The Relevance of Value Systems for Offensive Language Detection

Michael Wiegand
Digital Philology
Faculty of Philological and
Cultural Studies
University of Vienna
AT-1010 Vienna, Austria
michael.wiegand@univie.ac.at

Elisabeth Eder
Austrian Centre for
Digital Humanities
Austrian Academy of Sciences
AT-1010 Vienna, Austria
elisabeth.eder@oeaw.ac.at

Josef Ruppenhofer
Center of Advanced Technology for
Assisted Learning and Predictive Analytics
FernUniversität in Hagen
D-58097 Hagen, Germany
josef.ruppenhofer@fernuni-hagen.de

Abstract

Warning: *This paper contains content that may be offensive or upsetting.*

We examine in how far a person's value system has an impact on their perception of offensiveness. For instance, a scholar is likely to be offended by being accused of reporting unverified claims whereas many non-scholars would not feel that way. Thus, we move away from the assumption that offensiveness can be defined through a universal perspective. Ultimately, such research aims to support personalized approaches to content moderation. Our main contribution is the introduction of a dataset consisting of neutrally-phrased sentences on controversial topics, evaluated by individuals from 4 different value systems. This allows us to identify offensiveness patterns across value systems and conduct classification experiments.

1 Introduction

Offensive language is defined as hurtful, derogatory or obscene utterances made by one person to another person (1)-(2).[1] Closely related terms, e.g. *cyber bullying* (Zhong et al., 2016) or *hate speech* (Waseem and Hovy, 2016), are considered subtypes of offensive language and align with this definition.

(1) stop editing this, you dumbass.
(2) you stupid fucking idiot, fucking kill yourself

The rapid increase of offensive language on the Web makes automatic NLP methods necessary in order to prioritize and direct human moderation efforts toward the most relevant utterances.

If people are offended by some utterance, this may also be unintentional on the part of the speaker, particularly if the utterance does **not appear inherently offensive to everyone.** This kind of offensive language is often linked to people's beliefs and attitudes that a specific group of people all share. For example, a creative artist is likely to be offended by accusations of producing derivative work, while an accountant may be offended by accusations of lacking financial accuracy. Unintentional offense can occur if the speaker is unfamiliar with the beliefs or attitude of the addressed person.

Recent research in *perspectivism* has departed from the assumption of a unique perception of utterances (Kahneman et al., 2021; Rottger et al., 2022; Jhaver et al., 2023). In offensive language detection, for example, remarks about a specific identity group (e.g. *Jews*, *LGBTQ*, *black people* etc.) are often judged differently by their members compared to non-members (Sap et al., 2022; Fleisig et al., 2023; Kanclerz et al., 2022). Also, the cultural background of a person matters (Lee et al., 2024).

In this work, we address **value systems**, i.e. sets of values espoused by particular social groups sharing similar characteristics and a sense of unity. The values we focus on are core values, relatively stable cross-situational concepts of worth held by people. Values are thus abstract ideals (e.g. freedom, helpfulness) that function as important guiding principles (Rokeach, 1973; Schwartz, 1992). A person may hold several value systems that even overlap.

[1] http://thelawdictionary.org

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 6765–6789

	sentence type	example
favouring	approval	(S)he approves of consuming pornography.
	approval::intensive	(S)he is the most vocal supporter of consuming pornography.
	practice	(S)he consumes pornography.
	practice::intensive	Of all people I know, (s)he consumes pornography most frequently.
	let_others_participate	(S)he lets other people consume pornography.
	encourage_others_to_participate	(S)he encourages other people to consume pornography.
disfavouring	disapproval	(S)he disapproves of consuming pornography.
	disapproval::intensive	For him/her, any consumption of pornography is completely unacceptable.
	no_practice	(S)he does not consume pornography.
	no_practice::intensive	(S)he never consumes pornography.
	discourage_others_from_participating	(S)he discourages other people from consuming pornography.
	prevent_others_from_participating	(S)he prevents other people from consuming pornography.

Table 1: Illustration of the 12 sentence types divided into 2 major groups for the topic *pornography*.

In this paper, we study offensive language with respect to different value systems. Unlike previous work (Chulvi et al., 2023), we do not only measure differences in perceptions but we also **determine patterns** explaining the relationship between offensiveness and value systems. Moroever, we evaluate **classifiers tailored to particular value systems**.

We introduce a dataset where the same sentence is rated by representatives of 4 common groups of people that adhere to similar (Western) value systems: *believing Christians*, *conservatives*, *liberals* and *ecologically-conscious people*. In our setup, the raters should imagine that they overhear a conversation about a friend, henceforth referred to as the **target**, who shares the same values and whom they feel inclined to protect. We show that depending on the value system, **the same sentence can be perceived in quite different ways**. Our dataset is based on about 200 topics (e.g. *guns*, *pornography*) for which neutrally-phrased sentences are formed. Each sentence follows one of 12 types, which vary by the degree to which the target is engaged with the topic. This enables us to uncover patterns that explain the relationship between value systems and offensiveness in a structured manner.

Though our work is a **proof of concept**, we anticipate that our findings contribute to the development of personalized content moderation strategies (Jhaver et al., 2023) that are more finely tuned to individual sensitivities regarding offensive material.

In order to quantify the topics that are only offensive depending on one's value systems, we conducted a corpus study. We randomly sampled 500 sentences matching the phrase *accused of X* in the *WebAsCorpus* (Baroni et al., 2009). *X* typically matches a topic considered offensive by the person making the accusation (3). We had each topic rated by 5 crowdworkers from Prolific[2], all English native speakers, as to whether they think that topic is generally perceived offensive. We then computed the majority vote of those ratings. While the majority of the topics were considered generally offensive, e.g. *molesting a minor*, *fraud* or *murder*, 40%, which is a **significant share**, were only considered offensive due to a person's value system, e.g. *being gay*, *not loving Christ* or *being anti-Apple*.

(3) He was accused of molesting a minor.

Our task is framed as a **binary sentence-level classification problem** in which we distinguish between the classes *offensive* and *not offensive*. Unlike previous work, each of the 4 value systems has their distinctive labeling of the sentences.

Value systems introduce **additional contextual factors** to offensive language detection, complementing other interrelated aspects previously examined, such as the geographical origin of a micropost (Waseem and Hovy, 2016), its thread structure (Gao and Huang, 2017; Pavlopoulos et al., 2020), the author's gender (Waseem and Hovy, 2016), multimodal information (Kiela et al., 2020) or the situational embedding (Zhou et al., 2023).

