Identifying human values in arguments: an alternative approach based on topics detection

Mathias Cardarello Fierro

Università degli Studi di Milano.

Contributing authors: mathias.cardarellofierro@studenti.unimi.com;

Abstract

This work aims to develop an explainable multi-label classification model for labeling new argumentation texts with second-level human values categories. It utilizes the correlation between the topics discussed in the texts, detected using BERTopic, and descriptions of human values, using cosine similarity between arguments and human value embeddings. The performance of the model fitted with a Random Forest algorithm is validated using a test dataset from the SemEval 2023 task organizers, and further discussed focused on the influence of salience tokens present in the textual arguments.

Keywords: BERT, embedding, Human values, Topic modelling, Random Forest, SHAP

1 Introduction

The study of human values behind arguments is a complex task that implies the ability to detect implicit information from textual information, such as people's beliefs, which can vary across different cultures and contexts. According to (Schwartz 1994) a *"value is 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"*.

In recent years, there has been an increasingly attention on automatic human value detection in argument mining, which is a subfield of Natural Language Processing (NLP) that aims to identify and extract argumentative structures of inference and reasoning presented in natural language[1]. In fact, the 17th International Workshop

[1](Stab and Gurevych 2019)

on Semantic Evaluation (SemEval-2023) introduced the shared Task 4 titled *"Identification of Human Values behind Arguments"* to evaluate the performance of different models that automatically identify values within textual arguments.

(Schroter, Dementieva, and Groh 2023) presented the best-performing approach for the task of automatically detecting human values in arguments, consisting of an ensemble of 12 models that have been optimised for either loss minimisation or f1-score maximisation. Although the highest F1 scores for the shared task were obtained by large language models (LLMs)-based works, they were unable to have a deep understanding of more complex common knowledge.

The work is organized as follows: Section 2 provides more background about the task of human values detection from arguments, a formal definition of the problem, and an overview of our proposed approach; Section 3 discuss the performance of the whole framework applied; Section 4 sums up the most important takings of the work, provides a critical discussion on the experimental results, and discuss some ideas for future work. Appendix A includes a detailed description of the experimental methodology, and Appendix B contains some complementary plots. Additional tables and figures can be found in the Jupyter notebook (see Github repository) with the complete analysis results.

2 Research question and methodology

2.1 The taxonomy

To overcome the complexity of detecting human values, (Kiesel, Alshomary, Handke, et al. 2022) describe a multi-level taxonomy with 54 values that is in line with previous frameworks mainly based on Schwartz et al. (2012), Gouldner (1975), Brown and Crace (2002), and C. et al. (2020). In their paper, Kiesel, Alshomary, Handke, et al. outline a value taxonomy consisting of four levels. The first level includes 54 individual values, while the second level groups these values into 20 categories. The third level contains four higher-order values, and the last two levels specify base dichotomies.

2.2 The task

The SemEval 2023 Task 4 is a multi-label classification task that requiered predicting which of the 20 value categories defined in (Kiesel, Alshomary, Handke, et al. 2022) are present in a textual argument that is represented as a triple containing a conclusion, a stance, and a premise. The ultimate aim of the classification model is a vector indicating the presence or absence of each human value in an argument.

The task organizers provided a dataset created by (Mirzakhmedova et al. 2023) of 9,244 arguments in English from 6 different sources, labeled with one or multiple human values, and structured as follows:

The premise is a practical example of a situation for which someone could express an opinion, the stance indicates whether the conclusion statement is in favour or against the sentiment depicted in the premise, and the conclusion conveys an idea

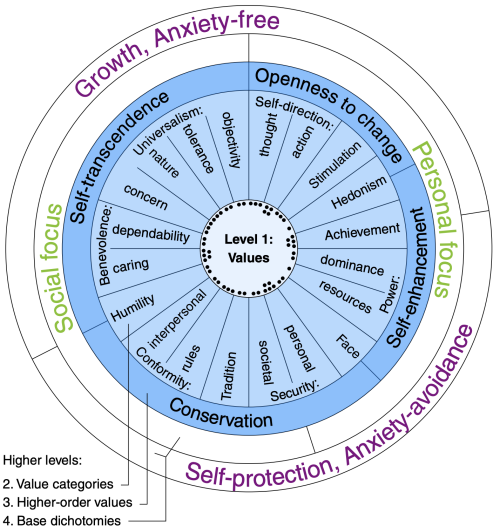

Fig. 1: Human values multi-level taxonomy. Illustration from (Kiesel, Alshomary, Handke, et al. 2022)

