TeamEC at SemEval-2023 Task 4: Transformers VS. Low-Resource Dictionaries, Expert Dictionary VS. Learned Dictionary

Nicolas Stefanovitch
Joint Research Centre
European Commission
nicolas.stefanovitch@ec.europa.eu

Mario Scharfbillig
Joint Research Centre
European Commission
mario.scharfbillig@ec.europa.eu

Bertrand de Longueville
Joint Research Centre
European Commission
bertrand.de-longueville@ec.europa.eu

Abstract

This paper describes the system we used to participate in the shared task (Kiesel et al., 2023), as well as additional experiments beyond the scope of the shared task, but using its data. Our primary goal is to compare the effectiveness of transformers model compared to low-resource dictionaries. Secondly, we compare the difference in performance of a learned dictionary and of a dictionary designed by experts in the field of values. Our findings surprisingly show that transformers perform on par with a dictionary containing less than 1k words, when evaluated with 19 fine-grained categories, and only outperform a dictionary-based approach in a coarse setting with 10 categories. Interestingly, the expert dictionary has a precision on par with the learned one, while its recall is clearly lower, potentially an indication of overfitting of topics to values in the shared task's dataset. Our findings should be of interest to both the NLP and Value scientific communities on the use of automated approaches for value classification.

1 Introduction and Background

This paper describes our participating system to the ValueEval shared task (SemEval 2023 task 4), which can be described as follows : "Given a textual argument and a human value category, classify whether or not the argument draws on that category." (Kiesel et al., 2023). The dataset for this task was built from existing datasets with the aim to cover a wide range of cultural sensitives, while English is the only language considered (Mirzakhmedova et al., 2023).

Automated detection of Human Values is seen by the authors as an important tool in a broader policy analysis toolbox. Indeed, recent research demonstrated that the design and acceptance of public policies is influenced by their underlying values and the way they are framed in values terms (Mair et al., 2019; Scharfbillig et al., 2021).

Our aim in participating in this task was to further explore the possible interplay between dictionary-based techniques - which have policy-related application (e.g. Scharfbillig et al., 2022)- learned dictionary approaches based on keyword extraction (Beliga, 2014) and more state-of-the-art methods involving large language models (Liu et al., 2019). Our contribution consists of:

- Using the task's annotated training set, extracting representative keywords for each of the 19 values classes, and using such learned dictionary to make a classifier using simple keyword-matching heuristics;
- Compare the results with fine-tuned RoBERTa transformers using the same data;
- Compare an expert designed dictionary and a learned one over a coarse grained setting, using only 10 values classes;

By doing so, we aim to stimulate debate and understanding on the respective merits of low/green tech approaches compared to state-of-the-art approaches based on large transformers models for complex classification tasks.

2 Systems and Experimental Setup

We used two approaches: a dictionary-based and a transformer-based. The only training data used was the new data created specifically for this shared task by the organisers. We use an 80/20 train/dev split of the training data, and used the same splits for both approaches. It has to be noted that we evaluated only the dictionary-based approach before the official end of the shared task, and that the transformer-based model was evaluated on the same dataset but outside the official competition.

For the dictionary-based approach, we preprocessed the text by lemmatizing it using NLTK part-of-speech tagger and lemmatizer, and removed

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 2107–2111

the stop words. We computed for each label the distribution over tokens, by splitting over white spaces, and filtering tokens with less than 3 occurrences in a class and removing all non-alphabetical tokens. Then, taking inspiration from TF-IDF we ranked the words: for each word, we took the inverse document frequency, considering each labelled text snippet as a document, and its TF-IDF value for each class by considering the count for that particular class. We then selected for each class the n most important words as measured by their TF-IDF weight. Using this dictionary, in order to perform predictions, we pre-processed the text to be predicted in the same way we pre-processed the training set, then we computed the score of each class, where we count 1 for each matched word. Because there was no easy way to set a threshold and that usually 10+ classes get predicted, we decided to rank the classes by their highest TF-IDF score, and to return the top 4. We observed that in the dataset, the average number of labels was 3.42, which suggests that a fixed number of classes close to it would heuristically optimise the quality. Experiments showed that 4 labels would optimise the score over the dev set. The best results were achieved by using the combination of conclusion and premise and not only the premise. We submitted two runs: for $n = 100$ and $n = 400$ because this approach seems to reach its best value for $n = 500$, that the performance increase with respect to 400 was negligible and we thought it would be better to take a lower value in order to avoid overfitting. These dictionaries had respectively a total of 679 and 2430 words. In another run, we also used the same setting, but split the tokens around stop words instead of spaces, a RAKE-based approach (Rose et al., 2010), which shared about 75% of the vocabulary with the dictionary-based ones, the rest being multi-word expressions.

