# A Discussion on Open Issues Regarding Human Value Detection in Arguments

Alfio Ferrara[1,†], Sergio Picascia[1,†] and Elisabetta Rocchetti[1,*,†]

[1]Università degli Studi di Milano, Department of Computer Science

Abstract

Human value detection consists in extracting human values from textual data. Being this a complex problem, the Semantic Evaluation 2023 workshop has dedicated one shared task, namely Task 4, to collect contributions and ideas on how to solve human value detection in arguments. This shared task has been organized as a challenge involving multiple teams, each of which have submitted an original solution. In this discussion paper, we present our team submission, reporting the system architecture employed and its performances[1]. By participating in SemEval 2023 Task 4, we noticed that none of the submitted solutions provide satisfying performances, hence we argue this task can still be considered an open issue. Therefore, we share the difficulties we experienced while trying to extract human values from arguments, and we provide a deep discussion on the types of error systems can make in this setting.

**Keywords**

natural language processing, human value detection

## 1. Introduction

Human Value Detection consists in identifying people's beliefs and values in textual data. Its intricate nature derives from the "implicitness" encoded in contextual information, which is difficult to identify and to extract using computational techniques. Due to this complexity, this task has been receiving more and more attention from the Natural Language Processing community. Indeed, human value detection in argumentative texts is one of the shared tasks hosted by the Semantic Evaluation Workshop in 2023. SemEval Task 4 is inspired by [1], the first work introducing human values detection from argumentation texts written in English. Technically, this task consists in producing a multi-label classification system predicting which among the 20 value categories defined in [1] are present in a textual argument. 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 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.

[1]Some sections and information are taken from the paper we submitted to SemEval 2023 Task 4.

*SEBD 2023: 31st Symposium on Advanced Database System, July 02–05, 2023, Galzignano Terme, Padua, Italy*

*Corresponding author.

†These authors contributed equally.

0000-0002-4991-4984 (A. Ferrara); 0000-0001-6863-0082 (S. Picascia); 0009-0000-5617-7612 (E. Rocchetti)

| | |
|---|---|
| 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 |
| $y$: | $[1, 1, 0, 0, \ldots, 0, 1, 0, 0, \ldots, 1, 0, 0, 0]$ |

**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*.

The main dataset is taken from the work by [2], which has 8865 instances; this data is divided in: training set, main validation set and 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. 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 contribution to SemEval Task 4 is an explainable value classification approach called SuperASKE: in this paper, we report the results achieved by our approach, discussing about its performances and providing a detailed error analysis. Furthermore, we want to open a discussion about complications and specificities characterizing human value detection. We argue that this task's complexity is highly related to the nature of the target of classification, namely values categories, which can be considered as implicit information conveyed in arguments. Moreover, it is important to point out that recognizing values in texts is not a trivial task also for human beings.

This work is structured as follows: Section 2 gathers the relevant literature about human values and zero-shot learning classification systems; Section 3 gives details about the proposed system, SuperASKE, and the experimental settings we implemented for SemEval 2023 Task 4 submission; error analysis and discussion can be found in Section 5.

## 2. Related Work

In this section we provide a brief background on human values definitions and schemes; moreover, we show some works done in the context of zero-shot learning (ZLS), which is the technique employed by our system SuperASKE to predict values from arguments.

### 2.1. Human Values

Human values have been studied in social sciences and they can be defined as "a belief pertaining to desirable end states or modes of conduct, that transcends specific situations, guides selection or evaluation of behavior, people, and events, and is ordered by importance relative to other values to form a system of value priorities" [3]. There are several value schemes; in the following, we present only those values from schemes considered by [1].