Table 1: Example argument about cosmetic surgery

ID	Conclusion	Stance	Premise	Labels
A02004	We should ban cosmetic surgery for minors	in favour of	children have not grown and truly formed their adult looks. cosmetic surgery for a child should be banned until as an adult the child can make a informed choice over their bodies not a whim of a child.	Benevolence: dependability, Universalism: objectivity

according to the respective premise and stance. Additionally, the available value taxonomy described above contains all value categories and their respective values described through sample sentences.

2.3 The methodology[2]

The aim of our work is to create an explainable multi-label classification model that can label new argumentation texts with second-level human values categories from premises, based on the correlation between the topics discussed in the textual arguments and the descriptions of the values. The ultimate goal is a critical discussion of the performance of the methodology, trying to understand whether mentioning certain topics drive the initial prediction towards certain human values.

The framework proposed in this work can be summarized as follows:

[2]See a detailed description in Appendix A

1. **Data preprocessing**: the first step consist in tokenize argumentation texts (premises), remove the stop words and, finally, perform lemmatization and stemming.
2. **Topic modelling**: the idea is to detect probable correlations between the topics discussed in the arguments and the human values, based on their descriptions. To do so we have identified the topics present in the argumentation texts, using the state-of-the-art topic modelling technique BERTopic that provide an efficient and effective way to extract topics from a collection of documents.
3. **Correlation analysis**: with the Cosine similarity between each topic and human value we were able to detect the human values that are more related to each topic, also checking whether there are statistically significant differences in the correlation scores between topics and human values.
4. **Multi-label classification**: in this final step, we fit a multi-label classification model with a Random Forest (RF) using as features the probability distribution over topics for each document (obtained in point 2). As we have imbalanced classes in our dataset, we also used the RandomForestClassifier from `scikit-learn` with balanced class weights, by independently training separate RF classifiers for each of the 20 labels and then combine their predictions.

3 Experimental results

3.1 Exploratory data analysis

After preprocessing the dataset we came out with a list of tokens in its root form for each argument, characterised by the following:

- Along the 7,389 premises we have a total of 5,505 unique tokens, with an average of almost 12 word per argument, with some cases composed by more than 70 tokens.
- The word people, or its stemmed version `peopl`, is by far the most frequent token in the dataset, with more than 1,400 mentions, followed by `would` with a frequency of over 600 cases.
- The training dataset shows a bias towards two specific human values: Security: personal and Universalism: concern, each representing 11.2% of the premises. Conversely, Hedonism and Conformity: interpersonal are underrepresented, each comprising only 1.1% of the premises. This class imbalance needs to be considered while fitting the RF classifier.

3.2 Topics

Initially, the analysis of topic modelling utilized the Non-Negative Matrix Factorization method, resulting in an identification of 35 'optimal' topics based on the Coherence score. However, the relation of some predicted topics to the premises was questionable. Consequently, the BERTopic method was later applied by optimising for Coherence and Silhouette scores. After testing various topic ranges from 2 to 50, settling on 26 topics proved to be a balanced choice, achieving a Coherence Score of 0.5764 and a Silhouette Score of 0.2878. The 857 documents that were detected as outliers by

