I2C Huelva at SemEval-2023 Task 4: A Resampling and Transformers Approach to Identify Human Values behind Arguments

Nordin El Balima Cordero, Jacinto Mata Vázquez,
Victoria Pachón Álvarez, Abel Pichardo Estévez
Escuela Técnica Superior de Ingeniería
Universidad de Huelva (Spain)
nordin.elbalima531@alu.uhu.es, mata@dti.uhu.es,
vpachon@dti.uhu.es, abel.pichardo107@alu.uhu.es

Abstract

This paper presents the approaches proposed for I2C Group to address the SemEval-2023 Task 4: Identification of Human Values behind Arguments (ValueEval)" (Kiesel et al., 2023), whose goal is to classify 20 different categories of human values given a textual argument. The dataset of this task consists of one argument per line, including its unique argument ID, conclusion, stance of the premise towards the conclusion and the premise text. To indicate whether the argument draws or not on that category a binary indication (1 or 0) is included. Participants can submit approaches that detect one, multiple, or all of these values in arguments. The task provides an opportunity for researchers to explore the use of automated techniques to identify human values in text and has potential applications in various domains such as social science, politics, and marketing. To deal with the imbalanced class distribution given, our approach undersamples the data. Additionally, the three components of the argument *(conclusion, stance and premise)* are used for training. The system outperformed the BERT baseline according to official evaluation metrics, achieving a *f1 score* of 0.46.

1 Introduction

Human values refers to the beliefs, principles and standards that individuals or groups hold to be important and worthwhile. These values guide people´s attitudes and behaviors; they can vary across cultures (Civitillo et al., 2019), communities, and individuals. According to a study of *Global Values Survey* (White et al., 2020), the most widely held values across the world are a sense of community, a sense of national pride, and a desire for social order. Other commonly held values include equality, respect for others, and a desire for a peaceful world.

In this context, the classification of human values in textual arguments is an important task in the field of *Natural Language Processing*. Understanding the values that underlie an argument could provide valuable insights into people´s beliefs, attitudes, and motivations. It could also be useful in various applications such as opinion analysis (Hemmatian and Sohrabi, 2019), argumentation mining, emotion recognition, and persuasive technology, among others.

Despite its importance, the automatic classification of human values in arguments remains a challenging problem. The task requires the ability to identify values in a text, understand the argument structure, and make a binary judgment about the presence of a value in the argument. It aims to advance the state of the art in human value classification in textual arguments , these textual arguments are compiled from the social science literature and described in detail in the accompanying ACL paper (Kiesel et al., 2022).

ValueEval had the advantage of bringing together 40 teams, taking in 112 different runs *(including the competition organizers)*. They provided a set of labelled data to solve the task. This dataset was highly imbalanced over the 20 categories, with a large number of instances belonging to the negative class. Undersampling (Arefeen et al., 2020) is a common technique used in machine learning to balance the class distribution in imbalanced datasets. To tackle this problem, we employed an undersampling strategy to decrease the number of instances within the negative class. On the other hand, we conducted experiments using various combinations of the information provided for the arguments.

Finally, our approach uses the premise, conclusion and stance of the argument as input features. These three components provide important information about the argument to help identify the values. In this work, we utilized transfer learning techniques by fine-tuning state-of-the-art pre-trained language models, using the transformers library. This approach allowed us to leverage the knowl-

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 1382–1387

edge learned from large-scale datasets and apply it to our specific task of argumentative text classification.

The results of our experiments show that our approach is effective in classifying arguments based on human values. The combination of undersampling and the use of all the data given leads to improved performance compared to other methods. Our findings contribute to a better understanding of how human values can be identified in arguments and have implications for a range of applications, including opinion analysis and argumentation (Lawrence and Reed, 2020). However, as is common in such tasks, some categories have a low number of positive instances, i.e. instances where the argument draws on that category. This low number of positive instances can pose a challenge for machine learning algorithms, as they may not have enough examples to learn from. This can result in overfitting, where the algorithm memorizes the training data instead of generalizing to new data, therefore the model may not be able to accurately predict the outcome for new instances.

