Friedrich Nietzsche at SemEval-2023 Task 4: Detection of Human Values from Text Using Machine Learning

Abdul Jawad Mohammed
M.S. Computer Science
Purdue University
Fort Wayne, Indiana
mohaa05@pfw.edu

Sruthi Sundharram
M.S. Computer Science
Purdue University
Fort Wayne, Indiana
sunds01@pfw.edu

Sanidhya Sharma
M.S. Computer Science
Purdue University
Fort Wayne, Indiana
shars05@pfw.edu

Abstract

Literature permeates almost every facet of our lives, whether through books, magazines, or internet articles. Moreover, every piece of written work contains ideas and opinions that we tend to relate to, accept or disregard, the debate over, or enlighten ourselves. However, the existence of subtle themes that are difficult to discern inspired us to utilize four machine learning algorithms: Decision Trees, Random Forest, Logistic Regression, and Support Vector Classifier (SVC) to aid in their detection. Trained on the ValueEval data set as a multi-label classification problem, the supervised machine learning models did not perform as well as expected, with F1 metrics hovering from 0.0 to 0.04 for each value. Noting this, our paper discusses our approach's limitations and weaknesses.

1 Introduction

The creation of language remains one of humanity's greatest achievements, through which concepts, knowledge, stories, memories and emotions are encapsulated into words and conveyed across generations. Among the multitude of ideas contained within the ever-growing expanse of written texts, whether it be news articles, novels, magazines or philosophical and religious scripture, there exists a commonality: The commentary on humanistic archetypes of freedom, creativity, self-preservation, etc. upon which rational beings base their actions and thoughts on. These notions are expressed textually in varying degrees, sometimes apparent while at other times too subtle and difficult for the layman to recognize. For example, advertisements and political campaigns try to identify with potential customers and voters by openly appealing to their human values such as religious beliefs and concerns about nature and the environment. At other times, however, ambiguous and obscure literature could foster misunderstandings, misinterpretations, and even conflict. Therefore, detecting and determining human values in texts could serve a potent role in many Natural Language Processing applications, including the categorization of text information for tag-based recommendation systems, opinion/stance detection in articles, and analysis of complex literature written by poets and philosophers. As part of the SemEval 2023 ValueEval task (Kiesel et al., 2023), we seek to explore the effectiveness of supervised earning classifiers for semantic analysis through the utilization of algorithms like Decision Trees, Logistic Regression, Support Vector Machines, and Random Forest in the extraction of implicit themes within the literature. With most research on semantic analysis being conducted using more powerful transformer/neural network based models, our paper chose to focus on more traditional machine learning techniques for their potential in being integrated with lightweight, portable recommendation systems and software based on thematic analysis of literature. Due to its vast array of libraries for machine learning/Data analysis/Natural Language Processing applications, Python was the primary language used for our study.

With the problem being of multi-label nature and the SemEval2023 dataset as our main subject of analysis, the data features underwent term frequency-inverse document frequency (TF-IDF) vectorization before being fed to ensembles of each machine learning model for training in order to predict 20 human value themes on the test data. The hyper-parameters for each model were also tuned to inspect for changes in performance.

Ultimately, we observed that the aforementioned techniques did not generally perform as well as their Deep learning counterparts, as reflected in our placing last in the rankings for scoring the main test set provided by SemEval. The limitations of the TF-IDF vectorization, coupled with the lower complexity of machine learning algorithms compared to neural networks, hindered the overall scoring

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 2179–2183
July 13-14, 2023 ©2023 Association for Computational Linguistics

performance of our ensembles. Nevertheless, the paper's primary purpose is to serve as both a cautionary and foundational study for future research on this subject.

2 Background

Related Work: Human values are of concern to most if not to all social sciences (Rokeach, 1973) and have also been integrated into computational frameworks of argumentation (Bench-Capon, 2003). In NLP, values have been analyzed for personality profiling (Maheshwari et al., 2017), but not yet for argument mining, as considered here. There has been significant progress recently in the creation of data sets for human values. One such data new data set is the ValueNet data set which contains human attitudes on 21,374 text scenarios (Qiu et al., 2022). ValueNet categorizes the text scenarios into 10 categories, namely, Universalism, Benevolence, Conformity, Tradition, Security, Power, Achievement, Hedonism, Stimulation, and Self-direction. Their work employs the data set to train a Transformer based regression model and apply it to dialogue systems.

The ETHICS data set (Hendrycks et al., 2020) is a new benchmark that can predict moral judgments. They also show to assess a language model's knowledge of morality.

The creation and evaluation of these data sets is a significant step in setting up benchmarks for human emotion and they pave the way for training models and subsequently training intelligent agents guided by ethically sound directives.