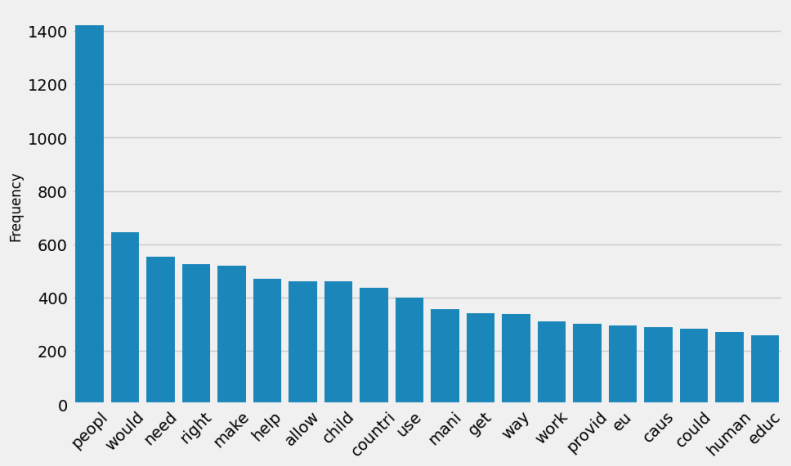

Fig. 2: Most frequent tokens in training dataset

BERTopic and not assigned to any topic in particular, were reassigned to the topic with the highest probability. Table 2 contains an overview of each topic, generated by `ChatGPT 3.5` and derived from the ten most frequent tokens, and their weight on the training dataset.

Table 2: Topics description

N.	Topic (weight)	Description
0	Societal Issues (25.2%)	Covers various societal and political issues, including taxation, drug policies, migration, and refugee matters.
1	European Union (EU), Politics and Economics (11.4%)	Focuses on EU, politics, economics, and sanctions, including aspects of work, military, and intellectual property rights.
2	Justice and Social Activism (7.3%)	Encompasses discussions on crime, justice, activism, and gender-related issues like prostitution.
3	Child Welfare and Education Policies (7.1%)	Covers topics related to child welfare, education policies, family dynamics, and social environments.
4	Surrogacy and Family Dynamics (5.3%)	Discusses surrogacy, family dynamics, and possibly religious or cultural perspectives.
5	Medical Practices and Risks (4.2%)	Involves discussions about medical practices, cosmetic procedures, and potential health risks.
6	Religious Freedom and Diversity (4.0%)	Focuses on religious beliefs, diversity, and societal acceptance of different faiths.
7	Political Ideologies and Systems (4.2%)	Discusses political ideologies, democratic processes, and societal choices.
8	National Security and Warfare (2.8%)	Covers issues related to national security, warfare, and constitutional considerations.
9	Environmental Concerns and Policy (2.6%)	Focuses on environmental concerns, policy debates, and safety considerations.

10	Industrial Practices and Health (3.3%)	Discusses industrial practices, health implications, agricultural impacts, and societal concerns about obesity.
11	Media and Communication Regulations (2.6%)	Covers media, communication, regulations, nuisance factors, and societal attitudes towards media consumption.
12	Information Dissemination and Bias (2.4%)	Focuses on information sources, journalism practices, biases, government involvement, and potential subsidies in media.
13	Financial Practices and Education (2.4%)	Discusses financial practices, student loans, subsidies, costs, and challenges faced by graduates.
14	Biomedical Research and Ethics (2.3%)	Involves debates on biomedical research, cloning, stem cells, ethics, medicine advancements, and funding sources.
15	Education and Intelligence Testing (1.6%)	Covers education, intelligence testing, potential harms, and focus on child development.
16	Language and Cultural Sensitivity (1.5%)	Discusses language, cultural sensitivity, gender-neutral language, and identity considerations.
17	Technological Advancements and Safety (1.5%)	Focuses on technological advancements, safety considerations, and impacts on commuting and disabilities.
18	Healthcare and Bioethics (1.4%)	Covers healthcare, bioethics, organ donation, market dynamics, and ethical considerations.
19	International Conflicts and Humanitarian Issues (1.4%)	Discusses international conflicts, humanitarian issues, and peace initiatives.
20	Technology and Economic Regulation (1.4%)	Involves discussions on technology, economic regulations, monopolies, and market manipulation.
21	Community Support and Societal Welfare (1.2%)	Covers community support, societal welfare, and the importance of collective efforts.
22	Sports and National Identity (1.2%)	Discusses sports, national identity, promotion, and spending on sporting events.
23	Freedom of Speech and Historical Sensitivity (1.3%)	Involves debates on freedom of speech, historical sensitivity, and anti-Semitic sentiments.
24	International Relations and Stability (0.3%)	Covers international relations, stability, and specific geopolitical issues.

