

Augustine Of Hippo at SemEval-2023 Task 4: An Explainable Knowledge Extraction Method to Identify Human Values in Arguments with SuperASKE

Alfio Ferrara, Sergio Picascia, Elisabetta Rocchetti

University of Milan,

Via Celoria 18, 20133, Milan, Italy

{alfio.ferrara, sergio.picascia, elisabetta.rocchetti}@unimi.it

Abstract

In this paper we present and discuss the results achieved by the "Augustine of Hippo" team at SemEval-2023 Task 4 about human value detection. In particular, we provide a quantitative and qualitative reviews of the results obtained by SuperASKE, discussing respectively performance metrics and classification errors. Finally, we present our main contribution: an explainable and unsupervised approach mapping arguments to concepts, followed by a supervised classification model mapping concepts to human values.

1 Introduction

Argumentation is a communication act requiring a strong pragmatical component to be entirely understood by humans. This high-level language characterization encodes contextual information which, in turn, can be partitioned into linguistic, cognitive, physical, semantic, and social aspects. Depending on actual and perceived context, individuals adjust their communication in order to convey a message as effectively as possible: this leads to the possibility of many concepts to remain implicit, especially those that can be deduced from context, which includes people's beliefs and values. SemEval-2023 Task 4 Kiesel et al. (2023a) aims at exploring novel methodologies for extracting human values from argumentation texts written in English. Focusing on this type of contextual information has the benefit of improving argumentation analysis and generation Kiesel et al. (2022).

Our contribution as "Augustine of Hippo" team consists in SuperASKE¹, an explainable system made of ASKE (Automated System for Knowledge Extraction) and Random Forest producing a binary output. Explanations are made possible by associating human values to *concepts* extracted by ASKE,

¹Docker container image available at sergiopicascia/semEval-superaske

which are ordered by Random Forest's feature importance. Our model has a median performance with respect to other runs submitted for the task. The work is organized as follows: Section 2 provides some background about the task of human values detection from arguments and the original ASKE model; Section 3 explains the behaviour of the whole framework; Section 4 specifies the details on how the experiment was run; Section 5 discusses the performances of the approach; Section 6 sums up the most important takings of the research.

2 Background

SemEval-2023 Task 4 is about human value detection. This task requires a classification model to recognise which among 20 values (or a subset of them) are present in a textual argument. Specifically, the subset of human values considered is taken from Kiesel et al. (2022) which, in turn, discusses a selection of values inspired by Schwartz et al. (2012), Gouldner (1975), Brown and Crace (2002) and C. et al. (2020). The authors in Kiesel et al. (2022) describe 4 levels contributing to the value taxonomy, which can be synthesized as follows: the first level contains 54 individual values; the second level specifies 20 values categories aggregating individual values; the third level contains 4 higher-order values; the last two levels specify base dichotomies². Our work focuses on extracting second-level value categories from arguments.

In particular, each argument is represented as a triple containing a conclusion, a stance and a premise; an example is depicted in Table 1. The premise represents a practical example of a situation for which someone could express an opinion. The stance indicates whether the conclusion statement is in favor or against the sentiment depicted

²The interested reader can refer to Kiesel et al. (2022) for a more detailed explanation of individual values and value taxonomy.

Conclusion:	We should prohibit school prayer
Stance:	against
Premise:	it should be allowed if the student wants to pray as long as it is not interfering with his classes
y :	[1, 1, 0, 0, ..., 0, 1, 0, 0, ..., 1, 0, 0, 0]

Table 1: Example representing an input argument. The last row reports the one-hot encoded label, which contains four values *self-direction: thought*, *self-direction: action*, *tradition* and *universalism: concern*.

in the premise. Finally, the conclusion conveys an idea according to the respective premise and stance. The target of classification is formulated as a vector $y = [0, 1]^{20}$ indicating the presence/absence of a value in an argument.