2 Background

The input data for the study consists of two tab-separated value files, "arguments-training.tsv" and "labels-training.tsv". Both contains 5,394 rows, on arguments each row represents a unique argument with its ID, conclusion, stance and premise, Figure 1 shows an example of an argument from the training dataset. . For the label each row corresponds to the unique argument ID, and one column for each of the 20 value categories, indicating whether the argument aligns with the particular value category (1) or not (0). The dataset used in this paper was extracted from the descriptions of articles in arXiv, as described in the source (Mirzakhmedova et al., 2023).

[Argument ID] *A01010*
[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.*

Figure 1: Example of argument from training dataset

The identification of human values behind arguments is an important aspect of argument mining. Some work have been researched in this area, which aims to extract natural language arguments and their relations from text (Cabrio and Villata, 2018), there are a lot of use cases like (Passon et al., 2018) predicting the usefulness of online reviews based solely on the amount of argumentative text that they contain, or finding relevant evidence (on argument premises) in the study of adjudication decisions about veteran´s claims for disability (Walker et al., 2018)

3 System Overview

The task of classifying textual arguments based on human values categories is challenging due to the subjective nature of human values. However, the system was able to address this challenge by using advanced deep learning algorithms such as BERT and RoBERTa.

3.1 Implemented Models

For our argument classification task, we employed the BERT and RoBERTa models. Both of these models are based on the transformer architecture and have been pre-trained on massive amounts of text data. BERT, short for Bidirectional Encoder Representations from Transformers, was introduced by (Devlin et al., 2019) and has achieved state-of-the-art performance on various natural language processing tasks. RoBERTa, a variant of BERT, was introduced by (Liu et al., 2019) and further improved the pre-training process by optimizing hyperparameters and using larger batch sizes. We used the Hugging Face[1] library to fine-tune the pre-trained BERT and RoBERTa models for our task. We employed the common hyperparameters, including batch size, learning rate, and weight decay, and used early stopping with a maximum of five epochs. We did not tune any specific hyperparameters.Our choice of these models was based on their proven success in various natural language processing tasks and their pre-training on large amounts of text data. We fine-tuned the models on our argument classification task to leverage their ability to understand and extract meaningful information from natural language text. Recent studies have demonstrated the effectiveness of these models in various tasks, such as sentiment analysis (Khan and Fu, 2021), question answering (Ju et al., 2019), and document classification (Liu et al., 2021).

[1]https://huggingface.co/

Class	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
0	1264	1786	990	1104	1934	781	800	857	2560	2211	726	1507	830	885	1706	1032	2664	958	850	1349
1	790	1116	198	138	1209	488	500	306	1600	1382	454	942	166	316	1066	645	1665	342	531	843

Table 1: Number of examples of each class after performing *Undersampling*.

3.2 Selected Inputs

We experimented with different input formats for our models, including using just the conclusion or just the premise as input. However, we found that using the conclusion stance and premise together as a single input yielded the best results in our experiments. This is because the conclusion stance provides important context for the premise, allowing the model to better understand the argument being made. To illustrate the impact of using different input formats, we show the F1 scores for each format in Table 2. As can be seen, the model using the combination of conclusion stance and premise as input achieved the highest F1 score.[2]

C+S+P	Premise	Conclusion
0.73	0.71	0.56

Table 2: Impact of Input Format on F1 Scores in minor class of label *"Security: Personal"*

3.3 Preprocessing Text

We have applied text preprocessing to clean and simplify the text:

- Conversion of all characters to lowercase: All the characters in the text were converted to lowercase for consistency and ease of processing.
- Expansion of all possible contractions in English: Contractions such as "don't" were expanded to "do not" and "can't" was expanded to "cannot" to ensure that the model could understand the text properly.
- Removal of special characters: Special characters such as punctuation marks and symbols were removed from the text. This helped to simplify the text and remove any unnecessary noise.
- Removal of multiple spaces between characters: Multiple spaces between characters were reduced to a single space. This was done to ensure that there was consistency in the text and that all the spaces were uniform.

By applying these text processing techniques, we were able to simplify and clean the text, making it easier to process and analyze.

3.4 Undersampling Techniques

In order to address the issue of class imbalance in the provided training dataset, we implemented an undersampling technique that reduces the size of the majority class to increase the representation of the minority class. Specifically, we used a multiplier on the size of the majority class to determine the number of samples to keep for each label, with the multiplier being determined by the size of the minority class. To ensure the models had sufficient data to train on, we trained at least with 1000 arguments. Table 1 shows the distribution of argument categories after performing undersampling. Overall, these preprocessing steps helped improve (see Table 3) the models' performance on the data.