3.3 Human values and topics correlation

As a next step, we obtain the embeddings for each human value and argument topic as embeddings, using `roberta-base` pre-trained RoBERTa model, and ensuring that them capture the semantic meaning of the values to be compared to the topics effectively. This allows us to compute Cosine similarity between each topic and human value. However, similarity values are high across all topics (over 0.87), indicating that probably cosine similarity metric may not be effectively discriminating between topics in relation to human values. Therefore, we normalize the cosine similarity scores to a range between 0 and 1, to emphasize the differences in similarity, by using min-max scaling.

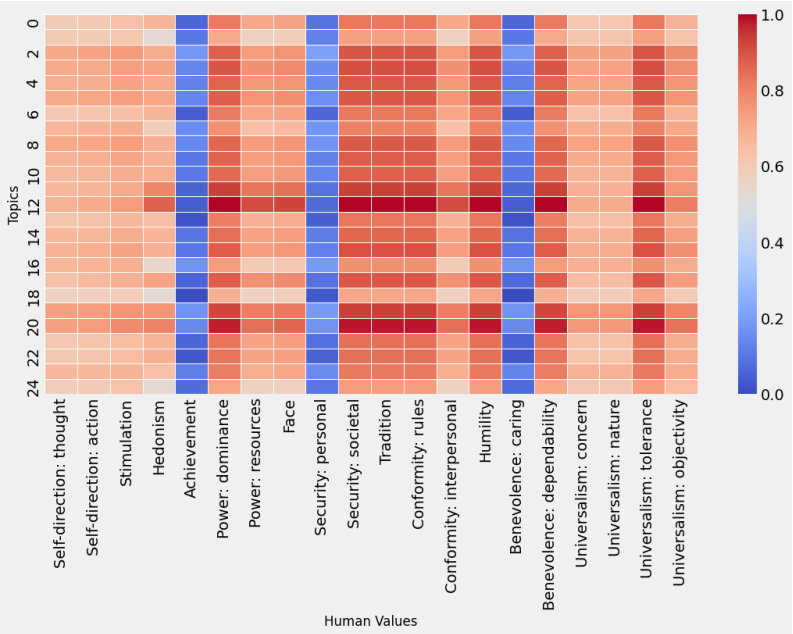

Fig. 3: Scaled Cosine Similarity between topics and human values

Figure 3 shows a heatmap with the Scaled Cosine Similarity between topics and human values we can observe that the values Achievement, Security: personal, and Benevolence: caring, have no strong correlation with any of the topics obtained in the previous step. The one-way ANOVA test verified that there is a significant difference in correlation scores for topics and Human values, while the pairwise t-tests performed to compare each topic or human value individually proved that for all human values there is a significant correlation with at least one of the 25 topics, at 5% significance in both tests.

3.4 Model results

The initial Random Forest (RF) model, which was trained on the original dataset consisting of 7,389 cases and tested on a dataset of 5,154 cases, indicated that the model's performance in precision, recall, and F1-score varied across different classes. The classes with fewer documents showed the poorest results, indicating the need to consider class imbalance to enhance the model's effectiveness.

Therefore, to assess the class imbalance problem, we train a RF binary classifier for each of the 20 labels corresponding to the human values. In this case, the weighted-average precision was 0.80, recall 0.83, and F1-score 0.81. However, the overall results suggest that the model performs well for class 0 but still relatively poorly for class 1. In particular, the classifier could not predict none of the arguments for labels 'Hedonism' and 'Face'. These human values are the ones with the lowest frequencies in the training dataset, representing only 1.1% and 2.1% of the arguments. The best performance was

Table 3: RF overall Classification Report

	Precision	Recall	F1-Score	Support
0	0.88	0.94	0.91	31946
1	0.33	0.19	0.24	5154
Accuracy			0.83	37100
Macro Avg	0.60	0.56	0.57	37100
Weighted Avg	0.80	0.83	0.81	37100

seen for 'Universalism: concern', the most frequent class in the dataset, with nearly 62% of precision at detecting the label.