The task Kiesel et al. (2023b) is organized as follows. The main dataset is taken from the work by Mirzakhmedova et al. (2023), which has 8865 instances; this data is divided in three splits: the training set, the main validation set and the main test set. For validating the robustness of approaches, there is an additional labeled collection including 100 arguments from the recommendation and hotlist section of the Chinese question-answering website Zhihu. Lastly, 279 arguments from the Nahj al-Balagha and 80 arguments from the New York Times articles related to the Coronavirus are made available as extra test sets. All arguments are written in English, even though source languages differ. Additionally to these datasets, a value taxonomy is available in json format: this file contains all value categories and their respective values described through sample sentences (see Listing 1 for an example).

```
{ "Self-direction: thought": {
  "Be creative": [ "allowing for more
    creativity or imagination",
    "being more creative", ... ],
  ... }, ... }
```

Listing 1: Value taxonomy example from json file.

Our solution is trained, validated and tested using all datasets and resources made available by the task organizers.

2.1 Automated System for Knowledge Extraction

ASKE (Automated System for Knowledge Extraction) Ferrara et al. (2023) is a framework focused on extracting structured knowledge from textual corpora through the exploitation of context-aware embeddings in a zero-shot setting. ASKE is an iterative process, meaning that its three main phases, namely zero-shot classification, terminology enrichment, and concept formation, can be repeated

for an arbitrary number of *generations*. It has also the advantage of being completely explainable, since the extracted knowledge is expressed in order to be completely understandable by human beings.

ACG: ASKE Conceptual Graph. In ASKE, all the relevant information are collected in a graph-based data structure, called ASKE Conceptual Graph (ACG). The nodes of the ACG can be of one of the following three kinds: (i) document chunks $K = \{(k, \mathbf{k})\}$, where k is a portion of text (*autonomous cars can make trips more relaxing and help the traveler arrive refreshed and happy*) and \mathbf{k} is its vector representation; (ii) terms $W = \{(w_s, w_d, \mathbf{w})\}$, where w_s is an n-gram extracted from the document chunks (*free time*), w_d is its definition retrieved from an external knowledge base (*time available for hobbies and other activities that you enjoy*), and \mathbf{w} is the vector representation of such definition; (iii) concepts $C = \{(c, \mathbf{c})\}$, representing cluster of terms (*free time = {free time, trip, travel, game}*), where c is the label assigned to the concept, given by the w_s of the closest term to the centroid, and \mathbf{c} is its vector representation, given by the centroid itself.

The edges linking these nodes are defined by the following relations: (i) classification, linking a document chunk and a concept that classifies it; (ii) origination, linking a term and the first concept to which it is assigned to; (iii) belonging, linking a concept and the terms occurring in its cluster; (iv) derivation, linking two concepts, one being the parent of the other.

At the beginning of the analysis, the ACG is initialized with the textual components k of text chunks K and one or more initial concepts C with the associated terms W . For example, one may define as initial concept a human value, i.e. *stimulation*, and link to it some dummy terms, i.e. *stimulation_1*, whose definition w_d is the one provided in the values taxonomy, i.e. *allowing people to experience foreign places*.

Together with their textual counterpart, also the vector representation are stored in the ACG, embedding the text using a large language model (LLM) capable of computing context-aware embeddings; in particular, we choose Sentence-BERT, a modified version of the original BERT model, which exploits siamese and triplets networks, in order to derive context-aware sentence embeddings, which are able to capture the semantic aspect of the em-

bedded text.

Zero-Shot Classification. In this phase text chunks are assigned to the concepts occurring in the ACG. Being the embeddings computed by the same embedding model, the sets of \mathbf{k} and \mathbf{c} exist together in the same semantic vector space. Therefore, ASKE is able to perform a zero-shot classification, meaning that it can assign text chunks to concepts, $f : K \rightarrow C$, without having been subject to a training process. This association is performed based on a similarity measure σ , i.e. cosine similarity, computed between the vector representations \mathbf{k} and \mathbf{c} . If this similarity is higher than a predefined threshold α , then k_j is classified as c_i .

For instance, the text chunk *autonomous cars can make trips more relaxing and help the traveler arrive refreshed and happy* is associated with the concept *stimulation* with a similarity score of 0.3.

Terminology Enrichment. From the text chunks K_i classified as a concept c_i , ASKE retrieves the set of terms W_i appearing in them. These terms are then projected in the same vector space of \mathbf{K} and \mathbf{c} , computing the vector representation of their definition retrieved from an external knowledge base, i.e. WordNet. Then, ASKE computes the similarity σ of \mathbf{w} w.r.t. the vectors representing the concept and the text chunks. The top γ candidates having a similarity greater than the threshold β are assigned to c_i .