For the transformer-based approach, we used the base model xlm-roberta-base (Conneau et al., 2019) that we fine tuned using the following hyperparameters: $1e-5$ learning rate, batch size 16 and early stopping with a patience of 3 and using micro F1 score as the optimisation criterion. Experiments using the train-dev set showed that both the F1 and the loss started to rise after 1500 optimisation steps. As such, we made a first submission using a model trained over 1500 steps over the full training dataset. The performance was below expectations, even lower than the dictionary method performance, therefore we submitted a new run by increasing the number of optimisations by a shift of 1000 steps until the model performance started to decrease on the test set. The optimal number of training steps was 4500 - surprisingly high considering our choice based on the dev set, and indicative of a possible severe overfitting of the model to the very specific data of the shared task. In terms of hyperparameter optimisation, we only tried increasing the number of steps, in an attempt to understand the unexpected behaviour of the transformer.

3 Experimental Results

We report in Table 1 the results on the test set of the two approaches and two set of parameters tested. Overall, all our results are at the level of the transformer baseline provided by the organizer or slightly higher.

The first striking result is that the performance of the Roberta classifier is not particularly high, reaching only 45.8 micro F1 at best, when overfitting the data of this shared task. This performance places it slightly above the median performance of runs submitted to the shared task, as there were a total of 111 official submissions. Also surprising was that the optimal number of training steps determined on the train-dev splits used, yielded performance lower than the one of dictionary based approaches. The third surprise was, that the transformer-based approach has a F1 of only 6.1 points above the dictionary-based approach which uses only 100 words per class, as similar previous experiences in a 3-class settings yielded a difference of about 20 points (Stefanovitch et al., 2022) in terms of micro F1. In Table 4 we see that the Roberta 1500 system did not train enough, as 4 of the 19 classes have 0 F1, while Roberta 4500, which is overfitting the test data, is only performing 7 points higher. The best performance of Roberta 4500 is 10 points below the best performing system, indicating significant room for improvement for our systems. However, given these results, and based on the fact that the language used and sentence structure in the data is very homogeneous, we would expect a classifier reaching top performance on this very dataset to have a significant performance drop when applied to other data.

The dictionary-based approaches have performances ranked a bit above the 3rd quartile of the shared task's leaderboard, indicating that they are

system	precision	recall	F1	rank
Dict 100	0.320	**0.514**	0.395	78
Dict 400	0.373	0.432	0.400	79
Rake 100	0.317	0.434	0.366	85
Rake 400	0.373	0.391	0.382	*81*
Roberta 1500	**0.514**	0.312	0.388	*80*
Roberta 4500	0.470	0.447	**0.458**	*48*

Table 1: Experimental results on the test set for the 3 different approaches and 2 different sets of parameters tested, F1 is micro, rank in italics indicates hypothetical ranks, as these submissions are not part of the official leaderboard.

system	precision	recall	$F1_{mic.}$	$F1_{mac.}$
Learned	0.49	**0.67**	**0.57**	**0.48**
Expert	**0.50**	0.29	0.37	0.32
Roberta	0.68	0.65	0.66	0.55

Table 2: Comparison of the performance of an expert designed dictionary and a learned dictionary evaluated on the dev split and mapping the 19 fine grained values to 10 coarse grained values. The performance of Roberta classifier is also provided for reference.

clearly underperforming. Nevertheless, the performance gap with our transformers system is not that significant, and it can even report better recall (for n=100) than our transformer-based approach. Interestingly, a smaller number of words yields a very low precision but a good recall of 0.51, probably competitive with the best system, whose F1 score is 0.56. On top of that, as seen in Table 4, the Dict 100 system has F1 higher for a 1/4 of the class than the Roberta 4500 system. The RAKE-based system did not provide any improvement over the pure word-based dictionary. On the contrary, they introduced a slight performance drop. The tokens in the rake dictionary are much less ambiguous, and are in this sense more interesting from an analyst's point of view. However, they are also more rare and as such not as well suited to building a classifier.