Original dataset	With Undersampling
0.53	0.57

Table 3: Impact of Undersampling on F1 Scores in minor class of label *"Security: societal"*

[2]Note C+S+P stands for Conclusion+Stance+Premise.

4 Experimental Setup

For our experimental setup, we used a train-validation-test split to evaluate the performance of our models. We split the dataset into 80% for training, 10% for validation, and 10% for testing. This allowed us to train our models on a sufficiently large amount of data while still having a separate set of data to test the models' generalization ability. We chose not to use the validation dataset during the training process because its arguments were substantially different from those in the test dataset, which could have affected the generalization ability of our models. Therefore, we solely relied on the training and test datasets for all the experiments. We used the PyTorch library and the transformers package to implement our models. We used the Hugging Face Transformers library to fine-tune pre-trained transformer models for the argument classification task. TIRA (Fröbe et al., 2023) is the platform used for the shared task that we submitted our system to.

For the experiment, the models were trained with 5 epochs, 32 batch size, 64 token length. Early stopping was used to avoid overfitting while training. Labels like *"Stimulation"* only left us with 198 positive examples and 4116 negatives for our training dataset. This made the identification of values a more difficult task.

By balancing it to an equal number of *0s* and *1s*, we have observed worse results compared to the original one. Then, to improve our approach which is identifying the human value, the dataset has been adjusted to bring at least 1000 arguments for training, which has shown better results as we can see in Table 4, our model gave us a *56%* improvement over the original one.

Dataset	Class 1	Class 0
Original	0.16	0.98
Equally balanced	0.15	0.78
With undersampling	0.30	0.94

Table 4: Impact on F1 Scores of label *"Stimulation"*

In addition to that, we have more balanced labels like *"Self-direction: action"* with 1116 positive examples and 3198 negatives. With our undersampling approach (see Table 5) we improved the result by *18%*.

Dataset	Class 1	Class 0
Original	0.51	0.85
Equally balanced	0.58	0.82
With undersampling	0.60	0.86

Table 5: Impact on F1 Scores of label *"Self-direction: action"*

5 Results

Our system achieved performance above the BERT baseline (shown in Figure 2), as measured by official evaluation metrics, we got better results overall. But the biggest gap is in more imbalanced classes like *"Stimulation", "Humility" and "Face"* where we nearly doubled the performance. Table 6 shows the overall results. In the competition, there were 39 participating teams. Our best-performing system was a *BERT* pretrained model with resampling. This approach achieved a *f1 score* of 0.46, obtaining the 24th position in the final leaderboard, demonstrating its effectiveness.

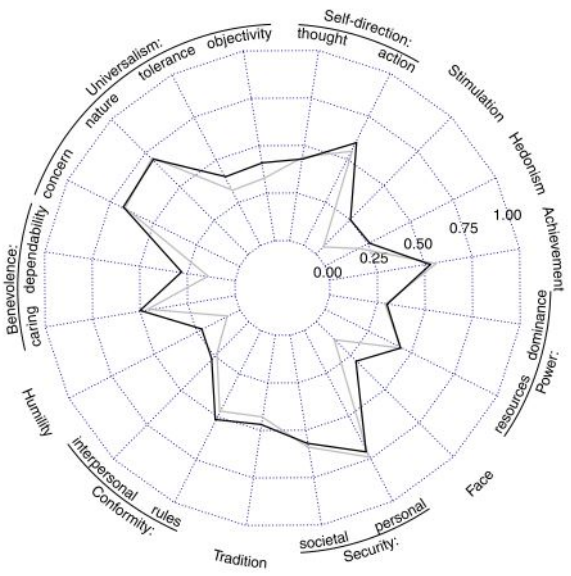

Figure 2: BERT *(grey)* baseline comparison with our best approach *(black)*

6 Conclusion

To conclude, our system for detection of human values behind arguments achieved competitive results, as it presents better results than the baseline BERT model. Our approach of utilizing undersampling to balance the dataset and focusing on the positive class proved to be effective, especially given