Considering the previous example, the terms *trip* and *traveler* are assigned to the concept *stimulation* respectively with similarity 0.96 and 0.97.

Concept Formation. In its final phase, ASKE runs a clustering algorithm, i.e. Affinity Propagation Frey and Dueck (2007), over the embedding vectors of the terms W_i belonging to a concept c_i . Based on the results, different operations can be enforced: (i) *conservation*: the original concept c_i is preserved, consisting in the cluster in which the term w_h corresponding to the label of c_i appears; (ii) *derivation*: the newly generated clusters, different from c_i , become new concepts; (iii) *pruning*: if a cluster c_j is made only of terms that belong also to c_i , the former concept is absorbed in the latter one. For example, from the concept *stimulation*, ASKE derived the concept *free time*, consisting of the following set of terms: $\{\textit{free time, trip, travel, games}\}$.

3 System Overview

The framework we propose for detecting human values is a concatenation of ASKE and Random Forest Breiman (2001), with the results of the analysis conducted by ASKE being employed as input for the RF model. Every instance of this framework is tailored on a single human value, meaning that it solves a binary classification task. The main advantage of this framework is having two explainable models: concepts in ASKE are described by the terms that compose them with the corresponding definitions, while also being features of the RF model, which provides the importance for each of them in the trees. This allows to identify which are the most influential concepts and how much they affect the final predictions.

Despite ASKE being a completely unsupervised model, running in a zero-shot setting, its flexibility gives us the chance of proposing it in its supervised version, SuperASKE, tuned for classification. First of all, we proceed fine-tuning the sentence embedding model employed for computing the vector representation of the ACG entities (all-MiniLM-L6-v2³). Being based on a siamese architecture, the model is fine-tuned by providing a pair of sentences and their corresponding semantic similarity. Therefore, we retrieve all the premises from the training set, and all the descriptions of the human values provided by the task organizers: if a premise p is classified with a certain human value v , all the possible pairs of p and the descriptions of v are given to the model with a similarity score of 1, otherwise the similarity is set to 0.

The fine-tuned embedding model is then employed to compute the vector representation of the initial ACG components. ASKE is initialized with only one concept, representing a single human value v , associated with some dummy terms, having as definitions the ones provided in the value taxonomy. As document chunks, we consider only the premise of each argument, excluding stances and conclusions which appeared to not benefit to the final results. Afterwards, the model is run as usual for multiple generations, providing it only the premises positively classified as v : in such a way, we ensure that the knowledge extracted by SuperASKE is relevant to the human value analyzed. The final version of the ACG is then exploited in order to compute the similarities between the concepts oc-

³Model available at <https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2>

curing in it and the whole set of premises, now including also the one negatively classified.

The RF model is employed for the final binary classification. The observations are represented by the document chunks K , i.e. the premises; the features are the concepts C in the ACG; the values are given by the similarities between K and C ; the labels are the ground truth determining if a premise is classified or not with a given human value.

4 Experimental Setup

We test our model as follows. First, data must be split into training set, validation set and test set. This procedure has been performed by the task organizers, who have provided a unique split for all participants to use: arguments are divided in training set (61%), validation set (21%) and test set (18%); moreover, classes distribution among splits is the same.

For each human value v , we retrieve only the set of premises classified as v and we run different configurations of SuperASKE, changing the hyperparameters α , β and g , with $\alpha \in \{-1, \dots, 0.5\}$ being the similarity threshold for the zero-shot classification phase, $\beta \in \{-1, \dots, 0.5\}$ being the similarity threshold for the terminology enrichment phase, and $g \in \{0, \dots, 8\}$ being the number of generations, i.e. the number of SuperASKE cycles performed. Each configuration learns a different version of the ACG, with its peculiar concepts and assigned terms. The different ACGs are then employed for computing the similarities between each premise in the entire dataset, considering also the one not classified as v , and each concept occurring in the ACG. These similarities are used as input for the RF model, trained in order to predict the correct label for each premise w.r.t. the human value v .

