

Representing Human Values in Argumentation: A Linguistically Grounded Comparison of Lexicons, Embeddings, and LLMs

Chaya Liebeskind

Jerusalem College of Technology

liebchaya@gmail.com

Barbara Lewandowska-Tomaszczyk

University of Applied Sciences in Konin

barbara.lewandowska-tomaszczyk@konin.edu.pl

Shmuel Liebeskind

Bar-Ilan University

israellieb@gmail.com

Abstract

Human values are enduring motivational principles that shape how people argue, justify, and evaluate social issues. While prior work has largely approached value detection as a standard classification problem, we investigate what values look like linguistically and whether computational models reflect aspects of the structure proposed by Schwartz’s theory of basic human values. Using the training partition of the Touché ValueEval corpus, we compare lexical resources, linguistic markers, supervised classifiers over TF-IDF and sentence embeddings, an automated model-search procedure, a zero-shot large language model (Gemini 3.5 Flash), and a transformer embedding baseline (all-MiniLM-L6-v2). On a shared 124-item test split, the three dense semantic approaches form a statistically indistinguishable high-performing tier (macro-F1 0.529–0.591) and achieve substantially higher scores than the sparse lexical methods, highlighting the importance of contextual semantic representation in this setting. A constructed diagnostic sanity check shows that deontic modals, universal quantifiers, and abstract nouns behave in the expected direction for value-bearing versus opinion language; however, these cues provide little information for distinguishing individual values in the Touché data. Performance is less stable for low-support value categories. Qualitative inspection of model confusions reveals several patterns consistent with theoretically close Schwartz values, alongside confusions between more distant categories; circumplex adjacency is not formally tested. Overall, the results suggest a distinction between surface linguistic cues associated with value-bearing language and the contextual semantic information involved in distinguishing specific values, while offering qualitative evidence that some model confusions align with relationships in Schwartz’s motivational structure.

1 Introduction

Values are the deeply ingrained principles by which people judge the world, distinct from *opinions*, which are time- and target-bound evaluations of specific entities. Their influence extends beyond abstract beliefs to consistent patterns of judgement and behaviour across contexts (Bardi and Schwartz, 2003). Identifying values in text has broad relevance across NLP: in political discourse analysis, values underlie policy preferences and frame ideological divides; in computational persuasion research, value appeals are a core rhetorical device; and in AI alignment and safety, where understanding and evaluating value representations has become an increasingly important research direction (Gabriel, 2020; Ren et al., 2024; Gupta et al., 2025). Characterising how human values are expressed in language is therefore an important step toward developing systems that can recognise and reason about value-laden content. Value detection also supports the study of cross-cultural communication, public opinion formation, and computational analyses of human values in discourse.

This paper investigates how human values are expressed in argumentative language and whether computational models reflect the motivational distinctions proposed by Schwartz’s theory of basic human values (Schwartz, 1992, 2012). While previous work has primarily focused on improving predictive performance, we approach value detection as a linguistic and semantic analysis problem rather than solely as a classification task. Accordingly, we examine the linguistic signals associated with value expression, conduct a comparison of lexical, supervised, embedding-based, zero-shot LLM, and unsupervised topic-modeling approaches within a unified experimental framework, and analyse whether model errors reflect the motivational relationships predicted by Schwartz’s theory.

We address three research questions:

- **RQ1.** Which linguistic signals distinguish value-bearing text from opinion?
- **RQ2.** Which representation strategies best capture Schwartz values for automatic classification?
- **RQ3.** Do model confusion patterns reflect the motivational structure predicted by Schwartz’s circumplex?

To answer these questions, we make three contributions. First, we conduct a systematic comparison of lexicon-based, classical supervised, embedding-based, zero-shot LLM, and unsupervised topic-modeling approaches within a unified evaluation framework. Second, rather than treating these models solely as classifiers, we use them as analytical tools to investigate the linguistic and semantic structure of value expression. Third, we identify qualitative confusion patterns that are consistent with aspects of Schwartz’s motivational structure, suggesting a possible relationship between model errors and theoretical proximity among values.

2 Related Work

Schwartz’s basic values. Schwartz’s theory of basic human values (Schwartz, 1992, 2012) identifies ten motivationally distinct values—power, achievement, hedonism, stimulation, self-direction, universalism, benevolence, tradition, conformity, and security—arranged in a circumplex defined by two opposing dimensions: openness-to-change versus conservation and self-enhancement versus self-transcendence. Adjacent values share motivational goals, whereas opposing values conflict. This structure implies that the categories are neither independent nor arbitrary, providing a theoretically grounded expectation for patterns of computational confusion.

Values vs. opinions. Values are enduring, abstract principles (e.g., “sustainability is important”), whereas opinions are target-specific evaluations (e.g., “this phone’s battery is terrible”). They are expressed through different linguistic markers: value-bearing language favors deontic modals, abstract nouns, and universal quantifiers, while opinion language more often employs evaluative adjectives, first-person hedges, and comparatives (Section 4). This distinction is particularly important in computational argumentation, where concrete evaluations are frequently intertwined with appeals to broader

principles. Distinguishing opinions from the values motivating them is therefore central to value-based reasoning. A classifier that conflates these dimensions risks attributing a general evaluative stance to a specific human value.

Other value frameworks. Beyond Schwartz’s model, Moral Foundations Theory (Graham et al., 2013) proposes a set of innate moral foundations (care/harm, fairness, loyalty, authority, and sanctity), whose associated Moral Foundations Dictionary serves as one of our lexical baselines. The World Values Survey provides a complementary cross-cultural perspective on human value priorities (Schwartz, 1992).

Value detection in NLP. Kiesel et al. (2022) introduced the Touché/ValueEval task and released the multi-label corpus, followed by transformer-based systems in the SemEval shared task (Kiesel et al., 2023). Earlier NLP research on stance detection (Mohammad et al., 2016) and subjectivity analysis provides important context, but value detection differs in that values are neither binary nor anchored to a specific target; instead, they are grounded in psychological theory rather than annotator consensus. Recent work has also shown that disagreement in value annotation reflects differences in both human and model reasoning, highlighting the inherently subjective nature of value identification (Falk and Lapesa, 2025). Overall, previous work has primarily emphasized predictive performance, while comparatively less attention has been paid to the linguistic and semantic structure of value expression.