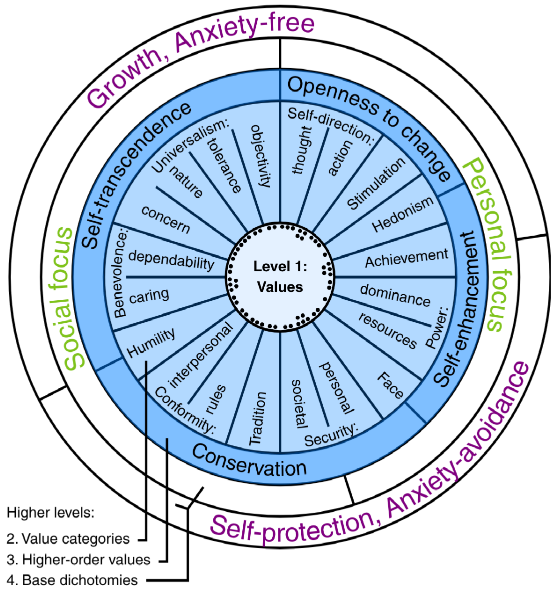


Figure 1: Value taxonomy proposed by [1]. This schema includes 54 values, which can be grouped into more abstract levels and categories. This illustration is adapted from [4].

The Schwartz Value Survey [4] is thought for cross-cultural analysis, and includes 48 values both and the individual level and at the society level. Furthermore, it is possible to establish relations and hierarchies between values according to [3]: this allows for a multi-level value analysis [1]. The Rokeach Value Survey [5] is built upon two concepts: (1)"a value is a belief pertaining to desirable end states or modes of conduct" and (2) "a value system is a prioritization of values based on cultural, social, and personal factors". The authors proposed a system of 36 values spanning desirable end states and desirable behaviour. The Life Values Inventory [6] focus on 14 values for psychological screening for human functioning; in particular, this schema is inspired by [5], but adds some more values such as *Spirituality*. The World Values Survey [7], which has been administered since 1981 in 51 regions, includes 41 values. In this work, we consider the value taxonomy proposed in [1], which is heavily inspired by the Schwartz Value Survey: indeed, 45 of 54 considered values are taken from [4]. The remaining 9 values are picked from the other three schemes mentioned above. Figure 1 depicts this value taxonomy structured in 4 levels, from single values (Level 1) to categorization on the more abstract levels (Levels 2, 3, 4a and 4b).

## 2.2. Zero-Shot Learning

ZSL is a problem setup in the field of machine learning, where a classifier is required to predict labels of examples extracted from classes that were never observed in the training phase. It was firstly referred to as *dataless classification* in 2008 [8] and has quickly become a subject of interest, particularly in the field of natural language processing. The great advantage of this approach consists in the resulting classifier being able to operate efficiently in a partially or totally unlabeled environment. According to [9], ZSL approaches can be classified according to three different criteria: the learning setting, the semantic space and the method. Firstly, the learning setting regards the nature of the analyzed dataset, which can be completely unlabeled, as in the original paper [8], or partially labeled like in [10]; with this last approach, called generalized ZSL, the classifier has to be capable of discriminating between observation of already seen classes and examples of unseen ones. Secondly, the semantic space can be engineered or learned: the former, designed by humans, can be constructed over a set of attributes [11] or a collection of keywords [12], while the latter is built on top of the results of a machine learning model, as in the case of a text-embedding space [13]. Finally, ZSL methods can be divided in instance-based [14], whose focus is on obtaining examples for unseen classes, and classifier-based [15], which instead aim at building a classifier for unlabeled instances.

ASKE [16] (Automated System for Knowledge Extraction) is our framework exploiting ZSL techniques and context-aware embeddings to extract semi-structured knowledge from textual corpora and collecting it in a graph-based data structure, called ASKE Conceptual Graph (ACG), whose nodes represent three kind of entities: document chunks $K$ (portions of the original documents), terms $W$ (n-grams find in the document chunks), and concepts $C$ (cluster of terms).

One of the components in ASKE is the embedding model used for computing the vectorial representations of the member of the ACG. In particular, we choose Sentence-BERT [17], a modification of BERT [18], that is specifically aimed at representing sentence meaning in a vector space. The model is capable of capturing the semantic aspect of the embedded tokens, since it computes context-aware embeddings which, in contrast with global ones, take into consideration the whole chunk of text in which they are used.