Based on the performances of the RF model on the validation set, we pick the best configuration of hyperparameters for both models, SuperASKE and RF. We then proceed repeating the same steps for each human value, training 20 different binary classifiers. Evaluation is performed in two ways: using F1, precision and recall measures for each class independently and computing macro-averages over all categories. Official evaluations have been done on TIRA platform Fröbe et al. (2023).

5 Results

Figure 1 depicts SuperASKE performances and how they compare to other significant models.

Other pictures and tables reporting performance measures are in the Appendix A.1. SuperASKE's F1 is placed near the median for the majority of the value categories. Though, this behaviour has some exceptions: "Self-direction: thought" and "Hedonism" F1 scores are in the first quartile, whereas "Power: resources", "Stimulation" and "Humility" F1 scores are in the fourth quartile. Considering precision and recall it can be noticed that our model is, on average, less precise than the median precision score; however, SuperASKE has a higher recall, on average, than the median recall score.

5.1 Error Analysis

Frequency-performance correlation. There is a positive correlation between models performance and value categories frequencies in datasets (see Appendix A.3 for correlation tables). In this perspective, it is curious to see such a high F1 score for "Universalism: nature". It can be hypothesized that this value category has a specific vocabulary, thus when nature-related words appear it is easier to guess the right class, both for humans and for automated models. Examples of words contained in "Universalism: nature" arguments are: "whaling", "human cloning", "nuclear weapons".

Confusion matrices. Our model has greater recall than others. Indeed, it can be noticed that false positives (FP) frequencies tend to be higher than false negative (FN) frequencies (confusion matrices are in Appendix A.2). Extracting some instances evaluated either as FP or FN in the training set, it is possible to qualitatively categorize the type of error made by SuperASKE. Let's consider the following argument:

ID: E04080

Conclusion: We need an inclusive and pluralistic European society.

Stance: in favor of

Premise: There need to be some rules for integration: Integration does not mean giving up European values and culture.

TP values: Tradition

FP values: Security: personal, Security: societal, Conformity: rules, Universalism: concern

In this argument the main topic discussed is integration for immigrants. Some FP value categories cannot be associated with this instance; however, "Conformity: rules" could be associated with "rules for integration", and "Universalism: concern" is about equality, which is a prerogative for integration measures. "Security: personal" and "Security: societal" are somewhat connected to immigration, but they are not directly involved in this argument.