Our **contributions** are the following:

- We introduce the first dataset relating offensive language detection to value systems. It comprises about 2500 sentences.
- Our dataset uses 12 sentence types with varying target involvement, enabling a structured analysis of offensive language perceptions.
- We show that previous classifiers fail to detect this type of offensive language.
- We show how a classifier needs to be configured to detect this type of offensive language.

[2] https://www.prolific.com

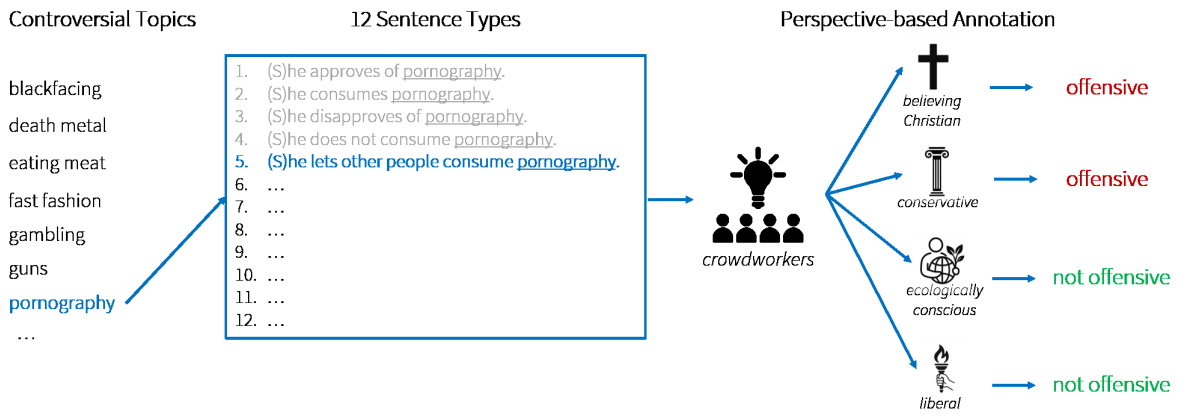


Figure 1: Illustration of the workflow to create the sentence-level dataset.

property	freq.
topics	212
sentences	2532
offensive sentences in at least 1 value system	1339
offensive sentences in all values systems	152

	Christian	Conservative	Ecological	Liberal
offensive	684	702	617	722
not offensive	1842	1825	1914	1803

Table 2: Statistics on the new dataset.

All data and annotation guidelines created as part of this research are **publicly available**.[3]

2 Data

Our dataset is based on a set of about **200 controversial topics** representing various areas of life. (*Appendix B includes the complete list of topics.*) Each topic was manually chosen and our focus was on those topics for which different value systems assign different sentiment. For example, for the topic *abortion*, liberal people predominantly take a pro-choice stance while conservatives tend to hold a pro-life view.

From these topics we derive a sentence-level dataset (§2.1) along with topic-related sentiment scores (§2.2).

2.1 Sentence-Level Dataset

In the following, we describe how we created the sentence-level dataset. The workflow is also illustrated in Figure 1.

For each topic, we formed **12 sentences** that connect a target represented by a third person singular pronoun (i.e. *(s)he*[4]) to that topic. Each sentence type adheres to a specific semantic pattern, as illustrated in Table 1. The **sentence types vary based on the target's degree of involvement with a given topic**. Our assumption is that it is not simply the mention of a topic that makes a sentence offensive to someone but the specific way in which the topic is related to a person in that very sentence.

The sentence types can also be divided into **2 groups** (6+6 sentence types). Either the individual referenced by the pronoun (i.e. the target) **favours** the topic (group 1) or **disfavours** it (group 2). For instance, for the topic *gun*, group 1 basically represents *gun freedom* while group 2 represents *gun restrictions*. This organization thus ensures that for each topic both orientations are considered.

In addition, our sentence types come in pairs, e.g. *approval* and *approval::intensity* which allow us to examine whether the intensity of the type also has an impact on the perceived offensiveness.[5] The motivation is that offensive language is often associated with expressions of higher polar intensity. For example, since *idiotic* has a stronger polar connotation than *less thought-through*, it is more likely to be perceived as offensive.

Each sentence was **formulated neutrally** in that none of them includes any unambiguously offensive word (e.g. *(S)he is stupid*). Neither is there any negative evaluation of the speaker towards the target (e.g. *(S)he behaves inappropriately*).

[3] https://github.com/miwieg/value_systems

[4] We had to use *(s)he* as a gender-neutral representation since crowdworkers tend to confuse the singular *they* (being the most preferred gender-neutral wording) with a plural *they*.

[5] Since the types *let_others_participate* and *prevent_others_from_participating* denote events that are not situated on some scale and thus cannot be intensified or diminished per se, we chose other events that are semantically related to approximate intensification or diminution, i.e. *encourage_others_to_participate* and *discourage_others_from_participating*.

value sys.	topics
Christian	casino gambling, sperm donation, death metal
Conserv.	redefinition of the family, critical race theory, asking personal questions
Ecological	driving big luxury cars, fast fashion, eating meat
Liberal	Breitbart News, Fox News, private prisons
all systems	urinating in public, caning children, blackfacing

Table 3: Topics predominantly offensive only to one value system.

sentence type	Chr.	Con.	Eco.	Lib.	All
favouring					
approval	39.6	33.0	23.6	25.1	30.3
approval::intensive	38.9	40.8	32.5	29.0	35.3
practice	38.6	38.6	21.8	22.3	29.1
practice::intensive	39.7	40.4	34.1	35.1	37.3
let_others_participate	37.8	29.9	20.3	19.3	26.6
encourage_others_to_part.	**42.5**	**41.0**	33.5	35.5	38.1
disfavouring					
disapproval	16.1	19.0	19.9	29.4	21.1
disapproval::intensive	13.2	19.8	22.2	32.5	21.9
no_practice	5.7	6.2	5.2	6.7	6.0
no_practice::intensive	9.2	8.7	6.8	4.8	7.4
discourage_others_from_p.	21.3	22.9	27.0	40.0	27.8
prevent_others_from_part.	22.9	37.7	**45.3**	**62.7**	**42.2**

Table 4: Percentage of offensiveness per sentence type.

The sentences were formulated by one co-author and subsequently cross-checked by another co-author; both are trained linguists. The resulting data were rated with regard to offensiveness via **crowdsourcing**. As a platform, we used Prolific. The crowdworkers should imagine that the sentences were part of a conversation they coincidentally overhear about their friend who also shares similar beliefs and values with the crowdworkers. For each sentence, they should decide whether they perceive the sentence as offensive. (*The full instruction is shown in Figure 4 in Appendix G.*) We conceived a scenario in which the crowdworkers overhear a conversation about a friend rather than themselves, as in this case, the crowdworkers tend to be protective and may more readily admit that they find a remark offensive (Rojas et al., 1996; McLeod et al., 2001; Zhang, 2023).