Argument mining, AI alignment, and model interpretability. Argument mining (Stab and Gurevych, 2017) focuses on identifying claims, premises, and their relations, whereas value detection addresses the motivational principles underlying an argument. Explicit representations of human values have also become increasingly important for AI alignment and interpretability (Gabriel, 2020), where systems must reason about value-laden content rather than only factual or topical information. Understanding how computational models represent and confuse human values therefore contributes both to computational argumentation and to the broader study of value-aware language models.

LLMs as zero-shot annotators. Recent studies suggest that large language models can both rec-

ognize and express value-related distinctions in text (Gilardi et al., 2023; Ren et al., 2024; Han et al., 2025). In particular, LLMs have been shown to annotate subjective and value-laden labels with accuracy competitive with human crowd workers (Gilardi et al., 2023), making them an attractive alternative for low-resource annotation tasks.

3 Dataset and Value Mapping

We use the training partition of the Touché ValueEval’23 corpus (Kiesel et al., 2022; Mirzakhmedova et al., 2024), containing 5,393 arguments. After mapping the 20 fine-grained labels to the ten classic Schwartz values, dropping *Face* and *Humility*, and retaining arguments expressing exactly one value, 636 instances remain. These instances are repartitioned for our experiments as described in Section 5.

The resulting distribution is strongly imbalanced (universalism 230, security 184, self-direction 113, . . . , hedonism and stimulation only one each; Figure 4, right), leaving several value categories with very limited support.

The 20→10 mapping aggregates the level-2 categories (e.g., *Self-direction: thought* and *Self-direction: action* both map to *self-direction*); *Face* and *Humility* are dropped because they have no counterpart among the classic ten. We restrict the analysis to single-value arguments to separate the study of value representation from the additional ambiguity introduced by multi-label annotation. Although this substantially reduces the available data, it provides a simplified setting for studying value identification before extending the analysis to the original multi-label task.

4 Linguistic Markers of Values

A central aim of this work is to understand *what linguistic signal expresses human values*, not merely to optimise classification accuracy (RQ1). To probe this, we design a targeted experiment contrasting value-bearing and opinion sentences.

We encode six feature groups—deontic modals, universal quantifiers, abstract nouns, first-person hedges, evaluative adjectives and comparatives—and fit a balanced logistic regression. Table 1 summarises the features with their theoretical loading.

We evaluate the marker classifier at two scales. First, we apply it to a hand-built diagnostic set of 60 sentences (50 value statements; 10 opinions). This constructed diagnostic set is intentionally cue-

Feature group	Direction	Example cues
Deontic modals	value+	must, should, ought
Universal quant.	value+	always, everyone, all
Abstract nouns	value+	justice, freedom, dignity
1st-person hedges	opinion−	I think, in my view
Evaluative adj.	opinion−	terrible, amazing, awful
Comparatives	opinion−	than, better, worse

Table 1: Linguistic marker features and their value/opinion loading.

driven and is used as a sanity check of the proposed linguistic markers rather than as an independent validation of their ability to distinguish naturally occurring value statements from opinions. The binary value/opinion classifier separates the two classes in the diagnostic set perfectly, and the learned coefficients recover the predicted direction: deontic modals, abstract nouns, and universal quantifiers push toward *value*; evaluative adjectives, hedges, and comparatives push toward *opinion*.

Second—and more importantly—we apply the same six features to the full 10-way value classification task on the Touché corpus, training on the 512-item pool and evaluating on the shared 124-item held-out test set (Marker LR in Table 3). The marker classifier achieves macro-F1 0.159 (95% CI [0.08, 0.22]), essentially matching the hand lexicon (0.161) and remaining far below the dense semantic approaches.

Taken together, the diagnostic and Touché results suggest a tentative distinction: the selected surface cues are sufficient to separate the constructed value/opinion examples, but provide little information for distinguishing among specific Schwartz values in naturally occurring arguments. This pattern is consistent with a two-level distinction between surface cues associated with value-bearing language and the semantic content that differentiates values such as *Universalism* and *Security*. The results therefore suggest that syntactic and stylistic cues may help indicate the presence of value-based reasoning, whereas finer-grained distinctions between motivational types benefit substantially from richer contextual semantic representations.

5 Experimental Setup

After filtering, we construct a stratified argument-level split with seed 0, yielding 388 training, 124 validation, and 124 test instances (approximately 60/20/20). Hyperparameters are selected on the validation partition, and final performance is evaluated

Value	Top data-mined terms
universalism	animals, cruel, whaling, species, zoos
self-direction	freedom, choice, vote, individual
security	health, dangerous, food, obesity
conformity	policy, zero-tolerance, schools, respect
achievement	knowledge, skills, work, financial
tradition	culture, religion, integration, banning

Table 2: Representative terms learned per value by the data-mined lexicon.

once on the 124-item held-out test set.

For classification, we report accuracy and macro-F1. Macro-F1 is computed over the eight value labels represented in the test set; *hedonism* and *stimulation* have no test instances and are therefore excluded from the macro average. For clustering, we report Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI). Bootstrap 95% confidence intervals are computed from 10,000 resamples of the held-out test set using a fixed set of the eight supported labels. Pairwise comparisons among the three strongest semantic approaches use paired bootstrap resampling over the same test instances.

Experiments use `gemini-3.5-flash`, `gemini-embedding-001`, `all-MiniLM-L6-v2`, and standard scikit-learn classifiers. The training and validation partitions together contain 512 of the 636 single-label arguments, with the remaining 124 reserved for testing. Random seeds are fixed throughout, and implementation details, prompts, split-generation code, and bootstrap procedures are provided in Appendix B.