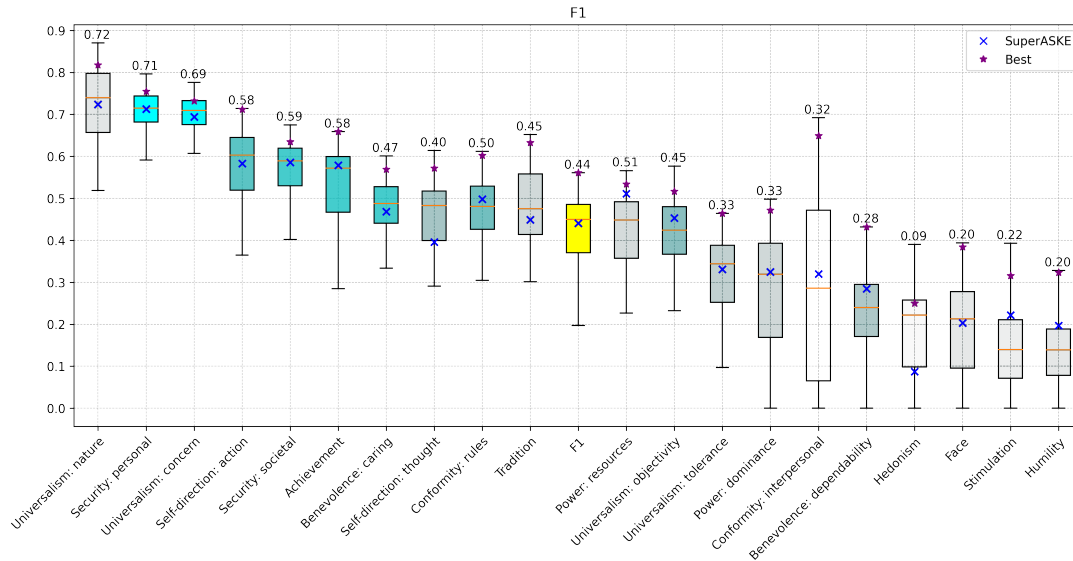


Figure 1: F1 scores distributions aggregating all the runs submitted to TIRA for each value category, computed using main test set predictions. The blue cross depicts SuperASKE’s F1 score, which is also reported above; the purple star depicts the best model’s F1 score. The yellow horizontal line represents the median value. Color intensity is proportional to the frequency of that value category. Plots are ordered according to the median values. Macro-average distribution is colored in yellow.

This fact highlights how subjective is to label arguments with human values, making it difficult to train NLP models effectively.

5.2 Explanation

Thanks to the explainable aspect of the framework, we are able to draw some conclusions about the behaviour of the two models. First of all, we analyse which are the most influential concepts for various values, reporting also their feature importance in parenthesis. For ‘Stimulation’, some of the discovered concepts are: (i) ‘act’ (0.25), containing terms related to theatrical performances and shows, such as ‘acting’ and ‘fair’; (ii) ‘free time’ (0.19), referring to leisure activities like ‘game’ and ‘trip’; (iii) ‘travel’ (0.1), including verbs regarding movements, such as ‘engage’, ‘carry’ or ‘play’. For ‘Power: resources’ the concepts are mostly related to financial activities, such as ‘trade’ (0.13), ‘raise’ (0.18) and ‘advance’ (0.11), or to economical conditions, i.e. ‘poor’ (0.07). Finally, we find a quite peculiar result for the value ‘Humility’, whose most relevant concepts are mostly time-related, i.e. ‘old’ (0.17), ‘conclusion’ (0.15) and ‘middle’ (0.14).


We also check for the relevance of the concepts discovered by SuperASKE, finding out that the more concepts it finds, the better the results. Indeed, according to the RF model, the importance of the original concept usually does not exceed 0.3, especially on values where our framework performs above the median value in terms of F1. This behaviour is also confirmed by the Pearson correlation test we run

between the number of concepts discovered and the number of outperformed models: the test shows a positive correlation of 0.41, with a p-value < 0.05 . The values on which SuperASKE performs below the median value are usually the ones on which, for reasons related to the hyperparameter tuning phase, it has been run for 0 generations, i.e. it does not discover any knowledge at all.

6 Conclusion

In this paper we used SuperASKE to compete in the SemEval-2023 Task 4 with the goal of classifying arguments with respect to human values. Even though our model has median performances compared to others, the results are still promising being it explainable and mostly unsupervised. On the explainability side, there could be possible improvements, such as inspecting decision trees rules to produce local explanation, i.e. how concepts extracted by ASKE are used by the RF classifier to compute predictions. Moreover, we are planning to extend the ASKE component of SuperASKE in order to make it work without an external knowledge base like WordNet. Finally, the SemEval-2023 Task 4 collective results show that the problem of human value detection in arguments is still an open issue that cannot be easily solved by state of the art language models. This motivates us to develop future research on this task by studying NLP systems leveraging pragmatics or by implementing a perspectivist approach in ground truthing to model the subjectivity affecting this type of issues.

7 Acknowledgments

 This project has received funding from the EU Horizon 2020 research and innovation programme under grant agreement No 101004949. This document reflects only the author’s view and the European Commission is not responsible for any use that may be made of the information it contains.