## 3. System Overview

The framework we propose for detecting human values, namely SuperASKE (Figure 2), is a concatenation of ASKE and Random Forest [19] with the output of the former being employed as input for the latter. Every instance of this framework is tailored on a single human value, meaning that it solves a binary classification task. The main advantage of SuperASKE is the presence of two explainable models: in ASKE, concepts are described by the terms that compose them with the corresponding definitions; in the RF model, the same concepts are treated as feature, and the importance of each of them in the trees can be computed. 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 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-v2[1]). 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. Each human value $v$ is associated with its SuperASKE instance; within each SuperASKE instance, the ASKE component is initialized with only one concept, representing the respective human value $v$, associated with some dummy terms, having as definitions the ones provided in the value taxonomy. As document chunks $K$, we consider only the premise of each argument, excluding stances and conclusions which appeared to not benefit to the final results. Afterwards, to 'train' ASKE, we run only on the premises positively classified as $v$: in such a way, we ensure

[1]Model available at https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

that the knowledge extracted by SuperASKE is relevant to the human value analyzed. Then, the 'learned' ACG is exploited to compute the similarities between the concepts occurring in it and the whole set of premises, now also including the ones negatively classified. The RF model is employed for the final binary classification. The matrix of similarities $X_{m \times n}$, with $m = |K|$ and $n = |C|$, is provided as input for the model, together with the ground truth $Y$, determining if a premise is classified or not with a given human value. The trained RF model generates the final predictions $\hat{Y}$.

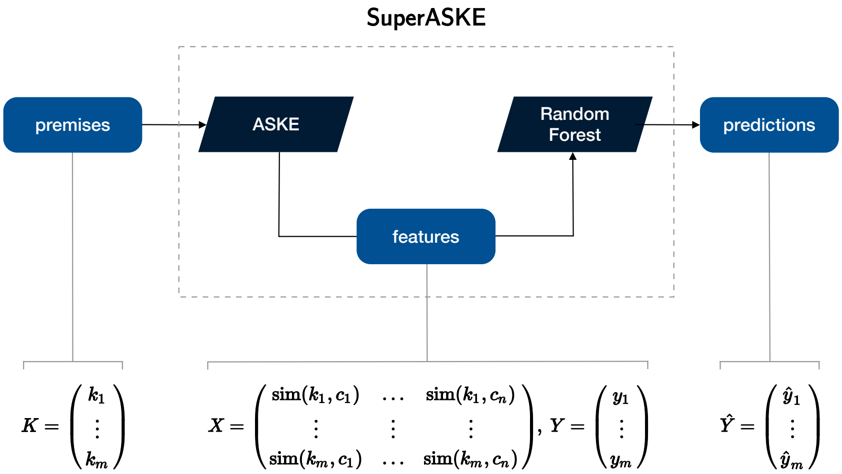


**Figure 2:** The SuperASKE workflow.

## 4. Experiments

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 $\alpha$, $\beta$ and $g$, with $\alpha \in \{-1, ..., 0.5\}$ being the similarity threshold for the zero-shot classification phase, $\beta \in \{-1, ..., 0.5\}$ being the similarity threshold for the terminology enrichment phase, and $g \in \{0, ..., 8\}$ being the number of generations, i.e. the number of ASKE 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 the models of SuperASKE, ASKE 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 the TIRA platform [20].