5.1 Lexical baselines

We compare two lexical approaches. The first is a hand-built lexicon combining Schwartz value keywords with the Moral Foundations Dictionary (Graham et al., 2013). The second is a data-driven lexicon learned from the training corpus using Rocchio weighting over TF-IDF. Representative learned terms are shown in Table 2; the complete construction procedure is described in Appendix C.

5.2 Supervised classification

We train linear classifiers over TF-IDF features using an automated hyperparameter search over feature representations and classifier families. After optimization, the best TF-IDF model achieves macro-F1 0.386 on the held-out test set. Replacing TF-IDF with `gemini-embedding-001` sen-

tence embeddings and a linear SVM improves performance to macro-F1 0.540 (+40% relative). Details of the search space and optimization procedure are provided in Appendix B.

5.3 Zero-shot LLM

We evaluate Gemini 3.5 Flash in a zero-shot setting by prompting it with the ten Schwartz value definitions and asking it to predict exactly one value for each argument. The complete prompt is given in Appendix A. Because the prompt provides explicit definitions of all ten Schwartz categories, this condition should be interpreted as task-informed zero-shot classification rather than as a definition-free representation comparison. The model is evaluated on the same held-out test set used for all supervised approaches.

5.4 Transformer baseline

As a non-proprietary embedding baseline, we encode each argument using `all-MiniLM-L6-v2` and classify the resulting sentence embeddings using the same model-selection procedure as for the Gemini embeddings.

5.5 Unsupervised topic modelling

We compare classical TF-IDF clustering (LDA and KMeans) with BERTopic (Grootendorst, 2022) over Gemini embeddings, evaluating cluster quality using Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI).

6 Results

Table 3 and Figure 1 summarise the comparison. All seven methods are evaluated on the same 124-item stratified held-out test set. The results reveal a clear separation between sparse lexical representations and dense semantic representations. All embedding-based and LLM approaches substantially outperform lexical methods, while differences within the semantic-representation group are comparatively small and not statistically significant.

Zero-shot Gemini 3.5 Flash achieves the highest macro-F1 (0.591), closely followed by Gemini embeddings (0.540) and MiniLM-L6-v2 (0.529). All LLM/embedding methods substantially exceed the sparse TF-IDF ceiling (0.386). Bootstrap 95% confidence intervals (Table 3) are wide—confirming that the small test size ($n=124$) limits statistical precision—and pairwise bootstrap tests between

Method	Representation	F1	95% CI
Marker LR	6 markers	0.159	[0.08, 0.22]
Hand lexicon	Schwartz+MFD	0.161	[0.07, 0.24]
Learned lex.	TF-IDF	0.363	[0.25, 0.45]
AutoResearch	TF-IDF	0.386	[0.27, 0.47]
Gemini emb.	emb.+SVM	0.540	[0.41, 0.63]
MiniLM	emb.+SVM	0.529	[0.40, 0.61]
Gemini CoT (0-shot)	LLM	0.591	[0.42, 0.69]

Table 3: Value identification on Touché ValueEval, macro-F1 over 8 test-supported labels ($n=124$, bootstrap $n=10,000$). Marker LR: 6 linguistic features (Section 4), balanced LR. MiniLM: all-MiniLM-L6-v2. Gemini emb.: gemini-embedding-001 (3072-d). Wide CIs reflect the small test set; pairwise differences within the LLM/embedding tier are not significant ($p>0.05$). Experiment shorthands exp5/exp6/exp8 are defined in Appendix D.

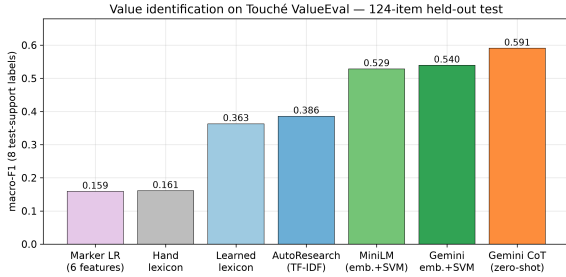


Figure 1: Macro-F1 of value-identification methods (held-out 124-item test set). Zero-shot Gemini CoT achieves the highest point estimate; Gemini embeddings and MiniLM-L6-v2 are close behind. Pairwise differences among the three LLM/embedding methods are not statistically significant ($p > 0.05$).

the three LLM/embedding methods are not significant at $\alpha = 0.05$ (all $p > 0.05$; $n = 10,000$ resamples), indicating that they form a single performance tier.

Per-class behaviour (RQ2). Per-class performance (Figure 2) is generally stronger for frequent categories, but substantial variability remains among low-support classes. Figure 3 illustrates this relationship: low-support categories tend to exhibit less stable performance, although the association between class frequency and F1 is not monotonic.

The gap between accuracy and macro-F1 is large: on the held-out test, the zero-shot LLM achieves accuracy 0.726 but macro-F1 0.591 (over the eight supported values), because frequent classes contribute more heavily to overall accuracy, whereas macro-F1 gives equal weight to each supported category. Table 4 gives the full per-class breakdown.

Value	P	R	F1	n
universalism	0.73	0.83	0.78	46
security	0.68	0.72	0.70	36
self-direction	0.81	0.77	0.79	22
achievement	0.75	0.50	0.60	6
conformity	0.67	0.40	0.50	5
tradition	0.75	1.00	0.86	3
power	1.00	0.33	0.50	3
benevolence	0.00	0.00	0.00	3
hedonism	—	—	—	0
stimulation	—	—	—	0
macro avg (8)	0.67	0.57	0.59	124

Table 4: Per-class precision, recall, F1 and support for zero-shot Gemini 3.5 Flash on the held-out test set ($n = 124$). The two singleton values have no test examples; macro averages are computed over the 8 supported values. Performance varies substantially across low-support classes.

Performance is strongest for several relatively frequent categories, including universalism, security, and self-direction, while some low-support categories remain difficult. At the same time, performance varies substantially even among similarly rare classes, indicating that class support is an important but not sufficient explanation of per-class performance.

Qualitative examples. Table 5 shows representative held-out examples illustrating correct and incorrect predictions. Several errors involve theoretically close values, while others span more distant regions of the value space, consistent with the qualitative analysis in Section 7.