Table 4: Precision, most correlated, and highest SHAP value topics by human value

Human value	**Precision (%)**	**Topic most correlated**	**Topic with Highest SHAP**
Universalism: nature	61.76	19	8
Security: personal	51.00	2	4
Universalism: concern	38.48	19	1
Universalism: objectivity	36.36	20	12
Security: societal	34.26	12	7
Tradition	33.82	12	5
Achievement	31.75	2	0
Benevolence: caring	28.64	2	0
Self-direction: action	28.57	20	3
Conformity: interpersonal	25.00	12	0
Self-direction: thought	23.39	19	5
Conformity: rules	22.73	12	1
Stimulation	18.18	20	0
Power: dominance	13.04	12	0
Power: resources	11.76	12	0
Benevolence: dependability	10.00	12	0
Universalism: tolerance	8.75	12	0
Humility	3.13	12	0
Face	0.00	12	0
Hedonism	0.00	12	0

It is interesting to note that the seven worst performing classes are the ones that share the same most correlated topic, Information Dissemination and Bias (2.4% of the dataset), based on similarity scores, which includes tokens like `wikipedia`, `inform`, `sourc`, `journalist`, `bias`, `subsidi`, `govern`, `pay`, and `articl`. However, the higher influential topic, based on SHAP contributions[3], is Societal Issues, the most frequent in the dataset (25.2%). The prevalence of Societal Issues in the dataset might result in the model being overly sensitive to this topic, possibly mitigating the importance of other features.

Through a token-level analysis by human value, we can determine which specific input tokens are significant for class prediction. The arguments for the best performing class, Universalism: nature, contain unique tokens such as `anim`, `farm`, `natur`, `nuclear`, `whale` and `zoo` that were only present in that documents. On the other side, most frequent tokens in the worst performing classes do not seem to be that specific for the human values implicit in the arguments, such as `people`, `social`, `child`, and `make`.

3.5 Other approaches

We started the current analysis by creating a Multi-label classification model to predict human values for each argument. To do this, we used a Random Forest algorithm and

[3]See the methodology in Appendix A.4

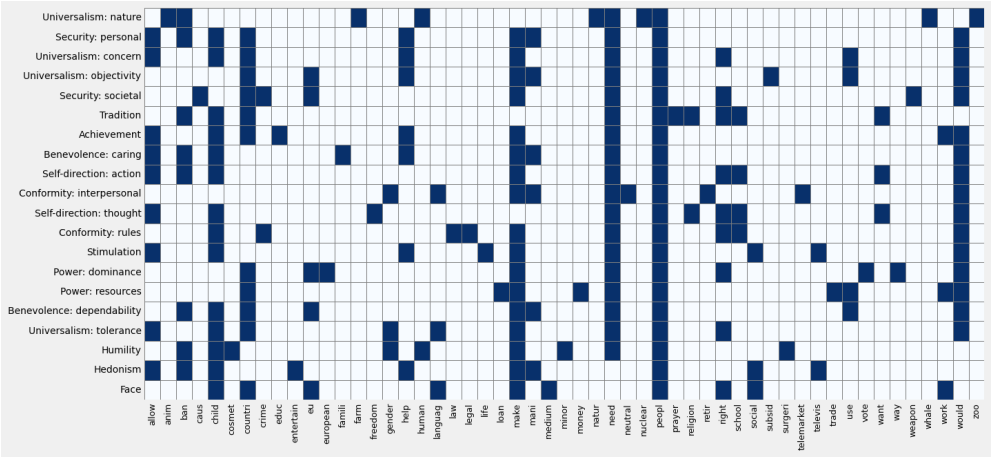

Fig. 4: Most frequent tokens by human value

BERT embeddings of arguments and topics, along with correlation scores between topics and human values as features. The model's overall weighted F1-score was 0.83, indicating that increasing the model's complexity does not significantly enhance its performance. Moreover, using embeddings as features may reduce interpretability since they are dense representations of the tokens' meanings rather than direct mappings to the original text.

4 Concluding remarks

Our approach has revealed that the performance of the model is strongly influenced by the frequency of each class in the dataset, as well as the specificity of the salience tokens present in the arguments. Other factors that may affect the model's performance should also be explored further.

If we decide to stick with a RF model due to its interpretability, it may be necessary to re-evaluate the feature selection process or adjust the model's hyperparameters to reduce the impact of dominant features such as Societal Issues, and to mitigate the potential impact of misleading features related to Information Dissemination and Bias. Furthermore, exploring the data and implementing effective feature engineering techniques could help identify and address any biases in the dataset that may be impacting the model's performance.

Alternatively, if the goal is to optimize for F1 score and prediction capabilities, we should consider applying other more performant approaches based on LLM's, as the ones published in SemEval 2023.