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
BERT (1st run)	.46	.43	.60	.25	.26	.53	.30	.44	.29	.71	.57	.47	.52	.27	.22	.50	.28	.68	.70	.40	.41
roBERTa (2nd run)	.41	.32	.58	.29	.18	.39	.26	.49	.14	.71	.58	.50	.48	.00	.07	.46	.23	.68	.73	.32	.52
BERT (3rd run)	.45	.42	.59	.23	.25	.58	.33	.46	.28	.70	.59	.43	.50	.28	.20	.49	.27	.70	.68	.38	.47

Table 6: Achieved F_1-score of team marquis-de-sade 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.

the nature of the competition where identifying it is more crucial than not identifying it. However, there is still room for improvement. One potential direction for future work is to explore data augmentation techniques to increase the size of our training arguments and potentially improve model performance. There is also the possibility of extending our system to other languages, which would require additional preprocessing and potentially the development of language-specific models. Finally, investigating the performance of other transformer-based models and their variants on this task could also lead to better results in future researches.

7 Acknowledgments

This paper is part of the I+D+i Project titled *"Conspiracy Theories and hate speech online: Comparison of patterns in narratives and social networks about COVID-19, immigrants, refugees and LGBTI people [NON-CONSPIRA-HATE!]", PID2021-123983OB-I00, funded by MCIN/AEI/10.13039/501100011033/ and by "ERDF A way of making Europe"*.

References

Md Adnan Arefeen, Sumaiya Tabassum Nimi, and M Sohel Rahman. 2020. Neural network-based undersampling techniques. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 52(2):1111–1120.

Elena Cabrio and Serena Villata. 2018. Five years of argument mining: A data-driven analysis. In *IJCAI*, volume 18, pages 5427–5433.

Sauro Civitillo, Linda P Juang, Marcel Badra, and Maja K Schachner. 2019. The interplay between culturally responsive teaching, cultural diversity beliefs, and self-reflection: A multiple case study. *Teaching and Teacher Education*, 77:341–351.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

Fatemeh Hemmatian and Mohammad Karim Sohrabi. 2019. A survey on classification techniques for opinion mining and sentiment analysis. *Artificial intelligence review*, 52(3):1495–1545.

Ying Ju, Fubang Zhao, Shijie Chen, Bowen Zheng, Xuefeng Yang, and Yunfeng Liu. 2019. Technical report on conversational question answering. *arXiv preprint arXiv:1909.10772*.

Zaid Khan and Yun Fu. 2021. Exploiting bert for multimodal target sentiment classification through input space translation. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3034–3042.

Johannes Kiesel, Milad Alshomary, Nicolas Handke, Xiaoni Cai, Henning Wachsmuth, and Benno Stein. 2022. Identifying the Human Values behind Arguments. In *60th Annual Meeting of the Association for Computational Linguistics (ACL 2022)*, pages 4459–4471. Association for Computational Linguistics.

Johannes Kiesel, Milad Alshomary, Nailia Mirzakhmedova, Maximilian Heinrich, Nicolas Handke, Henning Wachsmuth, and Benno Stein. 2023. Semeval-2023 task 4: Identification of human values behind arguments. In *17th International Workshop on Semantic Evaluation (SemEval-2023)*.

John Lawrence and Chris Reed. 2020. Argument mining: A survey. *Computational Linguistics*, 45(4):765–818.

Yang Liu, Hua Cheng, Russell Klopfer, Matthew R Gormley, and Thomas Schaaf. 2021. Effective convolutional attention network for multi-label clinical document classification. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5941–5953.

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*.

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.

Marco Passon, Marco Lippi, Giuseppe Serra, and Carlo Tasso. 2018. Predicting the usefulness of amazon reviews using off-the-shelf argumentation mining. *arXiv preprint arXiv:1809.08145*.

Vern R. Walker, Dina Foerster, Julia Monica Ponce, and Matthew Rosen. 2018. Evidence types, credibility factors, and patterns or soft rules for weighing conflicting evidence: Argument mining in the context of legal rules governing evidence assessment. In *Proceedings of the 5th Workshop on Argument Mining*, pages 68–78, Brussels, Belgium. Association for Computational Linguistics.

Cindel White, Michael Muthukrishna, and Ara Norenzayan. 2020. Worldwide evidence of cultural similarity among co-religionists within and across countries using the world values survey. *PsyArXiv. https://psyarxiv. com/uetg6*.