Unsupervised results. BERTopic obtained a higher NMI than TF-IDF clustering in this experiment (0.209 vs. 0.150), although neither method cleanly recovered the value categories (Figure 4). A 2-D embedding map (Figure 5) shows partial overlap and separation among value categories, consistent with the difficulty of recovering them as purely topical clusters.

7 Error Analysis and Value Structure

The row-normalised confusion matrix of the Gemini-embedding classifier (Figure 6, left) is partly, but not uniformly, consistent with Schwartz’s motivational structure. Several confusions occur between theoretically close values: power is misclassified as achievement (both within self-enhancement), benevolence as universalism (both within self-transcendence), and conformity

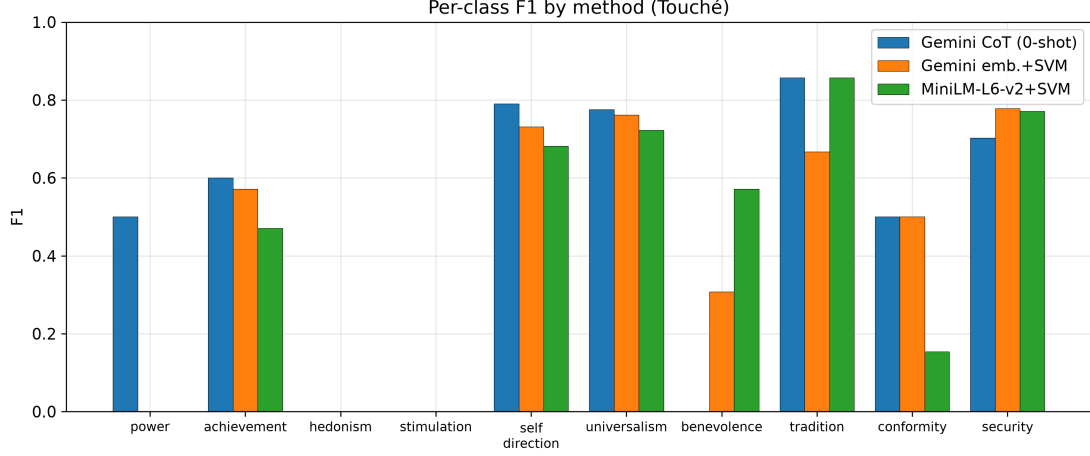


Figure 2: Per-class F1 for the three strongest methods.

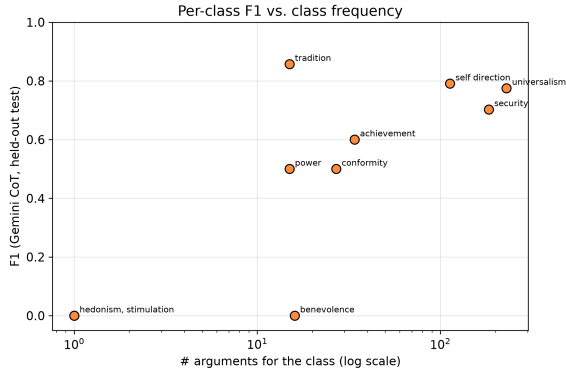


Figure 3: Per-class F1 (Gemini CoT) versus class frequency in the filtered corpus (log scale). Low-support categories exhibit substantially greater performance variability.

as security (both within conservation). Universalism and self-direction, which occupy neighbouring regions in the circumplex, are also confused in both directions.

At the same time, the largest off-diagonal masses connect universalism and security, two motivationally more distant values that are also the two largest categories in the held-out test set. These confusions may therefore partly reflect class-frequency effects rather than theoretical proximity. Direct confusions between opposing regions (e.g., achievement→benevolence and self-direction→security) occur only as isolated instances. We therefore interpret the confusion structure as qualitatively consistent with the value circumplex in specific cases, without claiming that the classifier recovers it systematically. Establishing a systematic relationship between theoretical

distance and model confusion would require a formal distance-based analysis controlling for class frequency, which we do not attempt here.

The close-value confusions are also semantically interpretable. Power and achievement share associations with success, influence, and status, whereas benevolence and universalism both concern the welfare of others but differ primarily in scope—from close others to broader concern for humanity and nature. Such cases illustrate why boundaries between neighbouring motivational concepts may be difficult for a classifier to distinguish, without implying that model errors systematically follow circumplex proximity.

Aggregating per-class performance by Schwartz’s higher-order value groups (Figure 6, right), conservation shows the highest descriptive mean F1 (0.69), whereas openness-to-change, self-enhancement, and self-transcendence obtain lower and broadly similar values (0.37–0.40). These group-level scores should be interpreted cautiously, however, because they are simple averages of per-class F1 and include categories with no test support, notably *hedonism* and *stimulation*. We therefore treat the higher-order aggregation as a descriptive summary rather than as evidence that some motivational regions are intrinsically easier to recover than others.

More generally, the analysis illustrates how model errors can complement aggregate performance metrics by highlighting where semantic category boundaries become difficult to distinguish. In this sense, confusion patterns provide a useful qualitative lens on computational representations of human values, while stronger theoretical con-

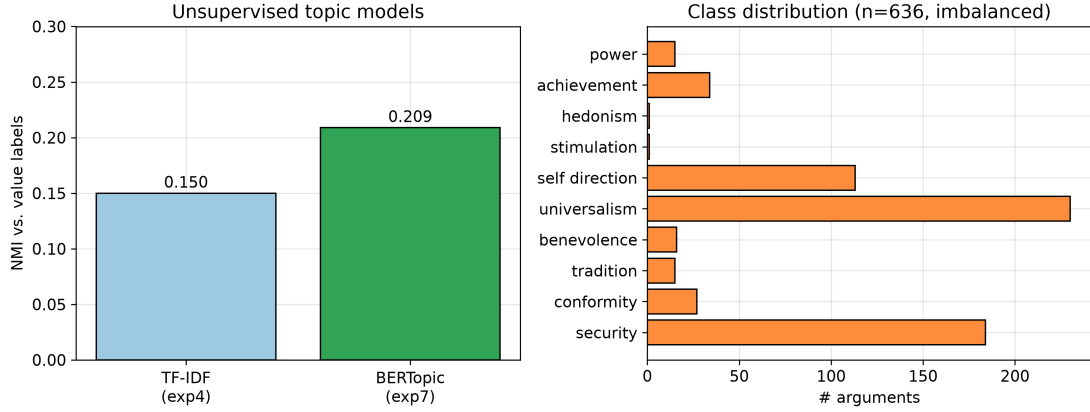


Figure 4: Left: cluster quality (NMI vs. value); BERTopic obtained a higher NMI than TF-IDF clustering in this experiment. Right: the imbalanced class distribution of the corpus.