Each sentence was rated by crowdworkers, English native speakers living in the US, who identify with one of 4 value systems: believing Christians[6], conservatives, liberals[7] or ecologically-conscious people.[8] We chose liberals and conservatives since they are regarded as the two major political leanings in the U.S. (Conover and Feldman, 1981). Christians were selected as they represent the largest religious group in the U.S. (Gallup, 2023). Ecologically-conscious individuals were included because a majority of people consider climate change a serious threat to humanity (Ritchie, 2024).

In selecting our value systems, we chose those that are distinct from one another. However, this does not imply that the members of different value systems are entirely separate, as **individuals can often adhere to multiple value systems**. Nonetheless, each crowdworker was recruited to represent exactly one of the four value systems. Beyond the fact that the 4 value systems represent distinct concepts (political ideology, religion and environmental awareness), they also exhibit clear behavioural differences. For example, while both conservatives and ecologically-conscious people may prioritize sustainability, conservatives often focus on economic sustainability through market solutions, whereas ecologically-conscious people prioritize environmental sustainability through regulations. In addition, we ensured that the value systems are mutually distinct by confirming that no value system is identical to, or a proper subset of, another (see statistics in Appendix D).

We do not claim that our chosen value systems cover any of the dimensions we consider, e.g. political ideologies or religion, completely. We merely want to show that distinctive value systems have an impact on the perception of offensiveness.

The sentences of our dataset were presented in random order. A sentence could be labeled as either *offensive*, *not offensive*, *cannot decide* or *not proper English*. **Each sentence is assigned 4 labels, where each label corresponds to the majority vote of the 5 crowdworkers sharing the same value system.** Thus, in total, each sentence is rated by 20 crowdworkers. We split our task in subtasks of 100 sentences so as to allow more than 20 crowdworkers annotate our dataset, in fact, approx. 250 participated. For the final dataset, **only sentences in which the majority vote was either *offensive* or *not offensive* were retained** but this number varies slightly among our 4 value systems.

Table 2 provides some statistics on the resulting dataset. While 1339 sentences have been rated offensive by at least one value system, only 152 out

[6]We recruited members of high-commitment Christian groups in order to exclude *nominal* Christians from our survey; see Appendix E for more details.

[7]The terms *conservative* and *liberal* were interpreted using Western, especially US, values.

[8]Appendix E provides more details on our crowdworkers.

of 2532 sentences have been found offensive by all value systems. This underscores that the perception of offensiveness varies quite notably across different value systems. For illustration, Table 3 lists some topics that are predominantly perceived offensive only by one of the 4 value systems.

The label distributions of Christians and conservatives exhibit the strongest similarities as do the distributions of liberal and ecologically-conscious people, i.e. a correlation of 0.58 and 0.57 using Spearman's ρ.[9] This can be explained by the notable ideological similarities within these two pairs.

We took a random sample of 200 sentences. To validate our approach, for each value system, we asked 5 additional crowdworkers (who identify with that value system) to rate the sentences. Then, we computed another majority vote based on their ratings. We reached a **substantial interannotator agreement** (McHugh, 2012) of at least Cohen's $\kappa = 0.61$ between the two different majority votes of each value system.[10]

Table 4 lists the proportion of offensiveness across the different sentence types. The table shows that the proportion varies notably. Often, it is larger on *favouring* sentence types. Intuitively, intensive variants of a sentence type are also rated offensive more often. Interestingly, the most offensive type for Christians and conservative people is *encourage_others_to_participate* while it is *prevent_others_from_participating* for ecologically-conscious and liberal people. It suggests that Christians and conservatives are more inclined to support bans. This aligns with psychological research suggesting a tendency linked to broader authoritarian traits (Altemeyer and Hunsberger, 2009; Ludeke et al., 2013). Both of the above sentence types also imply more **active involvement of the target**. This is known to meet with greater rejection than passive involvement (Thomson, 1976; Spranca et al., 1991).

2.2 Topic-Related Sentiment Scores

We also have our topics separately rated with regard to sentiment. This is done since we want to examine the relation between sentiment and offensiveness. For this annotation, crowdworkers identifying with either of the 4 value systems were asked to rate each topic *a priori* (out of context) with regard to their sentiment towards it on a Likert-scale, where 1 corresponds to *very negative* and 5 corresponds to *very positive*. In order not to bias this dataset, crowdworkers who had participated in the creation of the sentence-level dataset (§2.1) were excluded from this task. We averaged over ratings of 5 crowdworkers for each value system. The result is a dataset in which each topic has an average (prior) sentiment score according to each of the 4 value systems. For each value system, we also computed Spearman's ρ between the average sentiment score and an additional sentiment score that had been computed on the basis of 5 different crowdworkers sharing the same value system. For each value system, we obtained a **strong agreement**.[11]

Our research comprises analyses on both coarse-grained and fine-grained sentiment information. For the latter, we use the average sentiment scores. For the former, these average scores (ranging from 1 to 5) are converted into 3 discrete categories, i.e. positive, neutral and negative sentiment, by splitting the range into 3 equally-sized bins. The lower bin corresponds to negative, the middle bin to neutral and the upper bin to positive sentiment.

Table 5 displays, for each coarse-grained sentiment category, the proportion of offensive sentences across the different sentence types averaged over all value systems. The table shows a clear divide between favouring and disfavouring sentence types. For the former, negative sentiment tends to be most offensive, while for the latter, positive sentiment tends to be most offensive. This means that **sentences conveying a favouring attitude of the target towards topics that are considered negative are perceived offensive**. The same holds for sentences conveying a disfavouring attitude of the target towards topics that are considered positive.

3 Experiments

3.1 The Different Sentence Classifiers

For standard supervised classification, we use **DeBERTa** (Hea et al., 2021), one of the most advanced publicly available transformers. We **fine-tune** the pretrained model on the given training data using the FLAIR-framework (Akbik et al., 2019) with the hyperparameter settings from Wiegand et al. (2022), a study related to ours. We report the average over 5 training runs (+ standard deviation). ***Appendix A contains details on the settings of all classifiers.***

[9] Appendix C lists the correlation for all possible pairs.

[10] The individual scores are listed in Appendix F.1.

[11] The individual scores are listed in Appendix F.2.

	sentence type	pos.	neut.	neg.
favouring	approval	1.3	8.7	**15.7**
	approval::intensive	2.3	12.3	**15.9**
	practice	1.6	8.2	**14.8**
	practice::intensive	4.4	12.5	**15.4**
	let_others_participate	1.6	7.5	**13.5**
	encourage_others_to_participate	3.4	13.5	**16.5**
disfavouring	disapproval	**16.6**	5.0	7.3
	disapproval::intensive	**12.9**	7.3	2.2
	no_practice	**5.0**	0.9	0.4
	no_practice::intensive	**6.0**	0.9	0.3
	discourage_others_from_participat.	**20.2**	7.3	1.5
	prevent_others_from_participating	**24.8**	15.9	2.7

Table 5: Percentage of offensiveness within (topic-related) sentiment categories across sentence types.