### 4.1. Results

Figure 3 depicts SuperASKE performances and how they compare to other significant models. Other pictures and tables reporting performance measures are in the Appendix A. SuperASKE's

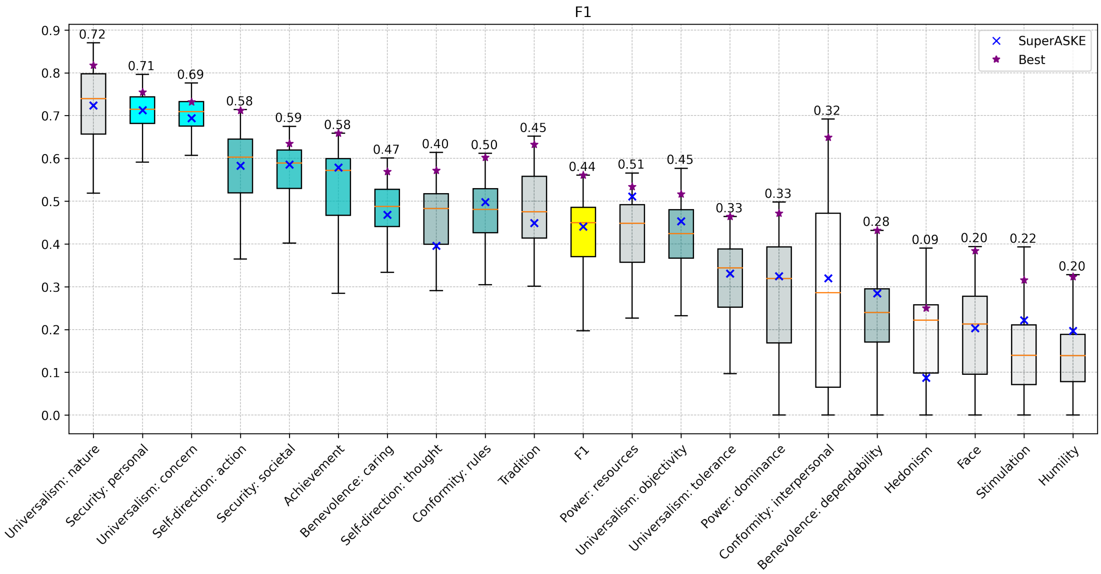


**Figure 3:** F1 scores distributions aggregating all the runs submitted to TIRA for each value category, computed using main test set predictions. The blue cross depicts SuperASKE's F1 score, which is also reported above; the purple star depicts the best model's F1 score. The yellow horizontal line 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.

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. Discussion

**Frequency-performance correlation.** There is a positive correlation between models performance and value categories frequencies in datasets. Considering our model, the correlation ranges from a Pearson statistic of 0.51 with a pvalue of 0.019 (measured on the *test-nahjalbalagha* dataset), to a Pearson statistics of 0.72 with a pvalue of 0.00 (measured on the *main test* dataset). 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", "zoos", "farming", "human cloning", "nuclear weapons".

**Confusion matrices.** Our model's recall is in the third quartile for most distributions (Figure 4): this information can be obtained also through the inspection of confusion matrices for each value category (see Table 3). As a symptom of this behaviour, it can be noticed that false positives frequencies tend to be higher than false negative frequencies. Furthermore, extracting some instances evaluated either as false positives or false negatives in the training set, it is possible to qualitatively categorize the type of error made by SuperASKE. Let's consider the following argument, followed by its true positive and false positive values:

> ***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**: Tradition*
> ***FP**: Security: personal, Security: societal, Conformity: rules, Universalism: concern*

In this argument the main topic discussed is integration for immigrants. Some false positive value categories cannot be associated with this instance, but there could be two exceptions: actually, "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. There are many other examples for which false positive value categories could make sense with the respective premises. This fact highlights how subjective and difficult is to label arguments with human values, thus training NLP models on this type of data leads to predictions for which it is hard to estimate a precise error measure. This problem arises also for true positive values which could be interpreted as false positives depending on the human annotator.

**Value categories correlations.** Comparing values categories correlations found in the ground truth and in the predicted labels, it is possible to study if and how SuperASKE learns correspondences among values. SuperASKE finds a higher number of correlated value-pairs than what is actually existing in the ground truth (7 pairs in predictions versus 3 pairs in true labels); moreover, correlations found in SuperASKE predictions are stronger than what can be seen in the ground truth (e.g. "Stimulation" and "Hedonism").