Argument premise (abbreviated)	Gold	Pred.	✓
<i>Correct predictions</i>			
Capital punishment is cruel and unusual.	univ.	univ.	✓
Banning cosmetic surgery is like banning freedom of speech.	s-dir.	s-dir.	✓
They will lead to more deaths on the roads.	sec.	sec.	✓
<i>Errors</i>			
Legalizing prostitution will further demean a woman or man and result in unnecessary violence.	conf.	sec.	×
As EU citizens we should never forget how privileged we are... we bear a human obligation to help others in need.	ben.	univ.	×
IP rights enable large corporations to have a stranglehold over creators' designs and work and should be abolished.	s-dir.	univ.	×
We should legalise organ trade because it could save so many lives.	sec.	univ.	×

Table 5: Illustrative examples drawn from the held-out test predictions of zero-shot Gemini. Examples are selected for qualitative illustration and do not constitute a formal test of circumplex adjacency.

clusions require formal testing of the relationship between model errors and motivational distance.

8 Discussion

Three findings emerge from our experiments. First, dense semantic approaches perform strongly in this setting (RQ2). As summarised in Table 6, the embedding- and LLM-based approaches achieve substantially higher macro-F1 than the lexical and

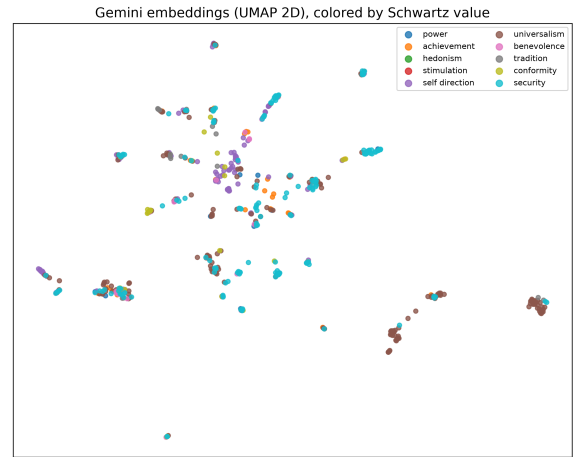


Figure 5: UMAP projection of the Gemini embeddings, coloured by gold value.

sparse TF-IDF methods, whereas additional optimization within the sparse TF-IDF family yields only modest gains. The three strongest semantic methods form a statistically indistinguishable performance tier. These results suggest that representation choice is an important factor in performance. However, the compared systems also differ in model family, training regime, and available task information, so the present comparison does not isolate representation as a single causal factor.

This pattern is consistent with previous work showing that distributional semantic representations can capture culturally meaningful concepts beyond lexical similarity (Kozłowski et al., 2019; Rodman, 2020).

Second, zero-shot LLMs provide a practical alternative for value annotation in low-resource settings. Although Gemini CoT achieves the high-

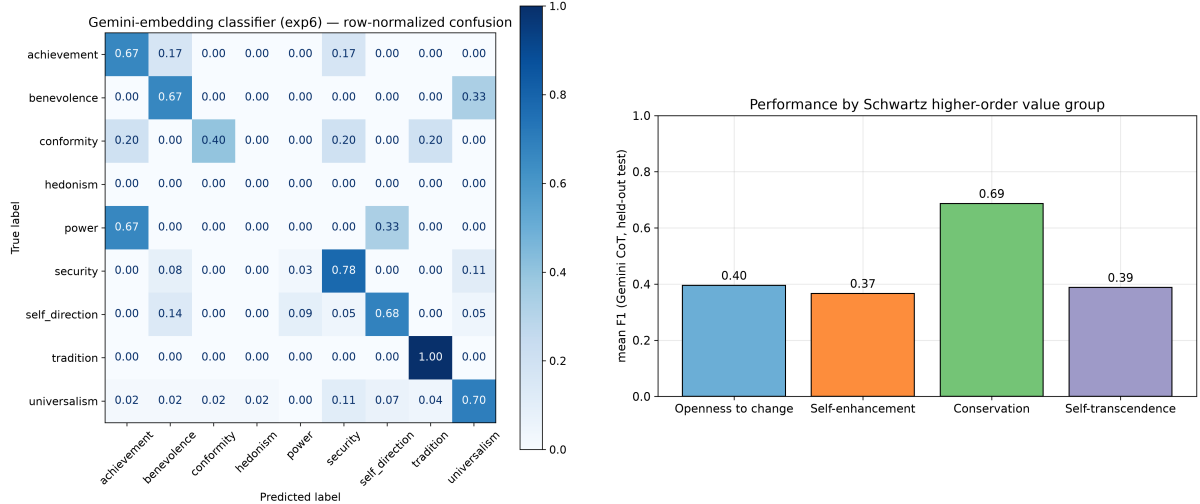


Figure 6: Left: row-normalised confusion matrix of the Gemini-embedding classifier. Right: descriptive mean per-class F1 by Schwartz higher-order value group for zero-shot Gemini CoT. Group-level means include categories with no test support and are intended as descriptive summaries.