Dataset Information:

Two datasets, one labelled for model refinement and the other being an unlabelled test dataset for the final submission, were provided by the SemEval Task 2023, containing arguments categorized under 20 labels, each representing one or more humanistic themes as shown in Figure 1. An argument sample consists of three attributes:

- *Premise*: A text feature showcasing the main argument.
- *Conclusion*: A text feature representing the conclusion inferred from the Premise.
- *Stance*: Value indicating if the conclusion is in 'favor of' or 'against' the premise.

Across the labelled data set, we observed words like 'abolish', 'ban', 'adopt', and 'legalize' were among the most frequent non-stop words for all categories, suggesting that a great portion of the conclusion-premise pairs revolved around legislation, moral discussions, and societal issues as seen in Figure 2.

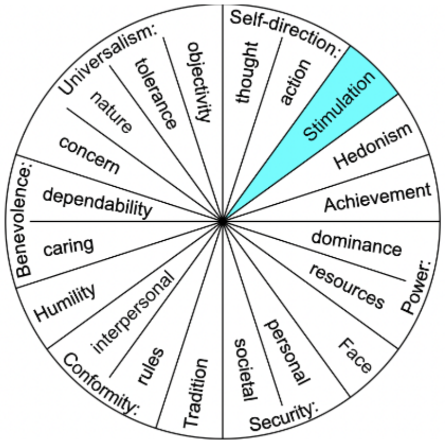


Figure 1: 20 Human Values

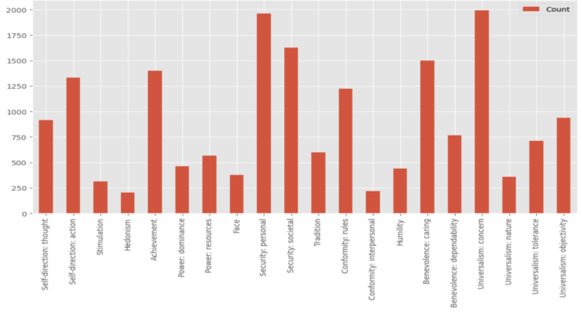


Figure 2: A bar plot of the most frequent terms in the labelled dataset

3 Experimental Setup

Due to the multi-label nature of the task and our data set comprising of both text and categorical features, feeding them into the models at once had given rise to various incompatibility issues that forced us to look for alternative approaches. The stages of our final decided approach will be explored in this section.

3.1 Experimental Procedure

Our experimental pipeline is as follows:

1. The three features are pre-processed according to their data type (text/categorical),
2. Four machine learning techniques were selected to be used for multi-label classification. Using sklearn, we were able to initialize the

multi-output versions of each learning algorithm we chose.

3. For each machine learning technique, three model copies were created, each trained on one of the three features. The predictions were made through the maximum voting process, similar to an ensemble classifier.

4. Each model was then trained on 70 per cent of the Labelled ValueEval dataset provided to us, before being made to predict the remaining 30 per cent. Due to imbalance in class frequencies observed within the dataset, we chose three specific metrics for evaluation:

   (a) Macro Average Precision: A type of precision that considers all classes with equal weight,

   (b) Weighted Average Precision: A precision calculation that adjusts the weight of each class according to the amount of samples belonging to it.

   (c) F1-Score: A harmonic mean of recall and precision, and the primary measure of performance in the competition.

5. Model hyper parameters are fine-tuned and observed for changes in performance,

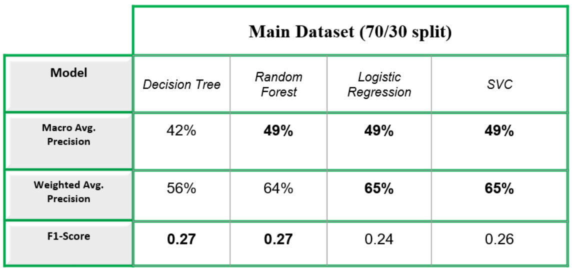
6. The overall best performing model is chosen for the prediction of the main test set provided by ValueEval.

3.2 Data Pre-processing

The ‘Conclusion’ and ‘Premise’ features, being in textual form, needed to be pre-processed before being fed into our machine learning models. To break them down into simpler, digestible components, we performed:

- Stopword removal to prevent potential bias,
- Punctuation removal to reduce noisy data,
- Tokenization,
- TF-IDF vectorization, a technique that takes word frequencies into account, giving more importance to uncommon words while placing a lower weight on very frequent terms. This type of vectorization converts text into numeric data to be understood by the model.

The ‘Stance’ feature was encoded into a binary value, as it represented only two types.