**Confounded value categories pairs.** Pointing out which value categories are confounded the most with other classes can further explain why SuperASKE produces errors. With the term "confounded" we mean that a value category $X$ is false negative for an instance having another value category $Y$ as false positive. However, as a matter of fact, there are not many instances for which SuperASKE confounds between value categories. In particular, the most confounded value categories for each dataset are: "Conformity interpersonal" and "Face" counting 9 confounded arguments in the validation set; "Face" and "Power: resources" having 13 confounded arguments in the training set. In the following we provide an example of confounding between "Power:

resources" and "Face".

> ***ID***: *A07097*
> ***Conclusion***: *We should ban whaling*
> ***Stance***: *against*
> ***Premise***: *whaling is an important part of the diet, tradition and economies of many countries, and it is not our place to dictate terms as outsiders to their culture.*
> ***True label***: *Power: resources*
> ***False label***: *Face*

In this case, SuperASKE does not recognize the economical perspective of this argument; on the other hand, "whaling" can also be considered as a threat to national public images, depending on whether the opinion on this topic is positive or negative. This latter point of view could offer a valid explanation to why this (and many others) argument is predicted with "Face" value category.

## 6. Conclusion

In this paper we present our solution to human value detection task, focusing on discussing issues about it. There are few points worth mentioning. First, we argue that a mechanism detecting implicit knowledge is needed. Personal human values could lead us to express a specific opinion about a discussed topic: however, arguments do not always embed the values or beliefs causing the stated opinion. For succeeding in solving tasks like human value detection, one must be able to extract causation paths linking implicit knowledge to expressed message. Second, one potential pitfall resides in the annotation procedure. Datasets used to train models are manually annotated. The problem arising here is that detecting human values in arguments is challenging also for humans. This leads to subjective annotations, which make the ground truth less reliable, and a lower-quality training procedure. We argue that a perspectivist approach in ground truthing would help modeling the subjectivity: this could translate into describing, in probability terms, if a certain value could be considered a ground truth label for an argument.

In conclusion, many improvements can be done to get closer and closer to achieving a good solution for human value detection. For instance, trying to map arguments to value categories though value categories descriptions could improve performances. Indeed, aligning value categories and argument terminology could help language models to produce better predictions.

## 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


[1] J. Kiesel, M. Alshomary, N. Handke, X. Cai, H. Wachsmuth, B. Stein, Identifying the Human Values behind Arguments, in: S. Muresan, P. Nakov, A. Villavicencio (Eds.), 60th Annual Meeting of the Association for Computational Linguistics (ACL 2022), Association for Computational Linguistics, 2022, pp. 4459–4471. doi:10.18653/v1/2022.acl-long.306.

[2] N. Mirzakhmedova, J. Kiesel, M. Alshomary, M. Heinrich, N. Handke, X. Cai, B. Valentin, D. Dastgheib, O. Ghahroodi, M. Sadraei, E. Asgari, L. Kawaletz, H. Wachsmuth, B. Stein, The Touché23-ValueEval Dataset for Identifying Human Values behind Arguments, CoRR abs/2301.13771 (2023). URL: https://arxiv.org/abs/2301.13771. doi:10.48550/ARXIV.2301.13771.

[3] S. H. Schwartz, Are there universal aspects in the structure and contents of human values?, Journal of Social Issues 50 (1994) 19–45. doi:https://doi.org/10.1111/j.1540-4560.1994.tb01196.x.

[4] S. Schwartz, J. Cieciuch, M. Vecchione, E. Davidov, R. Fischer, C. Beierlein, A. Ramos, M. Verkasalo, J.-E. Lönnqvist, K. Demirutku, O. dirilen gumus, M. Konty, Refining the theory of basic individual values, Journal of Personality and Social Psychology 103 (2012) 663–88. doi:10.1037/a0029393.