When increasing the number of words, we observe a slight increase of 5 points in precision at the cost of an important decrease of almost 10 points in recall. For the transformer-based approach, the opposite behavior is observed when augmenting the number of training steps: with 1500 steps the precision is at its highest of 0.51, while with 4500 steps there is a drop of about 5 points in precision but 13 points increase in recall.

Lastly, we decided to compare the performance of the expert made dictionary from (Ponizovskiy et al., 2020) and the learned dictionary. The expert dictionary is defined on a coarser 10-class version of the taxonomy. We therefore made the evaluation by integrating the more fine grained 19 values into the original 10 values (Face and Humility were respectively mapped onto Power and Conformity).

4 Results

4.1 Transformers vs. Dictionaries

Table 3 summarises a comparison between the results of the learned dictionary described above, and the expert-designed dictionary from (Ponizovskiy et al., 2020). If a word was shared between several classes, in order to respect the spirit of the expert dictionary, it was associated with the class for which it had the highest TF-IDF value, resulting in potentially less than 100 words per class. On average, the dictionary learned on the shared task data is somewhat smaller (583 words vs. 869 in lemmatized form). Common words in both dictionaries are usually few, the highest overlap being 13 words for Universalism. The matching is highest measured in the expert dictionary for Tradition and Hedonism (Achievement and Universalism), and lowest for Conformity and Security (Stimulation and Conformity). Interestingly, most of the values that are the lowest in their overlap in general (Conformity, Security, Power) are the ones that are least predicted when relating values in text (measured through the dictionary) and self-reported values (measured in a survey).

In Table 4 we can see that "Power: Dominance" class has particularly low performances. In the expert dictionary, the word "power" is used, but appears only once with that intent over all the instances of that word in the training dataset. Oppositely, the word "market" is the highest predictor of "Power: Dominance" class. This indicates maybe some differences, with experts tending to use more abstract concepts and one-word concepts very specific to that class, while learned dictionaries match any word used, potentially overfitting the topics in the dataset. Indeed, in Table 3 we can see that both dictionaries have the same precision, while only the recall of the learned one is higher.

Moreover, several terms in our dictionary, like *"market"*, *"drive"* or *"block"* are fundamentally polysemous and neutral by themselves on the plan of values, and without additional context, that only transformer-based systems are able to capture, can not sufficiently convey a particular value. The fact

	Ach.	Ben.	Conf.	Hed.	Pow.	Sec.	Self-Dir.	Stim.	Trad.
# Words in learned dict	36	32	83	24	74	118	48	15	23
# Words in expert dict	61	77	112	79	86	69	116	102	85
# common words	12	10	4	10	6	8	10	2	11
% of learned covered	25.0%	23.8%	4.6%	29.4%	7.5%	6.3%	17.2%	11.8%	32.4%
% of expert covered	16.4%	11.5%	3.4%	11.2%	6.5%	10.4%	7.9%	1.9%	11.5%

Table 3: Comparison of learned and expert dictionary for coarse grained values

that our dictionary- and transformer-based systems have comparable performances, could be indicative of an overfitting of topics to class, and a possible imbalance in the training data biasing the model in that direction.

Another possible explanation for the low performance difference between transformer- and dictionary-based systems is that the high number of labels for relatively short snippets: average size is 150 characters or 25 words, while half the snippets have 3 labels or less, the other half have up to 8 labels. It could also be that using too many labels tends to confuse the transformers. Testing these hypothesis is up to further research where one would check the correlation between labels used and the consistency of their use.

4.2 Expert vs. Learned Dictionaries

For the coarse-grained experiment, we report in Table 2 the comparison of the learned dictionary Dcit 100, the expert-based dictionary and a Roberta classifier trained on coarse label using the same previous settings. The precision for both dictionaries is roughly the same, while the recall of the Dict 100 is clearly higher: about 40 points above the one of the experts. The performance of the Roberta classifier is higher for coarse-grained than fine-grained, where it has a micro F1 11 points higher than Dict 100, which, However, still has a slightly higher recall.

Given the potential overfitting concerns of the ML models, the comparison to the expert dictionary, which was developed to capture values mentions in any kind of text, hints at the fact that the expert's judgment is still an appropriate approach. However, the low overlap between the learned- and expert dictionaries casts doubt on the conclusion that they may converge over more diverse datasets. Therefore, ML approaches seem to be better able to uncover weak signals for values embedded in arguments. In line with this finding is that for the categories for which the expert dictionary approach is least able to predict the values compared to surveys, the overlap between the ML and expert dictionary is the lowest. At the same time, the overall still low performance of the ML approaches and the fact that many words are used in multiple values with different weights makes the approach difficult to grasp for practitioners (e.g. communicators) who want to learn from the linguistic features directly.

