://pandas.pydata.org

[4] http://scikit-learn.org

[5] https://github.com/ThilinaRajapakse/simpletransformers/

[6] https://github.com/huggingface/transformers

a score of 0.12. The top score is seen in *Security: personal* with 0.77.

For the secondary test set, the label that got the highest F_1-score is *Achievement* with 0.66 while *Security: personal* was ranked third with 0.51. Similar to the main test set, *Stimulation* and *Humility* are among the lowest scoring labels, with 0.00 and 0.08 respectively.

In fact, 4 labels that each have $F_1 = 0.00$ are in this secondary test set: *Stimulation*, *Power:dominance*, *Power:resources* and *Universalism:tolerance*. Another 13 labels are below 0.35.

Across these two datasets, one can deduce that *Stimulation* and *Humility* were poorly predicted. There can be several reasons for that, but one that can be already seen in Mirzakhmedova et al. (2023) is that the data frequency for these two human values is relatively low for the main dataset.

However, though one can consider the influence of data frequency as a possible reason—which might be a factor for *Power:resources* having F_1-score of 0.00 for the secondary test—Table 2 in the Appendix shows another participant's system (code name r-m-haare) giving $F_1 = 0.33$ for this label on the secondary test set. It is intriguing to hypothesize the potential reasons behind this result.

Another question arises pertaining to the way of expressing ideas based on where the arguments were taken from. As mentioned earlier, the secondary test set has a nature of including covert messages embedded in its eloquent content, and that might be a challenge for the pre-trained models. For this, training on bigger data with more variety may help to better predict the hidden human values behind covert arguments.

6 Conclusion

This paper presented a system to predict 20 human values behind arguments by using a pre-trained language model, RoBERTa-Large, and with specific hyperparameter tuning; the main strategy employed was to use stance as a label, which meant that the system was set to predict the human values and the stance for each argument. For comparison, two other approaches were also described in the paper—one for comparing the effect of another pre-trained model and the other for the impact of including stance in the input.

Overall, it was shown that the main system outperformed the other two. Despite the reasonable rank that the system got in SemEval-2023 Task 4, a close look into where the system did not perform well was worth delving into.

Based on the secondary test set results, it seems that language use for expressing ideas may vary depending on the content. This can be a phenomenon to investigate in a future comparative study with a bigger and a more balanced dataset, which could potentially give an insight into the performance of pre-trained models for pragmatics and see whether or not human values in arguments can be predicted better despite implicity and covertness. Another suggestion is linked to transfer-learning: a model trained for ideology detection can be used on this task to see if it would contribute to the system performance on identifying human values behind arguments.

References

Khalid Al-Khatib, Yufang Hou, Henning Wachsmuth, Charles Jochim, Francesca Bonin, and Benno Stein. 2020. End-to-end argumentation knowledge graph construction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 7367–7374.

Filip Boltužić and Jan Šnajder. 2014. Back up your stance: Recognizing arguments in online discussions. In *Proceedings of the First Workshop on Argumentation Mining*, pages 49–58.

Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt, and Gaël Varoquaux. 2013. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pages 108–122.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

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.

Jonathan Kobbe, Ioana Hulpuș, and Heiner Stuckenschmidt. 2020. Unsupervised stance detection for arguments from consequences. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 50–60, Online. Association for Computational Linguistics.

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. *CoRR*, abs/1907.11692.

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.

T. C. Rajapakse. 2019. Simple transformers. https://github.com/ThilinaRajapakse/simpletransformers.

Haifeng Wang, Jiwei Li, Hua Wu, Eduard Hovy, and Yu Sun. 2022. Pre-trained language models and their applications. *Engineering*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

A Appendix

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	.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
Stanley-Grenz	.50	.55	.67	.10	.29	.61	.34	.49	.18	.77	.65	.62	.52	.29	.12	.57	.23	.75	.79	.38	.42
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	.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
Stanley-Grenz	.30	.12	.46	.00	.57	.66	.00	.00	.20	.51	.29	.41	.26	.11	.08	.41	.23	.24	.40	.00	.39

Table 1: Achieved F$_1$-score of team stanley-grenz per test dataset, from macro-precision and macro-recall (All) and for each of the 20 value categories. 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.

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	.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
r-m-haare	.51	.48	.66	.22	.23	.61	.43	.45	.32	.74	.63	.57	.54	.47	.15	.53	.36	.74	.81	.42	.55
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	.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
r-m-haare	.34	.08	.31	.17	.40	.62	.09	.33	.51	.49	.29	.45	.21	.14	.21	.28	.23	.27	.50	.00	.25

Table 2: Achieved F$_1$-score of team r-m-hare per test dataset, from macro-precision and macro-recall (All) and for each of the 20 value categories. 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.