	Main Dataset (70/30 split)			
Model	*Decision Tree*	*Random Forest*	*Logistic Regression*	*SVC*
Macro Avg. Precision	42%	**49%**	**49%**	**49%**
Weighted Avg. Precision	56%	64%	**65%**	**65%**
F1-Score	**0.27**	**0.27**	0.24	0.26

Figure 3: The Metrics obtained for the four machine learning algorithms when trained/tested on the labelled ValueEval Dataset

3.3 Models used

To aid in the extraction of implicit themes, we looked to traditional machine learning techniques as a foundation to which we base our experiments off. Taking the difficulty of multi-label classification into consideration, we decided to implement ’triplet ensembles’ of four learning algorithms.

3.3.1 Logistic Regression

A probabilistic classifier chosen for its versatility and simplicity. Final adjusted hyper-parameters: Penalty = ’l2’, C = 1.0, solver = ’lbfgs’

3.3.2 Decision Tree

A split-based algorithm that continuously divides samples based on feature importance. Final adjusted hyper-parameters: Criterion = ’gini’, splitter = ’best’, minimum sample split = 2

3.3.3 Random Forest

A combination of decision trees that have their predictions go through a maximum voting process to produce the final output. Final adjusted hyper-parameters: Number of estimators = 100, criterion = ’gini’, minimum sample split = 2, minimum leaf samples = 1

3.3.4 Support Vector Classifier (SVC)

A classification model that focuses on finding the margin that best segregates data points based on their classes. Final adjusted hyper-parameters: C = 1.0, rbf kernel, tolerance (tol) = 0.001

We have published our submission runs on Github in the following repository https://github.com/sanidhyaRsharma/2022-SemEval-Human-Value-Detection

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
2023-01-30-08-51-04	.01	.00	.00	.00	.00	.01	.00	.00	.00	.08	.03	.00	.00	.00	.00	.00	.00	.01	.04	.00	.00

Table 1: Achieved F_1-score of team friedrich-nietzsche 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.

4 Results and Analysis

As shown in Figure 3, the supervised learning algorithms achieved considerably significant average F1-Scores when trained and tested using a 70-30 split of the labelled ValueEval Dataset. Selecting the Support Vector Classifier as our competing model due to it obtaining slightly higher precision rates for the classes, we expected it to attain similar scores when predicting ValueEval's evaluation test set. However, Table 1 reflects the underwhelming results achieved on the unlabelled test dataset that placed us last in the competition, with F1-scores ranging from 0.0 to only 0.04 for each class. Despite the scores being insignificant when compared to others' approaches, they are signs of insight as the possible limitations with our tools are revealed.

One of the biggest problems with our method was the TF-IDF vectorization. Due to it only considering word frequency and converting terms into sparse vectors of simple ones and zeroes, the technique fails at capturing context depending on the phrase or sentence provided. Moreover, unlike unsupervised algorithms like GloVE and Word2Vec, TF-IDF cannot extract patterns on its own and solely relies on a static factor like word count to generate word vectors.

The simpler complexity of traditional machine learning models when put next to their deep learning counterparts cannot be ignored either. While BERT-based architectures consist of multiple layers with each serving a purpose in the semantic recognition of text data, the SVC we utilized is a more 'straightforward' method of predicting classes through margin distances.

5 Conclusion and Future Work

Finding human values in argumentative texts can be useful for a number of applications, including value-based argument generating, value-based personality profiles, and argument-faceted searches. A study of human values has the potential to expand prospective research in each of these applications. With this study being our first foray into Natural Language Processing, we observed that multi-label human value prediction can be convoluted due to steep pre-requisites of data preprocessing and need of more complex machine learning models for better extraction of semantic relations between arguments and conclusions. Despite the lackluster performance of our approach, we still believe in the potential of traditional machine learning algorithms to be a low-cost alternative to neural networks for classification of thematic literary values. As no field of research is ever truly conquered, we plan to further our research to include BERT-based models, unsupervised vectorization methods, and more literary data sets for not only enlightening machines, but ourselves to witness the beauty of our writings.

6 Acknowledgments

The authors would like to begin by expressing their utmost gratitude to Dr. Jonathan Rusert for his guidance and support throughout the research process. Secondly, we sincerely thank the organizers of SemEval 2023 for granting us a special opportunity to improve our research skills and expand our knowledge in the field of Semantic Analysis.

References

Trevor JM Bench-Capon. 2003. Persuasion in practical argument using value-based argumentation frameworks. *Journal of Logic and Computation*, 13(3):429–448.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Zheng Li, Dawn Xiaodong Song, and Jacob Steinhardt. 2020. Aligning ai with shared human values. *ArXiv*, abs/2008.02275.

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)*.

Tushar Maheshwari, Aishwarya Reganti, Upendra Kumar, Tanmoy Chakraborty, and Amitava Das. 2017. Semantic interpretation of social network communities. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Liang Qiu, Yizhou Zhao, Jinchao Li, Pan Lu, Baolin Peng, Jianfeng Gao, and Song-Chun Zhu. 2022. Valuenet. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):11183–11191.

Milton Rokeach. 1973. *The nature of human values*. Free press.