Representation type	Methods	F1 range
Marker features	Marker LR	0.159
Dictionary lookup	Hand lexicon	0.161
Sparse TF-IDF	Learned lex., AutoResearch	0.363–0.386
Dense emb.+SVM	Gemini emb., MiniLM	0.529–0.540
Zero-shot LLM	Gemini CoT	0.591

Table 6: Observed performance by representation type. The three strongest semantic methods—Gemini embeddings, MiniLM, and zero-shot Gemini CoT—do not differ significantly in paired bootstrap comparisons.

est point estimate, its confidence interval overlaps substantially with those of the embedding-based models, and paired bootstrap comparisons show no significant differences among the three strongest semantic approaches. The zero-shot condition requires no task-specific parameter training, although the prompt provides explicit definitions of the Schwartz value categories. The results therefore show that an LLM supplied with task definitions can identify value categories at a level comparable to supervised classifiers over dense embeddings, complementing recent work on value-aware LLMs (Han et al., 2025).

Third, the similar performance of Gemini embeddings and the substantially smaller public MiniLM model indicates that the advantage of dense semantic representations in this experiment is not specific to proprietary models. At the same time, data sparsity remains an important limitation: performance

is highly variable among low-support values, and several categories have too few examples for stable evaluation. The error analysis adds a complementary perspective. Some model confusions occur between theoretically close Schwartz values, while others span more distant regions of the value space. We therefore interpret the observed alignment with the circumplex as qualitative and case-specific rather than as evidence that the models systematically recover its motivational geometry.

More broadly, the results suggest that value identification depends on contextual semantic distinctions that are only weakly captured by surface lexical cues. In this sense, human values occupy a level of abstraction that is not well represented as simple topics or keyword sets. This perspective helps explain the strong performance of dense semantic approaches in our experiments and supports treating value identification primarily as a semantic representation problem rather than as conventional lexical text classification.

9 Conclusion

Human values are central to argumentation because they provide the motivational principles that arguments invoke, defend, and contest. This study examined how such values are expressed in language and how different computational approaches represent them. Across our experiments, dense semantic approaches achieved higher macro-F1 than lexical baselines, suggesting that contextual semantic information plays an important role in value identi-

fication.

Beyond predictive performance, the results suggest a distinction between surface cues associated with value-bearing language and the richer contextual semantics involved in distinguishing specific values. The constructed diagnostic analysis shows that selected syntactic and stylistic cues behave in the expected direction for value-bearing versus opinion language, but provide little information for separating individual Schwartz values. In addition, several model confusions are qualitatively consistent with relationships in Schwartz’s motivational structure, while others span more distant regions of the value space. The present analysis should therefore be viewed as suggestive rather than as a formal recovery of the circumplex.

Future work should extend the analysis to larger and more balanced multilingual corpora, evaluate the original multi-label ValueEval setting, and test whether the observed linguistic and confusion patterns generalize across domains, languages, and cultures.

Limitations

The 20→10 mapping drops *Face* and *Humility*, and restricting the data to single-label arguments discards most of the original examples and produces a strongly imbalanced distribution. Inputs are premise-only, English, and drawn from a single shared-task corpus.

The value/opinion analysis is based on a small, cue-driven constructed diagnostic set and should be interpreted as a sanity check rather than as independent validation on naturally occurring text. The supervised experiments use a custom argument-level repartitioning of the official ValueEval training split rather than preserving conclusion-level groups. As a result, different arguments associated with the same conclusion can appear in different partitions; in our split, this occurs for 45 conclusions shared between training and test. Because the models receive only the argument premise, this does not constitute duplicate or direct input leakage, but it may introduce topical overlap between training and evaluation data.

Bootstrap 95% confidence intervals (Table 3) confirm that the 124-item test set provides limited statistical precision. Pairwise bootstrap comparisons among the three strongest semantic approaches are not significant. The MiniLM result uses a frozen encoder with a linear classifier; task-

specific fine-tuning could yield different performance. BERTopic and the embedding pipeline contain stochastic components; although random seeds are fixed, absolute results may vary slightly across environments.

Finally, the observed correspondence between some model confusions and Schwartz’s circumplex is qualitative and should be interpreted as suggestive rather than confirmatory. We do not formally operationalize motivational distance or test whether adjacent values are confused more often than expected after controlling for class frequency. Establishing such relationships would require larger, more balanced datasets and a dedicated distance-based analysis.

Ethics Statement

We use the publicly released Touché ValueEval corpus (Kiesel et al., 2022); no new human-subject data were collected and the texts contain no personal identifiers. Automatic value detection is dual-use: it can support the analysis of public discourse, but it could also enable profiling or targeted persuasion. We report aggregate metrics only and discourage deployment for covert profiling of individuals.

AI Usage Disclosure

Large language models are an object of study in this work (gemini-3.5-flash for classification, gemini-embedding-001 for embeddings). An AI assistant was also used to help draft and edit the manuscript and the experiment code; all reported results were produced by the released scripts and verified by the authors, who take full responsibility for the content.

Data and Code Availability

The Touché ValueEval corpus is available from Zenodo (record 7879430). Code, random seeds, the 20→10 value mapping, prompts, cached embeddings, bootstrap evaluation, and scripts supporting the reported results and figures are available at <https://github.com/jct-nlp/LUHME-LinguisticallyLLM>.

Acknowledgments

We thank the organisers of the Touché ValueEval shared task for the data.