Appendix A System overview

In this section we provide a detailed description of the steps followed to accomplish our goal of making a prediction on argumentation texts, and the subsequent critical discussion of the results.

A.1 Data preprocessing

The premises extracted from the provided dataset of 9,244 labelled arguments from 6 different sources, were preprocessed by converting it to lowercase, removing unwanted characters like URLs and punctuation, tokenizing it into words, removing stopwords, lemmatizing each word to its base form, and stemming the words, using the SnowballStemmer, which reduces words to their root form. The result is a cleaned and standardized representation of the input text, in a list of tokens, ready for further analysis.

A.2 Topic modelling

As the emphasis is on detecting the possible correlations between discussed topics and its associated human value, we identified the topics present in the argumentation texts by using BERTopic. This unsupervised topic modeling technique leverages BERT (Bidirectional Encoder Representations from Transformers), a state-of-the-art pre-trained language model that provides an efficient and effective way to extract topics from a collection of documents. The methodology behind BERTopic involves several key steps:

1. **Embedding generation**: BERTopic first generates embeddings for each document in the corpus using a pre-trained BERT model. These embeddings capture the semantic context of words in the documents.
2. **Dimensionality reduction**: we applied UMAP (Uniform Manifold Approximation and Projection) to reduce the high-dimensional BERT embeddings into lower-dimensional representations. This step helps in visualizing and clustering the documents efficiently.
3. **Clustering**: After dimensionality reduction, BERTopic performs clustering on the reduced embeddings to group similar documents into topics. It utilizes hierarchical clustering to create a topic hierarchy based on the cosine similarity between document embeddings.
4. **Topic representation**: BERTopic represents each topic by identifying the most representative words within the cluster of documents. These representative words help interpret and label the topics.

The methodology involves an iterative process to find the optimal number of topics. It starts by defining a range of potential topic numbers and then fits a BERTopic model to the data for each number of topics in the range. Next, it evaluates the quality of the topics generated by the model using two metrics:

- Coherence Score: measures the semantic similarity between high-scoring words within a topic. Higher coherence scores indicate more coherent and interpretable topics.
- Silhouette Score: assesses the compactness and separation of the clusters formed by the topics. Higher silhouette scores suggest better-defined and well-separated clusters.

By computing and comparing these scores across different numbers of topics, we aim to identify the number of topics that maximizes both coherence and silhouette scores, indicating the most meaningful and well-separated topics for the given dataset. The scores were plotted against the range of topic numbers to visualize the relationship between the number of topics and each metric.

A.3 Correlation analysis

The aim is to identify correlations between the topics discussed in the arguments, obtained in the previous step, and the human values based on their descriptions. To do so, at first, we represented each human value and topic as embeddings generated by a pre-trained RoBERTa base model, ensuring that their representations capture the semantic meaning of the values. RoBERTa, or Robustly optimized BERT approach, is an enhanced version of BERT developed by Facebook AI that is trained on a much larger corpus of text data. These improvements lead to more robust and flexible language representations, rendering RoBERTa highly effective for diverse tasks in natural language understanding.

Then, with the Cosine similarity between each topic and human value we were able to detect the human values that are more related to each topic. Cosine similarity measures how similar two pieces of text are by evaluating their numerical vector representations or embeddings, and calculating the cosine of the angle between these two vectors in a multi-dimensional space. The similarity score ranges from -1 to 1, where 1 indicates perfect alignment, 0 means no similarity, and -1 signifies perfect misalignment.

Ultimately, we examined if the correlation scores between various topics and human values showed statistically significant variations by conducting a one-way ANOVA test across all topics and human values, followed by individual pairwise t-tests for each of them.

A.4 Multi-label classification

Finally, we fit a Random Forest Multi-label classification model with 100 decision trees and a specified random state for reproducibility. Random Forest is a machine learning algorithm that builds multiple decision trees and combines their predictions to improve accuracy. It works by randomly selecting subsets of data and features for each tree, which helps prevent overfitting and increases robustness. The final prediction is determined by averaging or voting among the individual trees' predictions.