References

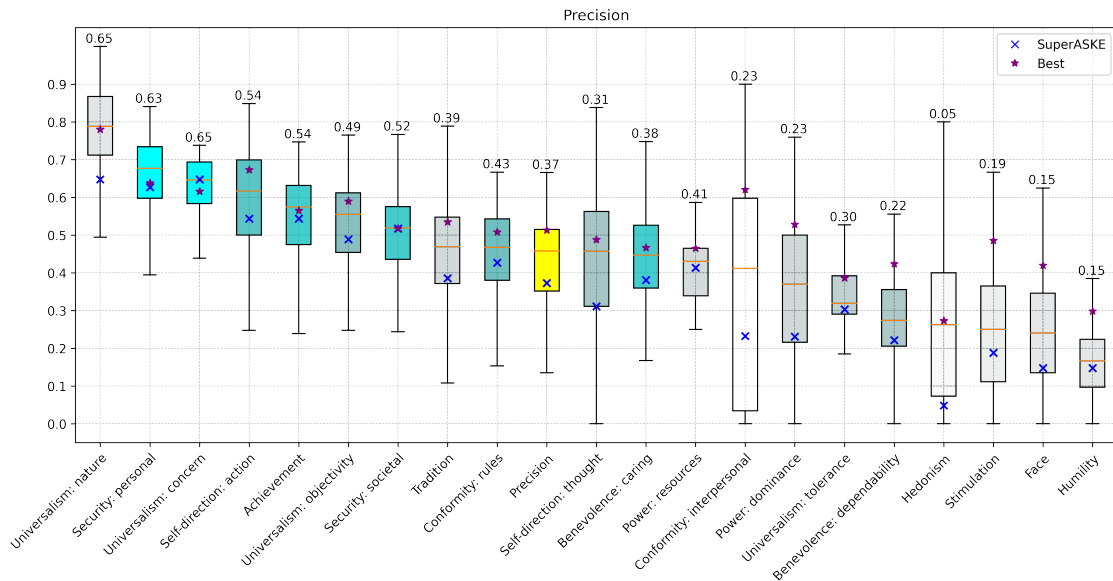
- Leo Breiman. 2001. *Machine Learning*, 45(1):5–32.
- Duane Brown and R. Kelly Crace. 2002. *Life values inventory facilitator’s guide*.
- Haerpfher C., Inglehart R., Moreno A., Welzel C., Kizilova K., Diez-Medrano J., M. Lagos, P. Norris, E. Ponarin, and B. Puranen (eds.). 2020. *World Values Survey: Round Seven - Country-Pooled Datafile*.
- Alfio Ferrara, Sergio Picascia, and Davide Riva. 2023. Context-aware knowledge extraction from legal documents through zero-shot classification. In *Advances in Conceptual Modeling: ER 2022 Workshops, CMLS, EmpER, and JUSMOD, Hyderabad, India, October 17–20, 2022, Proceedings*, pages 81–90. Springer.
- Brendan J Frey and Delbert Dueck. 2007. Clustering by passing messages between data points. *science*, 315(5814):972–976.
- Maik Fröbe, Matti Wiegmann, Nikolay Kolyada, Bastian Grahm, Theresa Elstner, Frank Loebe, Matthias Hagen, Benno Stein, and Martin Potthast. 2023. Continuous Integration for Reproducible Shared Tasks with TIRA.io. In *Advances in Information Retrieval. 45th European Conference on IR Research (ECIR 2023)*, Lecture Notes in Computer Science, Berlin Heidelberg New York. Springer.
- Helen Gouldner. 1975. *THE NATURE OF HUMAN VALUES*. By Milton Rokeach. New York: Free Press, 1973. 438 pp. *Social Forces*, 53(4):659–660.
- 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. 2023a. 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.
- Johannes Kiesel, Nailia Mirzakhmedova, Milad Alshomary, Maximilian Heinrich, Nicolas Handke, Xiaoni Cai, Valentin Barriere, Doratossadat Dastgheib, Omid Ghahroodi, Mohammad Ali Sadraei, Ehsaneddin Asgari, Lea Kawaletz, Henning Wachsmuth, and Benno Stein. 2023b. *Touché23-valueeval*.
- 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.
- Shalom Schwartz, Jan Cieciuch, Michele Vecchione, Eldad Davidov, Ronald Fischer, Constanze Beierlein, Alice Ramos, Markku Verkasalo, Jan-Erik Lönnqvist, Kursad Demirutku, Ozlem dirilen gumus, and Mark Konty. 2012. *Refining the theory of basic individual values*. *Journal of Personality and Social Psychology*, 103:663–88.