5 Conclusion

We took advantage of this shared task to answer two questions. Firstly, how well do transformer-based solution perform compared to simple dictionary approaches in the values domain? Secondly, how well does an expert designed dictionary perform vs. a learned one? We surprisingly find out that the transformers and dictionaries have the same level of performance on that specific dataset for a fine grained classification task, while performing better, but not drastically better, only in a coarse grained setting. We find that an expert dictionary has a similar precision, albeit lower recall than learned ones. These findings hint at a potential bias of topics toward specific values in the dataset, and call for further study when these models are applied to real world data. We believe this work provides valuable insights both to the NLP and the Value scientific communities on the use of automated approaches for value classification.

References

Slobodan Beliga. 2014. Keyword extraction: a review of methods and approaches. *University of Rijeka, Department of Informatics, Rijeka*, 1(9).

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

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*,

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
Rake 100*	.37	.40	.50	.11	.13	.45	.13	.32	.19	.62	.47	.39	.52	.35	.11	.44	.22	.61	.49	.26	.34
Rake 400*	.38	.42	.51	.00	.00	.47	.05	.36	.14	.65	.51	.47	.47	.07	.08	.41	.15	.66	.58	.27	.30
Dict 100	.40	.42	.49	.07	.11	.52	.17	.33	.14	.68	.53	.43	.48	.35	.21	.44	.29	.68	.54	.34	.45
Dict 400	.40	.42	.50	.04	.13	.53	.17	.36	.10	.68	.53	.47	.48	.21	.17	.46	.25	.68	.56	.34	.37
Roberta 1500*	.39	.48	.61	.07	.13	.59	.15	.36	.00	.74	.60	.54	.39	.00	.00	.51	.00	.73	.72	.12	.20
Roberta 4500*	.46	.57	.64	.15	.21	.60	.31	.49	.22	.71	.57	.49	.47	.32	.26	.46	.20	.74	.76	.31	.44

Table 4: Achieved F_1-score of team johann-georg-walch 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.

Toronto, Canada. 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. *arXiv preprint arXiv:1907.11692*.

David Mair, Laura Smillie, Giovanni La Placa, Florian Schwendinger, Milena Raykovska, Zsuzsanna Pasztor, and Rene Van Bavel. 2019. Understanding our political nature: How to put knowledge and reason at the heart of political decision-making. Scientific analysis or review KJ-NA-29783-EN-N (online),KJ-NA-29783-EN-C (print), Joint Research Centre, Luxembourg (Luxembourg).

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.

Vladimir Ponizovskiy, Murat Ardag, Lusine Grigoryan, Ryan Boyd, Henrik Dobewall, and Peter Holtz. 2020. Development and validation of the personal values dictionary: A theory–driven tool for investigating references to basic human values in text. *European Journal of Personality*, 34(5):885–902.

Stuart Rose, Dave Engel, Nick Cramer, and Wendy Cowley. 2010. Automatic keyword extraction from individual documents. *Text mining: applications and theory*, pages 1–20.

Mario Scharfbillig, Vladimimir Ponizovskiy, Zsuzsanna Pasztor, Julian Keimer, and Giuseppe Tirone. 2022. Monitoring social values in online media articles on child vaccinations. Scientific analysis or review KJ-NA-31-324-EN-N (online), Joint Research Centre, Luxembourg (Luxembourg).

Mario Scharfbillig, Laura Smillie, David Mair, Marta Sienkiewicz, Julian Keimer, Raquel Pinho Dos Santos, Hélder Vinagreiro Alves, Elisa Vecchione, and Laurenz Scheunemann. 2021. Values and identities - a policymaker?s guide. Scientific analysis or review KJ-NA-30800-EN-N (online),KJ-NA-30800-EN-C (print),KJ-NB-30800-EN-Q, Joint Research Centre, Luxembourg (Luxembourg).

Nicolas Stefanovitch, Jakub Piskorski, and Sopho Kharazi. 2022. Resources and experiments on sentiment classification for georgian. In *International Conference on Language Resources and Evaluation*.