The benefit of this approach is that it enables us to comprehend how predictions were made and which characteristics played a vital role in determining the outcome. To achieve this level of interpretability, we use the SHAP (SHapley Additive exPlanations) interpreter. By examining the SHAP values that indicate how much each feature contributed to the prediction outcome, obtained with a subset of 1,000 cases for computational purposes, we can determine which topics had the most significant impact on the model's decision-making process.

Moreover, as we have imbalanced classes in our dataset, we also fit the Random Forest classifier from `scikit-learn` with balanced class weights, by independently

training separate random forest classifiers for each of the 20 labels and then combine their predictions.

Finally, the results are validated with the already defined test dataset provided by the SemEval 2023 task organizers, using macro precision, recall, and F1-score as the evaluation metrics. Precision measures the proportion of true positive predictions among all positive predictions, while recall measures the proportion of true positive predictions among all actual positive instances, and F1 score balances precision and recall to provide a single metric for evaluating the classification model performance.

Appendix B Additional plots

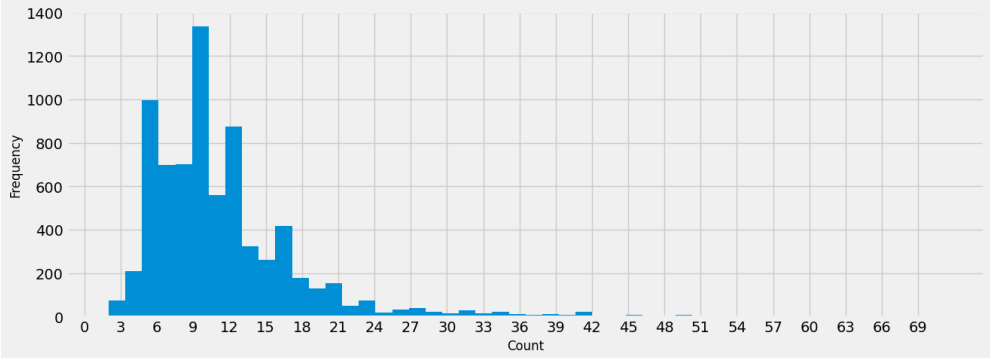

Fig. B1: Histogram of tokens count

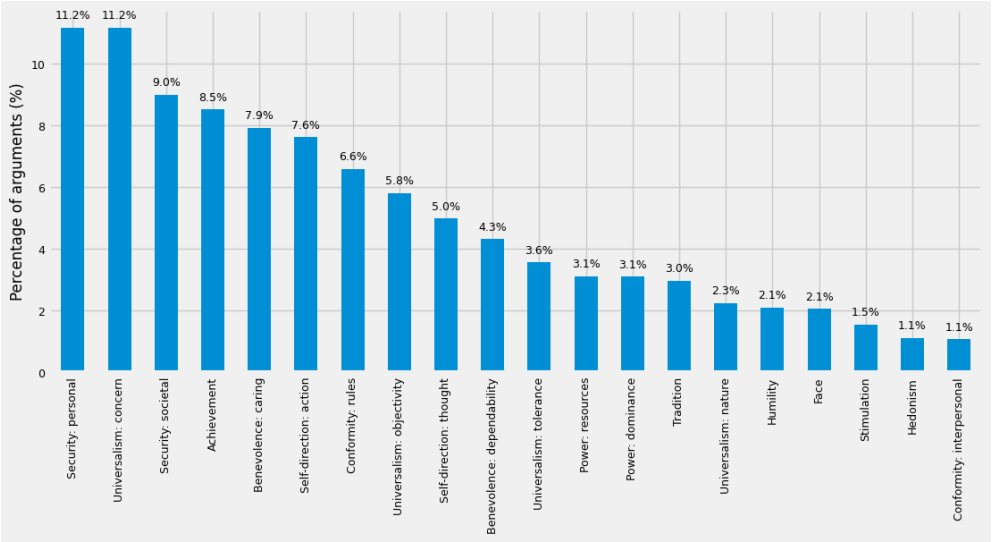

Fig. B2: Class Distribution of Human values

References

Albanese, Nicolo Cosimo (Sept. 2022). "Topic Modeling with LSA, pLSA, LDA, NMF, BERTopic, Top2Vec: a Comparison". In: *Towards Data Science*. URL: https://towardsdatascience.com/topic-modeling-with-lsa-plsa-lda-nmf-bertopic-top2vec-a-comparison-5e6ce4b1e4a5.