A Appendix

A.1 Additional pictures and tables

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
<i>Main</i>																					
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
SuperASKE	.44	.40	.58	.22	.09	.58	.33	.51	.20	.71	.59	.45	.50	.32	.20	.47	.28	.69	.72	.33	.45
<i>Nahj al-Balagha</i>																					
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
SuperASKE	.23	.08	.16	.00	.10	.55	.09	.10	.39	.47	.14	.50	.23	.00	.10	.28	.23	.27	.08	.00	.27
<i>New York Times</i>																					
Best per category	.50	.50	.22	.00	.03	.54	.40	.00	.50	.59	.52	.22	.33	1.00	.57	.33	.40	.62	1.00	.03	.46
Best approach	.34	.22	.22	.00	.00	.48	.40	.00	.00	.53	.44	.00	.18	1.00	.20	.12	.29	.55	.33	.00	.36
BERT	.24	.00	.00	.00	.00	.29	.00	.00	.00	.53	.43	.00	.00	.00	.57	.26	.27	.36	.50	.00	.32
1-Baseline	.15	.05	.03	.00	.03	.28	.03	.00	.05	.51	.20	.00	.07	.03	.12	.12	.26	.24	.03	.03	.33
SuperASKE	.23	.29	.07	-	.29	.4	.14	-	.0	.47	.28	-	.2	.0	.15	.19	.32	.28	.13	.0	.23

Table 2: Achieved F₁-score of team augustine-of-hippo 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. Notice that there are no arguments that resort to "Stimulation", "Power: Resources", or "Tradition" in the New York Times dataset: for this reason, values are substituted with "-".



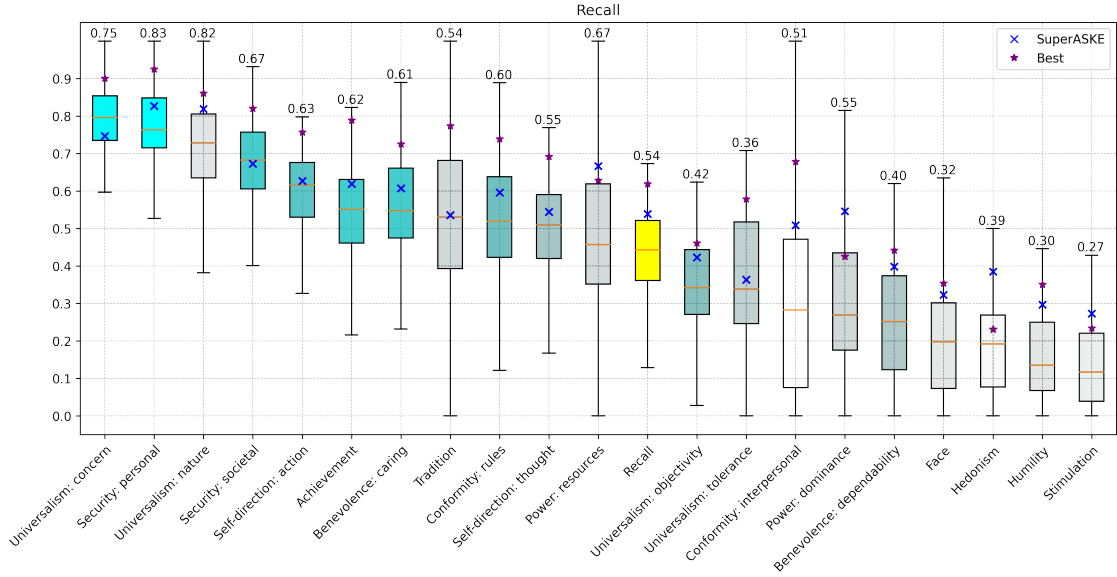


Figure 3: Recall scores distributions for each value category, computed using main test set predictions. Each boxplot represents recall score distribution for all the runs submitted to TIRA. In particular, two points are highlighted in each boxplot: the blue cross depicts SuperASKE’s recall score for the respective category, which is also reported above; the purple star depicts the best model’s recall score for the respective value category. Color intensity is proportional to value category frequency, and it goes from bright light blue, meaning high frequency, to white, meaning low frequency. Finally, plots are ordered according to the median recall score for the respective value category. Macro-average distribution is colored in yellow.