References

- Anat Bardi and Shalom H Schwartz. 2003. Values and behavior: Strength and structure of relations. *Personality and social psychology bulletin*, 29(10):1207–1220.
- Neele Falk and Gabriella Lapesa. 2025. Mining the uncertainty patterns of humans and models in the annotation of moral foundations and human values. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22898–22921.
- Iason Gabriel. 2020. [Artificial intelligence, values, and alignment](#). *Minds and Machines*, 30(3):411–437.
- Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. [ChatGPT outperforms crowd workers for text-annotation tasks](#). *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.
- Jesse Graham, Jonathan Haidt, Sena Koleva, Matt Motyl, Ravi Iyer, Sean P. Wojcik, and Peter H. Ditto. 2013. [Moral foundations theory: The pragmatic validity of moral pluralism](#). *Advances in Experimental Social Psychology*, 47:55–130.
- Maarten Grootendorst. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*.
- Aman Gupta, Denny O’Shea, and Fazl Barez. 2025. Val-bench: Measuring value alignment in language models. *arXiv preprint arXiv:2510.05465*.
- Jongwook Han, Dongmin Choi, Woojung Song, Eun-Ju Lee, and Yohan Jo. 2025. Value portrait: Assessing language models’ values through psychometrically and ecologically valid items. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 17119–17159.
- Johannes Kiesel, Milad Alshomary, Nicolas Handke, Xiaoni Cai, Henning Wachsmuth, and Benno Stein. 2022. [Identifying the human values behind arguments](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4459–4471.
- 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 (SemEval-2023)*, pages 2287–2303.
- Austin C Kozlowski, Matt Taddy, and James A Evans. 2019. The geometry of culture: Analyzing the meanings of class through word embeddings. *American Sociological Review*, 84(5):905–949.
- Nailia Mirzakhmedova, Johannes Kiesel, Milad Alshomary, Maximilian Heinrich, Nicolas Handke, Xiaoni Cai, Valentin Barriere, Doratossadat Dastgheib, Omid Ghahroodi, Mohammad Ali Sadraei Javaheri, Ehsaneddin Asgari, Lea Kawaletz, Henning Wachsmuth, and Benno Stein. 2024. [The touché23-ValueEval dataset for identifying human values behind arguments](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16121–16134, Torino, Italia. ELRA and ICCL.
- Saif M. Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. [SemEval-2016 task 6: Detecting stance in tweets](#). In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41.
- Yuanyi Ren, Haoran Ye, Hanjun Fang, Xin Zhang, and Guojie Song. 2024. Valuebench: Towards comprehensively evaluating value orientations and understanding of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2015–2040.
- Emma Rodman. 2020. A timely intervention: Tracking the changing meanings of political concepts with word vectors. *Political Analysis*, 28(1):87–111.
- Shalom H. Schwartz. 1992. [Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries](#). In *Advances in Experimental Social Psychology*, volume 25, pages 1–65. Elsevier.
- Shalom H Schwartz. 2012. An overview of the schwartz theory of basic values. *Online readings in Psychology and Culture*, 2(1).
- Christian Stab and Iryna Gurevych. 2017. [Parsing argumentation structures in persuasive essays](#). *Computational Linguistics*, 43(3):619–659.

A Zero-shot Prompt

The Gemini 3.5 Flash classifier receives the following prompt, with `{text}` replaced by the argument premise:

Classify the sentence into exactly one of the 10 Schwartz basic human values:

- **power:** social status, prestige, control or dominance over people and resources
- **achievement:** personal success through demonstrating competence by social standards
- **hedonism:** pleasure and sensuous gratification for oneself
- **stimulation:** excitement, novelty, and challenge in life
- **self-direction:** independent thought and action; choosing, creating, exploring
- **universalism:** understanding, tolerance, and protection for all people and nature
- **benevolence:** preserving and enhancing the welfare of close others

- **tradition**: respect and commitment to customs of one’s culture or religion
- **conformity**: restraint of actions that upset others or violate social norms
- **security**: safety, harmony, and stability of society, relationships, and self

Sentence: {text}. Reason step by step, then end with a line exactly: value: <one_value>

B Implementation Details

B.1 Experimental configuration

All experiments were implemented using scikit-learn, LangChain, and the Gemini API. Random seeds were fixed throughout all experiments, and Gemini embeddings were cached to support reproducibility.

After filtering, the 636 single-label arguments are repartitioned at the argument level using a stratified split with seed 0, yielding 388 training, 124 validation, and 124 held-out test instances. Singleton classes are assigned to the training partition.

Bootstrap confidence intervals are computed using 10,000 resamples of the 124-item held-out test set with seed 0, sampling test instances with replacement. Macro-F1 is computed in every resample over the same fixed set of eight labels represented in the test set; *hedonism* and *stimulation* have no test instances and are therefore excluded from the macro average. Labels absent from an individual bootstrap resample remain part of this fixed evaluation set and contribute an F1 of zero.

For pairwise comparisons among zero-shot Gemini, Gemini embeddings, and MiniLM, the same resampled test indices are used for both models in each bootstrap replicate, yielding paired distributions of macro-F1 differences. The resulting two-sided p -values are 0.634 for zero-shot Gemini versus Gemini embeddings, 0.515 for zero-shot Gemini versus MiniLM, and 0.771 for Gemini embeddings versus MiniLM; none of the three differences is significant at $\alpha = 0.05$.

B.2 Hyperparameter search

The TF-IDF classifier is optimized using an automated hill-climbing search. The search explores

- word and character n -gram features,
- document-frequency thresholds,
- sublinear TF weighting,
- classifier family (Linear SVM, Logistic Regression, Complement Naïve Bayes, SGD, Nearest Centroid),

- regularization,
- class weighting.

Candidates are evaluated using the mean macro-F1 over five resamples of the training pool. The search runs for 150 iterations with periodic random restarts.

For the embedding models, a balanced Linear SVM or Logistic Regression is selected on the validation split and retrained on the combined training+validation data before evaluation on the held-out test set.

C Data-mined Lexicon Construction

The data-driven lexicon is constructed using Rocchio weighting over TF-IDF. For each Schwartz value, we identify the terms whose TF-IDF mass is most concentrated within that value relative to the corpus-wide mean. The top 50 terms are retained and L2-normalized to form the class lexicon.