[5] M. Rokeach, The nature of human values., New York, Free Press (1973).

[6] D. Brown, R. K. Crace, Life values inventory facilitator's guide (2002). URL: https://www.lifevaluesinventory.org/LifeValuesInventory.org%20-%20Facilitators%20Guide%20Sample.pdf.

[7] H. C., I. R., M. A., W. C., K. K., D.-M. J., M. Lagos, P. Norris, E. Ponarin, B. P. (eds.), World Values Survey: Round Seven - Country-Pooled Datafile (2020). URL: https://www.worldvaluessurvey.org/WVSDocumentationWV7.jsp.

[8] M.-W. Chang, L.-A. Ratinov, D. Roth, V. Srikumar, Importance of semantic representation: Dataless classification., in: Aaai, volume 2, Citeseer, 2008, pp. 830–835.

[9] W. Wang, V. W. Zheng, H. Yu, C. Miao, A survey of zero-shot learning: Settings, methods, and applications, ACM Transactions on Intelligent Systems and Technology (TIST) 10 (2019) 1–37.

[10] Y. Xian, C. H. Lampert, B. Schiele, Z. Akata, Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly, IEEE transactions on pattern analysis and machine intelligence 41 (2018) 2251–2265.

[11] C. H. Lampert, H. Nickisch, S. Harmeling, Learning to detect unseen object classes by between-class attribute transfer, in: 2009 IEEE conference on computer vision and pattern recognition, IEEE, 2009, pp. 951–958.

[12] R. Qiao, L. Liu, C. Shen, A. Van Den Hengel, Less is more: zero-shot learning from online textual documents with noise suppression, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2249–2257.

[13] Y. Xian, Z. Akata, G. Sharma, Q. Nguyen, M. Hein, B. Schiele, Latent embeddings for zero-shot classification, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 69–77.

[14] X. Xu, T. Hospedales, S. Gong, Transductive zero-shot action recognition by word-vector

embedding, International Journal of Computer Vision 123 (2017) 309–333.
[15] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, M. Ranzato, T. Mikolov, Devise: A deep visual-semantic embedding model, Advances in neural information processing systems 26 (2013).
[16] A. Ferrara, S. Picascia, D. Riva, 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, Springer, 2023, pp. 81–90.
[17] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, arXiv preprint arXiv:1908.10084 (2019).
[18] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).
[19] L. Breiman, Machine Learning 45 (2001) 5–32. doi:10.1023/a:1010933404324.
[20] M. Fröbe, M. Wiegmann, N. Kolyada, B. Grahm, T. Elstner, F. Loebe, M. Hagen, B. Stein, M. Potthast, 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, Springer, Berlin Heidelberg New York, 2023.

# A. 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 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *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 |
| SuperASKE | .44 | .40 | .58 | .22 | .09 | .58 | .33 | .51 | .20 | .71 | .59 | .45 | .50 | .32 | .20 | .47 | .28 | .69 | .72 | .33 | .45 |
| *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 |
| SuperASKE | .23 | .08 | .16 | .00 | .10 | .55 | .09 | .10 | .39 | .47 | .14 | .50 | .23 | .00 | .10 | .28 | .23 | .27 | .08 | .00 | .27 |
| *New York Times* | | | | | | | | | | | | | | | | | | | | | |
| 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 $F_1$-score of team augustine-of-hippo (SuperASKE) 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. 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 "–".