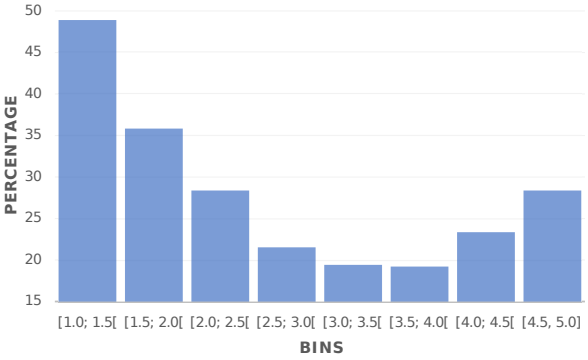


Figure 3: Percentage of offensive sentences within bins of a particular range of sentiment scores.

```
procedure classify(sentence, valueSystem)
    offensive ← FALSE
    topic ← getTopic(sentence)
    sentiment ← getSentiment(topic, valueSystem)
    group ← getSentenceTypeGroup(sentence)
    if group == FAVOUR and sentiment == NEG then
        offensive ← TRUE
    else if group == DISFAVOUR and sentiment == POS then
        offensive ← TRUE
    return offensive
```

Figure 2: Rule-based sentiment classifier.

LM	prompt	Chr.	Con.	Eco.	Lib.	Avg.
DeepSeek V3	value-agnostic	59.4	62.0	66.8	66.1	63.6
DeepSeek V3	value-specific	60.2	69.1	63.2	65.4	64.5
Qwen3	value-agnostic	60.8	63.0	68.5	70.8	65.8
LLaMA 4	value-specific	64.6	69.7	61.6	68.8	66.2
LLaMA 4	value-agnostic	62.1	64.3	69.2	70.3	66.5
GPT-5	value-agnostic	65.0	68.1	70.2	70.8	68.5
Qwen3	value-specific	70.0	72.9	64.4	70.8	69.5
GPT-4	value-agnostic	66.1	67.5	72.1	72.9	69.7
GPT-4	value-specific	71.0	**76.4**	68.5	75.4	72.8
GPT-5	value-specific	**72.0**	74.4	**74.7**	**78.8**	**75.0**

Table 6: F1-score of zero-shot classifiers.

In the following subsections, we describe the specific classifiers we examine in this work:

3.1.1 Human Classifier

As an upper bound, we tested a human classifier in which we randomly sampled the judgment of one individual annotator of the respective value system from the crowdsourced gold-standard annotation. This individual judgement may notably differ from the gold standard label which is the majority label of 5 annotators.

3.1.2 Shared Offensiveness

Since existing classification approaches for offensive language detection typically only consider one definite class label for each utterance, we created a baseline that reflects this behaviour. We examine **2 variants**:

The **automatic** variant is a classifier trained on the training data of our dataset in which only those instances are considered offensive that have been judged as offensive by all 4 value systems.

The **oracle** variant is not trained on training data but draws its knowledge directly from the gold-standard annotation of our dataset. It classifies a test instance as offensive if it has been judged as offensive by all 4 value systems in the gold standard.

3.1.3 Sentence-Type Classifier

Since we observed a notable difference between the proportion of offense among the different sentence types (Table 4), we construct a feature-based classifier (more specifically, we used *logistic regression*) that is exclusively trained on the sentence type of a sentence.

3.1.4 Topic Classifier

Our topic classifier considers all sentences associated with a specific topic either as offensive or not. Thus, it represents a baseline that ignores the specific contextualization of a topic. We consider a topic as offensive if more than 50% of its related sentences are offensive according to the gold standard label for the given value system.

3.1.5 Sentiment Analysis

We employ a classifier based on topic-related sentiment. Figure 2 outlines this rule-based classifier that simply encodes the regularity from Table 5 as discussed in §2.2. Specifically, sentences that express a favouring attitude of the target toward topics considered negative are predicted as offensive, as are sentences that express a disfavouring attitude of the target toward topics considered positive.

3.1.6 Existing Classifiers for Offensive Language Detection

We consider 2 publicly available tools: **PerspectiveAPI**,[12] a tool for the general detection of offensive language, and the most recent transformer for implicitly offensive language detection focusing on identity groups from Hartvigsen et al. (2022), i.e. HateBERT fine-tuned on **ToxiGen**.

3.1.7 Existing Datasets for Offensive Language Detection

We also fine-tune a transformer on two existing datasets for offensive language detection. We chose **OLID** (Zampieri et al., 2019) and the dataset introduced by **Founta et al. (2018)**. Unlike many other available datasets, these do not focus exclusively on hate speech directed at specific identity groups (e.g. sexism, racism or antisemitism). Therefore, these two datasets better align with our broader notion of offensive language.

3.1.8 GPT-4 & GPT-5

Given the high effectiveness of recent GPT-models (Laskar et al., 2023), we also examine classifiers based on the language models GPT-4 and GPT-5. The prompt we use for the 4 value systems is (4). As a baseline we use a *value-agnostic* prompt (5) in order to measure whether there is any benefit of value-specific prompts.

(4) **value-specific prompt:** Would `<value_system>` find it offensive if this were said about them or their like-minded friends? Please answer with "Yes" or "No". In addition, please briefly motivate your answer.

(5) **value-agnostic prompt:** Would the average person find it offensive if this were said about them or their like-minded friends? Please answer with "Yes" or "No". In addition, please briefly motivate your answer without differentiating between individuals.

Our first classifier based on GPT is a **zero-shot** approach (Plaza-del-arco et al., 2023) that simply maps completions to the prompt containing *Yes* as *offensive* and those containing *No* as *not offensive*.

In our second method, we are **augmenting** each sentence from our dataset by the completion we obtained from the zero-shot approach. We then fine-tune and test a transformer on these augmented instances. The augmented text may give a classifier additional helpful clues. Instead of just concatenating the full completion to the original sentence, we also consider a variant in which we only concatenate the **immediate response** (i.e. *Yes*/*No*).

[12] https://perspectiveapi.com

3.1.9 Other LLMs

Since GPT-4 and GPT-5 are proprietary models, we also use the following recent open-weight LLMs: DeepSeek V3 (DeepSeek-AI et al., 2025), LLaMA 4 (Touvron et al., 2023) and Qwen3 (Yang et al., 2025). All models are evaluated using the same prompts as those applied to GPT-4 and GPT-5.