A.2 Confusion matrices

Value	TP	FP	FN	TN
Security: personal	3363	30	0	2000
Universalism: concern	3290	22	1	2080
Achievement	3861	20	0	1512
Benevolence: caring	4037	24	1	1331
Security: societal	3636	29	16	1712
Self-direction: action	3980	18	8	1387
Conformity: rules	4193	23	9	1168
Universalism: objectivity	4306	33	2	1052
Self-direction: thought	4393	12	2	986
Benevolence: dependability	4556	31	22	784
Universalism: tolerance	4715	14	18	646
Power: dominance	4760	23	41	569
Tradition	4806	19	11	557
Power: resources	4736	32	66	559
Universalism: nature	4959	7	3	424
Humility	4991	7	20	375
Face	4793	218	56	326
Stimulation	5114	32	10	237
Hedonism	4990	231	25	147
Conformity: interpersonal	5032	154	44	163

Table 3: Confusion matrices for all value categories, computed using training set labels and training set predictions. TP stands for True Positive, TN for True Negative, FP for False Positive and FN for False Negative.

Value	TP	FP	FN	TN
Security: personal	1131	6	0	759
Universalism: concern	1204	5	0	687
Achievement	1315	6	1	574
Benevolence: caring	1255	8	0	633
Security: societal	1392	16	5	483
Self-direction: action	1392	8	1	495
Conformity: rules	1432	9	7	448
Universalism: objectivity	1517	8	1	370
Self-direction: thought	1643	2	1	250
Benevolence: dependability	1614	14	9	259
Universalism: tolerance	1663	10	7	216
Power: dominance	1712	20	11	153
Tradition	1716	8	6	166
Power: resources	1726	38	10	122
Universalism: nature	1768	1	0	127
Humility	1764	5	12	115
Face	1679	87	26	104
Stimulation	1755	3	11	127
Hedonism	1728	65	27	76
Conformity: interpersonal	1771	65	18	42

Table 4: Confusion matrices for all value categories, computed using validation set labels and validation set predictions. TP stands for True Positive, TN for True Negative, FP for False Positive and FN for False Negative.

A.3 Value categories correlation

True labels correlations: training set			
<i>Value1</i>	<i>Value2</i>	<i>Pearsonr</i>	<i>pvalue</i>
Self-direction: thought	Self-direction: action	0.3	0.0
Stimulation	Hedonism	0.25	0.0
Achievement	Power: resources	0.2	0.0
Predictions correlations: training set			
<i>Value1</i>	<i>Value2</i>	<i>Pearsonr</i>	<i>pvalue</i>
Self-direction: thought	Security: societal	-0.2	0.0
Self-direction: thought	Self-direction: action	0.3	0.0
Self-direction: action	Hedonism	0.21	0.0
Stimulation	Hedonism	0.65	0.0
Face	Conformity: interpersonal	0.53	0.0
Face	Universalism: tolerance	0.25	0.0
Conformity: interpersonal	Universalism: tolerance	0.21	0.0

Table 5: Value categories correlations in training set. On top, the table depicts correlations between value categories found in the ground truth. On bottom, the table depicts correlations between value categories found in SuperASKE predictions.

True labels correlations: validation set			
<i>Value1</i>	<i>Value2</i>	<i>Pearsonr</i>	<i>pvalue</i>
Self-direction: thought	Self-direction: action	0.25	0.0
Stimulation	Hedonism	0.35	0.0
Security: societal	Universalism: concern	0.22	0.0
Predictions correlations: validation set			
<i>Value1</i>	<i>Value2</i>	<i>Pearsonr</i>	<i>pvalue</i>
Stimulation	Face	0.21	0.0
Stimulation	Hedonism	0.69	0.0
Hedonism	Face	0.23	0.0
Power: resources	Security: personal	0.21	0.0
Face	Conformity: interpersonal	0.54	0.0
Face	Universalism: tolerance	0.29	0.0
Security: societal	Universalism: concern	0.23	0.0
Security: societal	Conformity: rules	0.2	0.0
Conformity: interpersonal	Universalism: tolerance	0.21	0.0

Table 6: Value categories correlations in validation set. On top, the table depicts correlations between value categories found in the ground truth. On bottom, the table depicts correlations between value categories found in SuperASKE predictions.