Top learned terms per value (Rocchio over TF-IDF). Frequent values yield coherent terms; the singleton classes (*hedonism*, *stimulation*) are noise.

power: european, space, useful, trading, economy, council
achievement: knowledge, gain, skills, work, financial, products
hedonism: (noise: trying, spread, world, ago)
stimulation: (noise: world, talent, olympic, games)
self-direction: freedom, choice, vote, individual, voting, bodies
universalism: animals, zoos, whaling, cruel, inhumane, species
benevolence: safe spaces, help, minors, cosmetic surgery, nepotism
tradition: culture, religion, integration, banning, whaling
conformity: policy, zero tolerance, schools, pride parades, respect
security: health, dangerous, food, obesity, accidents, surgery

D Per-class Supervised Results

Table 7 reports per-class F1 for the three supervised models on the shared 124-item held-out test set (singleton values have no test examples). Gemini embeddings (exp6) improve small classes over TF-IDF (e.g. tradition 0.00 $\rightarrow$ 0.67); MiniLM (exp8) further improves benevolence (0.31 $\rightarrow$ 0.57) and tradition (0.67 $\rightarrow$ 0.86) while lagging on conformity (0.50 $\rightarrow$ 0.15).

Value	n	exp5 F1	exp6 F1	exp8 F1
universalism	46	0.73	0.76	0.72
security	36	0.70	0.78	0.77
self-direction	22	0.62	0.73	0.68
achievement	6	0.18	0.57	0.47
conformity	5	0.18	0.50	0.15
benevolence	3	0.67	0.31	0.57
power	3	0.00	0.00	0.00
tradition	3	0.00	0.67	0.86
macro (8)		0.39	0.54	0.53

Table 7: Per-class F1 on the shared held-out test set: AutoResearch TF-IDF (exp5) vs. Gemini embeddings (exp6) vs. MiniLM-L6-v2 (exp8).

E Linguistic-Marker Features

The value/opinion classifier uses six substring-matched feature groups. Each feature is the count of cue terms found in the sentence.

Deontic modals (value+): must, should, ought, have to, shall, need to

Universal quantifiers (value+): always, never, everyone, all, every, fundamental

Abstract nouns (value+): justice, freedom, equality, tolerance, security, heritage, dignity, welfare, autonomy, peace

First-person hedges (opinion-): I think, I feel, in my view, honestly, seems, to me

Evaluative adjectives (opinion-): terrible, amazing, awful, boring, cheap, slow, expensive, disappointing

Comparatives (opinion-): than, compared, better, worse, faster, slower

A logistic regression over these six counts separates the diagnostic set perfectly; the sign next to each group is the direction of the learned coefficient, matching the theoretical markers in Section 4.

F Diagnostic Sentence Set

The 60 hand-built sentences used to evaluate the value/opinion marker classifier (Section 4). The set contains 50 value statements (5 per Schwartz basic value) and 10 opinion sentences. A logistic regression over the six marker features (Appendix E) separates the two classes perfectly.

Value statements. Power (self-enhancement): Those in authority must maintain control over resources to preserve social order. · Status and prestige should always be earned through dominance over rivals. · A true leader ought to command respect and control those who serve under them. · Power over others is what ultimately determines one's rightful place in society. · We must hold authority firmly because control is the foundation of all influence.

Achievement (self-enhancement): Personal success must be pursued relentlessly through demonstrated competence. · One should always strive to achieve ambitious goals and prove their ability. · Hard work and accomplishment are what truly define a worthy and capable life. · We ought to set challenging targets and excel far beyond every expectation. · Excellence and competence should always be rewarded above mediocrity.

Hedonism (self-enhancement): Life should be enjoyed and pleasure is the highest good one can ever seek. · We must indulge in the gratification

of our senses whenever the chance arises. · Enjoyment and personal pleasure ought to guide how we spend all our days. · One should always savor delight and seek sensual comfort in every moment. · Pleasure and fun are what make human existence truly worthwhile.

Stimulation (openness): Life must be full of excitement, novelty, and daring adventure at every turn. · We should always seek thrilling challenges that push us beyond our limits. · A bold, adventurous life of constant novelty is what everyone truly needs. · One ought to embrace risk and chase exhilarating new experiences fearlessly. · Variety, thrill, and surprise are essential to any life worth living.

Self-direction (openness): Everyone must be free to think independently and choose their own path. · Creativity and autonomy should always be protected as fundamental freedoms. · We ought to explore our own ideas without anyone else controlling us. · Independence of thought is the very foundation of a meaningful life. · People must always be free to create and decide things for themselves.

Universalism (self-transcendence): We must protect the environment and preserve nature for all future generations. · Justice, equality, and tolerance should be guaranteed for everyone everywhere. · Peace and social justice are foundational rights that must never be denied. · All people deserve equal dignity and we ought to defend the welfare of the planet. · Tolerance and understanding among all peoples must always be upheld.

Benevolence (self-transcendence): We must always care for our family and help those who are closest to us. · Kindness and loyalty toward our friends should guide every single action. · One ought to nurture and support the people in their own community. · Helping our loved ones and protecting their welfare is a sacred duty. · We should always be generous and compassionate toward those we love.

Tradition (conservation): We must preserve our heritage and honor the customs of our ancestors. · Tradition and religious custom should always be respected and faithfully upheld. · One ought to maintain the time-honored ways that define our shared culture. · Respect for our heritage must never be abandoned for the sake of novelty. · The customs handed down to us should always be observed without question.

Conformity (conservation): Everyone must obey the rules and conform to society's clear expect-

tations. · We should always restrain our impulses and follow established social norms. · Politeness and obedience to authority ought to be maintained at all times. · One must never violate the social norms that keep daily order intact. · Following the rules and respecting expectations is what holds us all together.

Security (conservation): We must ensure safety and stability for our society at almost any cost. · National security and public order should always be our highest priority. · A stable, harmonious, and safe community is what everyone must work to protect. · Order and protection from every threat ought to be guaranteed for all. · Safety and stability are the foundation that must never be compromised.

Opinion sentences. I think this phone's battery life is terrible compared to my old one. · The new restaurant downtown is amazing but it is honestly way too expensive. · In my view the latest update made the app much slower than it was before. · That movie was boring and far too long if you ask me. · I feel the mayor is doing a terrible job handling the traffic this year. · This laptop is noticeably faster and smoother than the model I had last year. · The coffee here tastes awful and the service seems painfully slow to me. · I think their new logo looks cheap and much worse than the original one. · This hotel room felt cramped and overpriced for what you actually get. · Honestly the sequel was disappointing and not nearly as good as the first.