3.2 Evaluation on the Sentence-Level Dataset

We always report macro-average F1-score.

Table 6 compares the different zero-shot classifiers. GPT-5 outperforms the other language models. Notably, only for GPT-5 do we observe an increase in performance with value-specific prompts compared to value-agnostic ones across all value systems. Due to its superior performance, we are utilizing the GPT-5-based value-specific variant for GPT-5::augmenting in upcoming experiments.

Table 7 displays the performance of the further classifiers. For all classifiers that are both trained and tested on our dataset, we carry out a 5-fold cross-validation. The folds were arranged so that similar topics are kept within the same fold. This enables a strict evaluation in which the test fold comprises topics not observed during training.

Table 7 shows that all classifiers trained on other existing datasets perform poorly. Despite their simplicity, the sentence-type and the topic classifier outperform several of these classifiers which underscores that sentence-type and topic information are relevant for this task. Even more noteworthy is the performance of the sentiment classifier. Regarding shared offensiveness, we find that there is huge difference between a realistic automatic version and the oracle version. A simple plain supervised classifier is not very effective. Training on the dataset augmented with (untruncated) GPT-5 completions represents the best fully automatic classifier. If we add to this classifier, which is apparently still too restrictive, all those offensive-predictions of the oracle version of the shared-offensiveness classifier, we obtain performance close to humans. Apparently, classifiers tailored for a particular value system still seem to lack the notion of what is generally considered offensive. Our research suggests this typically involves concepts of lesser significance, e.g. (6).

(6) (S)he eats horse meat. (*offensive in all value systems*)

3.3 Discussion

Although our sentiment classifier suggests that associating people with things they dislike, or disso-

classifier	Christian	Conservative	Ecological	Liberal	Average
majority-class classifier	42.2	41.9	43.1	41.7	42.2
Founta et al. (2018)*	47.0 (±5.9)	46.5 (±5.7)	46.0 (±3.7)	45.3 (±4.9)	46.2 (±5.0)
OLID*	55.5 (±6.7)	54.5 (±0.5)	51.3 (±4.2)	52.3 (±5.3)	53.4 (±4.2)
ToxiGen	56.0	54.3	53.9	53.9	54.5
shared offensiveness::automatic*	58.8 (±1.6)	61.1 (±1.9)	56.8 (±3.1)	58.3 (±2.1)	58.7 (±2.2)
topic classifier (*each topic either offensive or not*)°	61.0	61.2	58.6	58.9	59.9
sentence type†°	60.1	60.2	59.6	61.4	60.3
PerspectiveAPI	63.5	63.4	63.9	63.3	63.5
plain supervised classification on the given dataset*°	65.5 (±2.9)	66.8 (±2.5)	66.8 (±1.0)	71.0 (±2.0)	67.5 (±2.4)
sentiment analysis°	74.5	68.7	67.0	70.0	70.0
GPT-5::augmenting (*only immediate response: 'Yes'/'No'*)*°	67.8 (±3.2)	74.5 (±1.7)	69.8 (±1.8)	77.0 (±1.8)	72.3 (±1.7)
shared offensiveness::oracle	72.4	72.1	73.2	71.7	72.4
GPT-5::zero-shot::value-specific°	72.0	74.4	74.7	78.8	75.0
GPT-5::augmenting*° (*using the full completion*)	76.0 (±0.8)	76.8 (±0.2)	74.8 (±0.4)	80.1 (±0.4)	76.9 (±0.5)
GPT-5::augmenting*° + shared offensiveness::oracle	**78.7** (±1.1)	**78.4** (±0.2)	**77.2** (±0.5)	**82.7** (±0.3)	**79.2** (±0.5)
human classifier°	81.1	81.4	78.8	83.7	81.2

Table 7: F1-score of different classifiers (**: fine-tuned with DeBERTa; †: trained using logistic regression; °: for each value system there is a distinct classifier*).

ciating them from things they like, might result in offensive statements, Figure 3 indicates the degree to which a person likes or dislikes something also plays a significant role. For instance, associating someone with an item they *mildly* dislike is unlikely to produce an offensive statement. Moreover, since the sentiment classifier lags behind human performance (Table 7), we conclude that **sentiment analysis and our task are not the same**.

The language models we use, particularly in zero-shot classification (Table 6), tend to perform best on the liberal value system. This is in line with previous work (Motoki et al., 2023; Santurkar et al., 2023). However, this **liberal bias** has not yet been reported for offensive language detection.

3.4 Using Texts Invented by Crowdworkers

The dataset that we proposed for this study may be criticized for being artificial in that we created it based on predefined sentence types. In order to show that this makes our dataset more difficult, we created a small dataset (Table 8) of similar offensive utterances in which we asked crowdworkers to invent such offensive remarks for the same situation we used for our original dataset (§2.1). The resulting sentences were rated by further crowdworkers in the same way as our proposed dataset.

Table 9 compares key classifiers between our proposed dataset and the crowdworkers' dataset.[13] In contrast to our dataset, the best (fully automatic) GPT-based classifier performs on a par with human evaluation, indicating that automatic classification is simpler on that dataset. We observed that this dataset exhibits properties that facilitates correct predictions by language models. Unlike our proposed dataset, the invented dataset is more skewed towards a small set of very generally discussed topics. For instance, 18% of the sentences address *pollution* while in our proposed dataset, it is only 7%. For *pollution*, it is easier for a language model to know what value systems feel offended than for rare topics, such as *hormone replacement therapy* (only part of our proposed dataset). Moreover, by manual inspection we found that over 25% of the offensive sentences created by the crowdworkers exhibit extra clues, such as negative sentiment towards the target (7) or contradictory behavior (8). Such clues are absent in our proposed dataset that solely includes sentences formulated neutrally.

(7) (S)he is obsessed with reducing your carbon footprint.
(8) (S)he's always talking about climate change, but I bet (s)he still takes long showers.

4 Related Work

There has been a large body of research for offensive language detection (Nobata et al., 2016; Badjatiya et al., 2017; Fortuna and Nunes, 2018). The detection of implicitly offensive language (Waseem et al., 2017; ElSherief et al., 2021), i.e. offensive language not conveyed by offensive words (e.g. slurs), remains challenging (van Aken et al., 2018; Wiegand et al., 2021b; Ocampo et al., 2023). Our research is situated in implicitly offensive language. Previous work has focused on other subtypes, such as stereotypes (Davani et al., 2022; Hartvigsen et al., 2022), dogwhistles (Mendelsohn et al., 2023) or comparisons (Wiegand et al., 2021a).