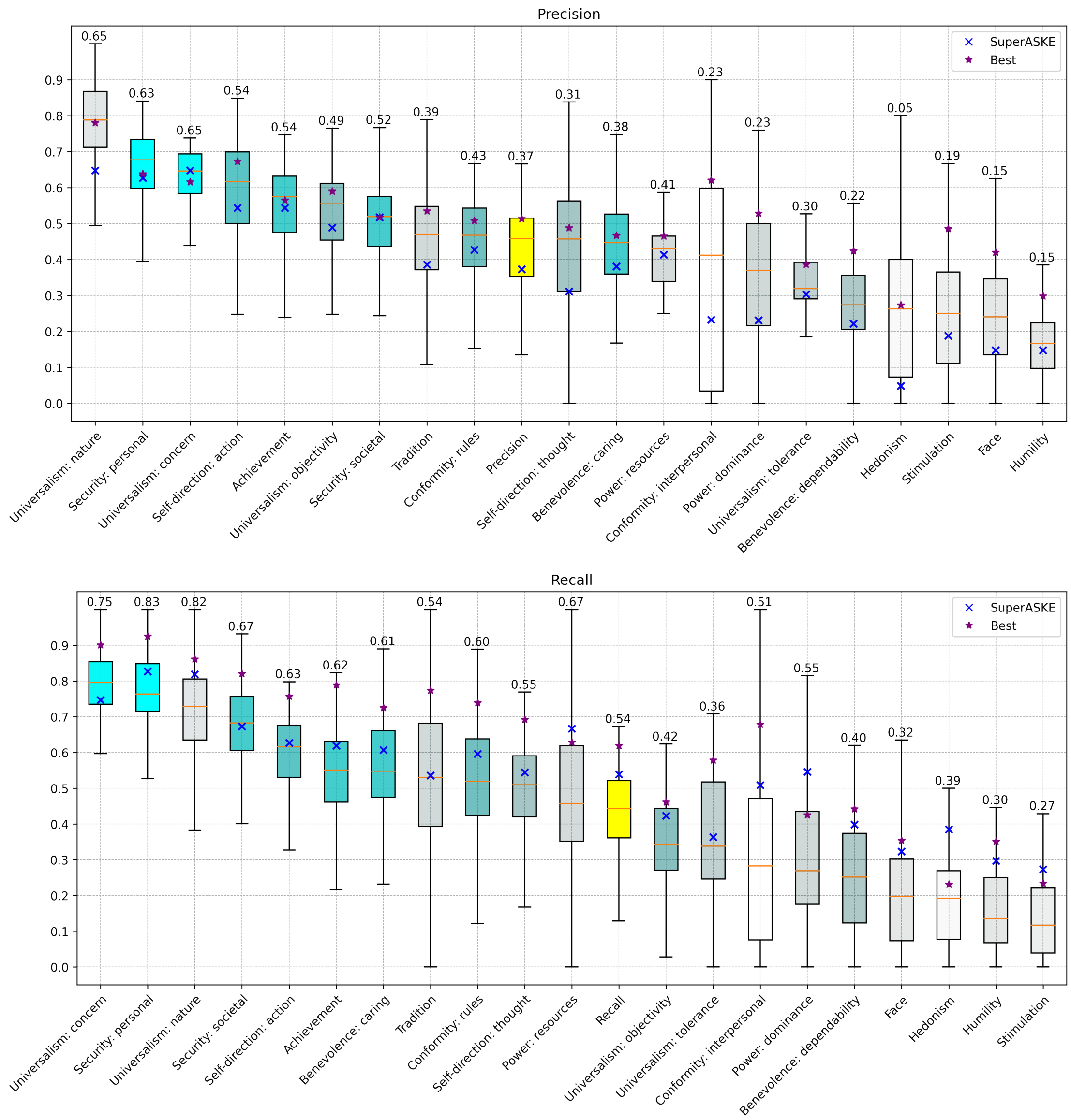


**Figure 4:** Precision scores distributions (top) and recall scores distributions (bottom) for each value category, computed using main test set predictions. Each boxplot represents precision score distribution for all the runs submitted to TIRA. In particular, two points are highlighted in each boxplot: the blue cross depicts SuperASKE's precision score for the respective category, which is also reported above; the purple star depicts the best model's precision 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 precision score for the respective value category. Macro-average distribution is colored in yellow.

| 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 3**
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.