Balikas, Georgios (2023). "John-Arthur at SemEval-2023 Task 4: Fine-Tuning Large Language Models for Arguments Classification". In: *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*.

Efimov, Vyacheslav (2023). "Large Language Models: SBERT — Sentence-BERT". In: *Towards Data Science*. URL: https://towardsdatascience.com/sbert-deb3d4aef8a4.

Ferrara, Alfio, Sergio Picascia, and Elisabetta Rocchetti (2023). "Augustine Of Hippo at SemEval-2023 Task 4: An Explainable Knowledge Extraction Method to Identify Human Values in Arguments with SuperASKE". In: *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*.

Jipeng, Qiang et al. (2019). *Short text topic modeling techniques, applications, and performance: A survey*. arXiv: 1904.07695[cs.IR].

Kiesel, Johannes, Milad Alshomary, Nicolas Handke, et al. (May 2022). "Identifying the human values behind arguments". In: *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: Long papers)*. Dublin, Ireland: Association for Computational Linguistics, pp. 4459–4471. DOI: 10.18653/v1/2022.acl-long.306. URL: https://aclanthology.org/2022.acl-long.306.

Kiesel, Johannes, Milad Alshomary, Nailia Mirzakhmedova, et al. (July 2023). "SemEval-2023 Task 4: ValueEval: Identification of Human Values Behind Arguments". In: *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*. Ed. by Atul Kr. Ojha et al. Toronto, Canada: Association for

Computational Linguistics, pp. 2287–2303. DOI: 10.18653/v1/2023.semeval-1.313. URL: https://aclanthology.org/2023.semeval-1.313.

Koehrsen, Will (2018). "Neural Network Embeddings Explained: How deep learning can represent War and Peace as a vector". In: *Towards Data Science*. URL: https://towardsdatascience.com/neural-network-embeddings-explained-4d028e6f0526.

Mirzakhmedova, Nailia et al. (Dec. 2023). "Unveiling the Power of Argument Arrangement in Online Persuasive Discussions". In: *Findings of the Association for Computational Linguistics: EMNLP 2023*. Ed. by Houda Bouamor, Juan Pino, and Kalika Bali. Singapore: Association for Computational Linguistics, pp. 15659–15671. DOI: 10.18653/v1/2023.findings-emnlp.1048. URL: https://aclanthology.org/2023.findings-emnlp.1048.

Murshed, Belal Abdullah Hezam et al. (June 1, 2023). "Short text topic modelling approaches in the context of big data: taxonomy, survey, and analysis". In: *Artificial Intelligence Review* 56.6, pp. 5133–5260. ISSN: 1573-7462. DOI: 10.1007/s10462-022-10254-w. URL: https://doi.org/10.1007/s10462-022-10254-w.

Schroter, Daniel, Daryna Dementieva, and Georg Groh (2023). "Adam-Smith at SemEval-2023 Task 4: Discovering Human Values in Arguments with Ensembles of Transformer-based Models". In: *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*.

Schwartz, Shalom H. (1994). "Are There Universal Aspects in the Structure and Contents of Human Values?" In: *Journal of Social Issues* 50.4, pp. 19–45. DOI: https://doi.org/10.1111/j.1540-4560.1994.tb01196.x. eprint: https://spssi.onlinelibrary.wiley.com/doi/pdf/10.1111/j.1540-4560.1994.tb01196.x. URL: https://spssi.onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-4560.1994.tb01196.x.

Shi, Sheng, Xinfeng Zhang, and Wei Fan (Feb. 18, 2020). *A Modified Perturbed Sampling Method for Local Interpretable Model-agnostic Explanation*. arXiv:2002.07434. type: article. arXiv. arXiv: 2002.07434[cs,stat]. URL: http://arxiv.org/abs/2002.07434 (visited on 11/25/2023).

Stab, Christian and Iryna Gurevych (2019). "Argument Mining: A Survey". In: *Computational Linguistics* 45.4, pp. 765–810. DOI: 10.1162/coli_a_00390. URL: https://direct.mit.edu/coli/article/45/4/765/93362/Argument-Mining-A-Survey.