[13] Many previously used classifiers cannot be applied to this new dataset due to the unrestricted sentence construction.

Within offensive language detection research, our work is most closely related to Weerasooriya et al. (2023), who show that both human and machine moderators often disagree on what constitutes offensive speech, especially due to political leanings. Our study differs in several key ways: we cover 4 value systems instead of 2, and more than 200 topics instead of just 2. By using a structured set of 12 sentence types, we can identify specific perception patterns across these value systems (Table 4). We also introduce the central sentiment rule (Figure 2), which is absent in Weerasooriya et al. (2023). Finally, while they asked their annotators to judge from a political opponent's perspective, we asked ours to judge from their own.

Our work is also related to stance detection (Mohammad et al., 2016; Küçük and Can, 2020) where the task is to determine a person's viewpoint towards an entity or proposition in a text. (This is also referred to as ideology (Iyyer et al., 2014; Shen and Rose, 2021).) In our work, we consider the perspective of a *listener* hearing a statement. The crucial element is determining whether the listener perceives the utterance as offensive based on their stance on what is being said. Unlike stance detection, we do not deduce a fixed attitude from the statement. Instead, we explore variable sentiment interpretations based on specific value systems.

The topics that form the basis of our dataset, e.g. *pornography* or *guns*, do not convey an unambiguous sentiment, yet, given a person's value system, they are considered *positive*, *neutral* or *negative*. In that respect, our work also bears some resemblance to implicit sentiment (Deng et al., 2013; Ding and Riloff, 2018; Zhou et al., 2021).

Value systems are also connected to moral values, i.e. principles that govern what is considered right and wrong behavior. Morality has also recently been addressed in NLP (Hoover et al., 2020; Hendrycks et al., 2021; Liscio et al., 2022, 2023; Reinig et al., 2024). However, the relationship between morality and offensiveness has only been examined in a descriptive way (Kennedy et al., 2023).

We differentiate values from attitudes and ideologies. Following Maio et al. (2003), we take attitudes as tendencies to evaluate an object, while we understand ideologies as systems of attitudes and values that are organized around an abstract theme (e.g. liberalism). Importantly, attitudes, values and ideologies are evaluative constructs that are not disconnected: influences potentially flow *upwards* from attitudes via values to ideology or *downwards* from ideology via values to attitudes.

property				**freq.**
sentences				753
offensive sentences in at least 1 value system				517
offensive sentences in all values systems				36
	Christian	Conservative	Ecological	Liberal
offensive	167	236	241	314
not offensive	576	508	481	414

Table 8: Statistics on the smaller crowdworkers' dataset.

classifier	**proposed**	**crowdw.**
majority-class classifier	42.2	40.1
Founta et al. (2018)*	46.2 (±5.0)	50.7 (±6.6)
shared offensiveness::automatic*	58.7 (±2.2)	51.0 (±7.5)
PerspectiveAPI	63.5	52.1
ToxiGen	54.5	55.5
OLID*	53.4 (±4.2)	58.9 (±1.5)
plain supervised classification*	67.5 (±2.4)	64.9 (±2.3)
shared offensiveness::oracle	72.4	69.0
GPT-5::augmenting*	76.9	77.2
human classifier	81.2	78.1

Table 9: Comparison of F1-scores (averaged over all value systems) between the proposed and the crowdworkers' dataset *(*: fine-tuned with DeBERTa)*.

5 Conclusion

We introduced a dataset for evaluating the perception of offensiveness across different value systems. We found that sentences are deemed offensive when they link individuals to disliked concepts or separate them from favoured ones. The level of engagement also plays a crucial role. Classifiers developed using previous datasets fail to accurately predict offensiveness across diverse value systems. Only a few sophisticated LLMs, such as GPT-5, demonstrated an ability to classify offensiveness effectively.

6 Limitations

Our dataset **only addresses one subset of offensive language**, i.e. language that is only potentially offensive to people depending on their value system. Therefore, classifiers trained on our new data will only be capable to detect this subset of offensive language rather than offensive language, in general. Thus, we follow Wiegand et al. (2021b) who argue that a *divide-and-conquer approach* is the only reasonable approach to complex phenomena such as implicitly offensive language.

Our newly curated **dataset encompasses just 4 value systems** prevalent in the US. While there are numerous other potential value systems, each

with its unique sentiment and offensiveness profile, it was not feasible to include them all within the scope of this study. Our selection was further limited by the availability of crowdworkers on Prolific. For more specialized value systems, we lacked the requisite number of crowdworkers to evaluate the entire dataset comprehensively. It is essential to emphasize that our research is a **proof of concept**. We do not purport to provide an exhaustive representation.

At first glance, our model suggests that there are two main value systems: conservatives and liberals, with Christians and ecologically-conscious individuals being subsets of these two. However, this *oversimplification* overlooks other combinations, such as ecologically-conscious conservatives (*Green Conservatives*). Individuals can embody multiple value systems simultaneously, like being both Christian and conservative. Moreover, we provided statistics in Appendix §D showing that 4 value systems are distinctive categories and that there is no subset-relationship between pairs of these 4 systems. Unfortunately, the sparsity of crowdworkers prevented a deeper exploration of other possible value system combinations.

In this study, we estimate potentially offensive content based on the **assumption that an individual's value system is primarily influenced by their political or religious beliefs**. While this provides a foundational understanding, it is admittedly a simplification. There are myriad factors, such as personal experiences, that can influence an individual's perception of offensive remarks. Nonetheless, it remains challenging for NLP to capture personal attitudes and beliefs at such a detailed level.

Our dataset was created by embedding a set of controversial topics into a pre-defined set of sentence structures. Additionally, each sentence is expressed within a specific situational context. Despite potential criticisms due to its **artificial nature**, this methodology enables a systematic exploration of perception shifts across different sentence forms and we believe that the sentence types we selected largely encompass the most distinct ways of introducing a given topic within a sentence. Direct extraction of sentences from textual corpora would introduce a prohibitive level of variability, obstructing our ability to isolate the effects of specific linguistic alterations while keeping the remainder of the sentence constant. Furthermore, data sourced directly would likely resemble our alternative dataset, where crowdworkers generated sentences spontaneously (§3.4). This resulted in significantly simplified classifications due to additional contextual clues and topic biases.

7 Ethical Considerations

Most of our new gold standard data was created with the help of crowdsourcing. All crowdworkers were compensated following the wage recommended by the crowdsourcing platform Prolific (i.e. $12 per hour). We inserted a warning of the offensive nature in the task advertisement.

Our research involves **categorizing individuals based on shared value systems**. While such categorization may raise ethical concerns, particularly regarding the risk of stereotyping or prejudgment, it is important to clarify the underlying motivation of our work. Our primary objective is foundational research aimed at developing software that can alert users when their written content may be offensive to others due to differing value systems. We explicitly do not aim to generate, reinforce or legitimize exaggerated or arbitrary associations between specific groups and particular topics.

Furthermore, our work aims to highlight that the prior assumption of a unanimous agreement on what is deemed offensive is an oversimplification.

For our proposed dataset (§2.1), we actually created sentences that are potentially offensive to people ourselves. In §3.4, we have crowdworkers create similar textual data. Creating morally disputable content as part of research is not unheard of. Both in plagiarism detection (Potthast et al., 2010), deception detection (Ott et al., 2011) and offensive language detection itself (Vidgen et al., 2021; Wiegand et al., 2021a) a procedure similar to ours was pursued. Moreover, we found no alternative method that would yield a dataset with a comparable size and quality.

8 Acknowledgements

The authors were partially supported by the Austrian Science Fund (FWF): P 35467-G. The authors would like to thank Julia Jaremko for their feedback on earlier drafts of this paper.

References

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. FLAIR: An easy-to-use framework for state-of-the-art NLP. In *Proceedings of the Human Language*

Technology Conference of the North American Chapter of the ACL (HLT/NAACL), pages 54–59, Minneapolis, MN, USA.

Bob Altemeyer and Bruce Hunsberger. 2009. Authoritarianism, Religious Fundamentalism, Quest, and Prejudice. *The International Journal for the Psychology of Religion*, 2(2):113–133.

Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. Deep Learning for Hate Speech Detection in Tweets. In *Proceedings of the International Conference on World Wide Web (WWW)*, pages 759–760, Perth, Australia.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetti. 2009. The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora. *Language Resources and Evaluation*, 43(3):209–226.

Berta Chulvi, Lara Fontanella, Roberto Labadie-Tamayo, and Paolo Rosso. 2023. Social or Individual Disagreement? Perspectivism in the Annotation of Sexist Jokes. In *Proceedings of the Workshop on Perspectivist Approaches to NLP*, Kraków, Poland.

Pamela Johnston Conover and Stanley Feldman. 1981. The Origins and Meaning of Liberal/Conservative Self-Identifications. *American Journal of Political Science*, 25(4):617–645.

Aida Mostafazadeh Davani, Mohammad Atari, Brendan Kennedy, and Morteza Dehghani. 2022. Hate Speech Classifiers Learn Normative Social Stereotypes. *Transactions of the Association for Computational Linguistics (TACL)*, 11:300–319.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Haowei Zhang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Li, Hui Qu, J. L. Cai, Jian Liang, Jianzhong Guo, Jiaqi Ni, Jiashi Li, Jiawei Wang, Jin Chen, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, Junxiao Song, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Lei Xu, Leyi Xia, Liang Zhao, Litong Wang, Liyue Zhang, Meng Li, Miaojun Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Ning Tian, Panpan Huang, Peiyi Wang, Peng Zhang, Qiancheng Wang, Qihao Zhu, Qinyu Chen, Qiushi Du, R. J. Chen, R. L. Jin, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, Runxin Xu, Ruoyu Zhang, Ruyi Chen, S. S. Li, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shaoqing Wu, Shengfeng Ye, Shengfeng Ye, Shirong Ma, Shiyu Wang, Shuang Zhou, Shuiping Yu, Shunfeng Zhou, Shuting Pan, T. Wang, Tao Yun, Tian Pei, Tianyu Sun, W. L. Xiao, Wangding Zeng, Wanjia Zhao, Wei An, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, X. Q. Li, Xiangyue Jin, Xianzu Wang, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaojin Shen, Xiaokang Chen, Xiaokang Zhang, Xiaosha Chen, Xiaotao Nie, Xiaowen Sun, Xiaoxiang Wang, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xingkai Yu, Xinnan Song, Xinxia Shan, Xinyi Zhou, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, Y. K. Li, Y. Q. Wang, Y. X. Wei, Y. X. Zhu, Yang Zhang, Yanhong Xu, Yanhong Xu, Yanping Huang, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Li, Yaohui Wang, Yi Yu, Yi Zheng, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Ying Tang, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yu Wu, Yuan Ou, Yuchen Zhu, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yukun Zha, Yunfan Xiong, Yunxian Ma, Yuting Yan, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Z. F. Wu, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhen Huang, Zhen Zhang, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhibin Gou, Zhicheng Ma, Zhigang Yan, Zhihong Shao, Zhipeng Xu, Zhiyu Wu, Zhongyu Zhang, Zhuoshu Li, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Ziyi Gao, and Zizheng Pan. 2025. DeepSeek-V3 Technical Report. *Preprint*, arXiv:2412.19437.

Lingjia Deng, Yoonjung Choi, and Janyce Wiebe. 2013. Benefactive/Malefactive Event and Writer Attitude Annotation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 120–125, Sofia, Bulgaria.

Haibo Ding and Ellen Riloff. 2018. Weakly Supervsied Induction of Affective Events by Optimizing Semantic Consistency. In *Proceedings of the National Conference on Artificial Intelligence (AAAI)*, pages 5763–5770, New Orleans, LA, USA.

Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. 2021. Latent Hatred: A Benchmark for Understanding Implicit Hate Speech. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 345–363, Online and Punta Cana, Dominican Republic.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. LIBLINEAR: A Library for Large Linear Classification. *Journal of Machine Learning Research*, 9:1871–1874.

Eve Fleisig, Rediet Abebe, and Dan Klein. 2023. When the Majority is Wrong: Modeling Annotator Disagreement for Subjective Tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6715–6726, Singapore.

Paula Fortuna and Sérgio Nunes. 2018. A Survey on Automatic Detection of Hate Speech in Text. *ACM Computing Surveys*, 51(4):85:1–85:30.

Antigoni-Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael

Sirivianos, and Nicolas Kourtellis. 2018. Large Scale Crowdsourcing and Characterization of Twitter Abusive Behavior. In *Proceedings of the International AAAI Conference on Weblogs and Social Media (ICWSM)*, Stanford, CA, USA.

Gallup. 2023. How Religious Are Americans? https://news.gallup.com/poll/358364/religious-americans.aspx. Accessed: 26-Aug-2024.

Gallup. 2024. Church Attendance Has Declined in Most U.S. Religious Groups. https://news.gallup.com/poll/642548/church-attendance-declined-religious-groups.aspx. Accessed: 27-Aug-2024.

Lei Gao and Ruihong Huang. 2017. Detecting Online Hate Speech Using Context Aware Models. In *Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP)*, pages 260–266, Varna, Bulgaria.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. ToxiGen: A Large-Scale Machine-Generated Dataset for Adversarial and Implicit Hate Speech Detection. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 3309–3326, Dublin, Ireland.

Pengcheng Hea, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTa: Decoding-enhanced BERT with Disentangled Attention. In *International Conference on Learning Representations*.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2021. Aligning AI with Shared Human Values. In *Proceedings of the International Conference on Learning Representations (ICLR)*, Virtual Event, Austria.

Joe Hoover, Gwenyth Portillo-Wightman, Leigh Yeh, Shreya Havaldar, Aida Mostafazadeh, Ying Lin, Davani, Brendan Kennedy, Mohammad Atari, Zahra Kamel, Madelyn Mendlen, Gabriela Moreno, Christina Park, Tingyee E. Chang, Jenna Chin, Christian Leung, Jun Yen Leung, Arineh Mirinjian, and Morteza Dehghani. 2020. Moral Foundations Twitter Corpus: A collection of 35k tweets annotated for moral sentiment. *Social Psychological and Personality Science*, 11(8):1057–1071.

Mohit Iyyer, Peter Enns, Jordan Boyd-Graber, and Philip Resnik. 2014. Politicial Ideology Detection Using Recursive Neural Networks. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1113–1122, Baltimore, MD, USA.

Shagun Jhaver, Alice Qian Zhang, Quan Ze Chen, Nikhila Natarajan, Ruotong Wang, and Amy X. Zhang. 2023. Personalizing Content Moderation on Social Media: User Perspectives on Moderation Choices, Interface Design, and Labor. In *Proceedings of the ACM on Human-Computer Interaction*, volume 7, pages 1–33.

Daniel Kahneman, Olivier Sibony, and Cass R. Sunstein. 2021. *Noise: A Flaw in Human Judgment*. Little, Brown and Company.

Kamil Kanclerz, Marcin Gruza, Konrad Karanowski, Julita Bielaniewicz, Piotr Milkowski, Jan Kocon, and Przemyslaw Kazienko. 2022. What if Ground Tuth is Subjective? Personalized Deep Neural Hate Speech Detection. In *Proceedings of the Workshop on Perspectivist Approaches to NLP*, pages 37–45, Marseille, France.

Brendan Kennedy, Preni Golazizian, Jackson Trager Mohammad Atari, Joe Hoover, Aida Mostafazadeh Davani, and Morteza Dehghani. 2023. The (moral) language of hate. *PNAS Nexus*, 2:1–16.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes. In *Advances in Neural Information Processing Systems*, volume 33, pages 2611–2624, Vancouver, Canada.

Dilek Küçük and Fazli Can. 2020. Stance Detection: A Survey. *ACM Computing Surveys*, 53(1):12:1–12:37.

Md Tahmid Rahman Laskar, M Saiful Bari, Mizanur Rahman, Md Amran Hossen Bhuiyan, Shafiq Joty, and Jimmy Huang. 2023. A Systematic Study and Comprehensive Evaluation of ChatGPT on Benchmark Datasets. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 431–469, Toronto, Canada.

Nayeon Lee, Chani Jung, Junho Myung, Jiho Jin, Jose Camacho-Collados, Juho Kim, and Alice Oh. 2024. Exploring Cross-Cultural Differences in English Hate Speech Annotations: From Dataset Construction to Analysis. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4205–4224, Mexico City, Mexico.

Enrico Liscio, Oscar Araque, Lorenzo Gatti, Ionut Constantinescu, Catholijn Jonker, Kyriaki Kalimeri, and Pradeep Kumar Murukannaiah. 2023. What does a Text Classifier Learn about Morality? An Explainable Method for Cross-Domain Comparison of Moral Rhetoric. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 14113–14132, Toronto, Canada.

Enrico Liscio, Alin Dondera, Andrei Geadau, Catholijn Jonker, and Pradeep Murukannaiah. 2022. Cross-Domain Classification of Moral Values. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2727–2745, Seattle, WA, USA.

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*.

Steven Ludeke, Wendy Johnson, and Thomas J. Bouchard Jr. 2013. "Obedience to traditional authority:" a heritable factor underlying authoritarianism, conservatism and religiousness. *Personality and Individual Differences*, 55(4):375–380.

Gregory R. Maio, James M. Olsen, Mark M. Bernard, and Michelle A. Luke. 2003. Ideologies, Values, Attitudes, and Behavior. In John DeLamater, editor, *Handbook of Ssocial Psychology*, pages 283–308. Kluwer Academic/Plenum Publishers.

Mary L. McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia Medica*, 22(3):276–282.

Douglas M. McLeod, Benjamin H. Detenber, and William P. Eveland Jr. 2001. Behind the Third-Person Effect: Differentiating Perceptual Processes for Self and Other. *Journal of Communication*, 51(4):678–695.

Julia Mendelsohn, Rona Le Bras, Yejin Choi, and Maarten Sap. 2023. From Dogwhistles to Bullhorns: Unveiling Coded Rhetoric with Language Models. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 15162–15180, Toronto, Canada.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. SemEval-2016 Task 6: Detecting Stance in Tweets. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval)*, pages 31–41, San Diego, CA, USA.

Fabio Motoki, Valdemar Pinho Neto, and Victor Rodrigues. 2023. More human than human: measuring ChatGPT political bias. *Public Choice*.

Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive Language Detection in Online User Content. In *Proceedings of the International Conference on World Wide Web (WWW)*, pages 145–153, Republic and Canton of Geneva, Switzerland.

Nicolas Ocampo, Ekaterina Sviridova, Elena Cabrio, and Serena Villata. 2023. An In-depth Analysis of Implicit and Subtle Hate Speech Messages. In *Proceedings of the Conference on European Chapter of the Association for Computational Linguistics (EACL)*, pages 1989–2005, Dubrovnik, Croatia.

Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T. Hancock. 2011. Finding Deceptive Opinion Spam by Any Stretch of the Imagination. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 309—-319, Portland, OR, USA.

John Pavlopoulos, Jeffrey Sorensen, Lucas Dixon, Nithum Thain, and Ion Androutsopoulos. 2020. Toxicity Detection: Does Context Really Matter? In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4296–4305, Online.

Pew Research Center. 2014a. The Religious Landscape Study: Political ideology among Christians. https://www.pewresearch.org/religious-landscape-study/#political-ideology. Accessed: 26-Aug-2024.

Pew Research Center. 2014b. The Religious Landscape Study: Views about environmental regulation among Christians. https://www.pewresearch.org/religious-landscape-study/#views-about-environmental-regulation. Accessed: 26-Aug-2024.

Pew Research Center. 2023. How Americans View Future Harms From Climate Change in Their Community and Around the U.S. https://www.pewresearch.org/science/2023/10/25/how-americans-view-future-harms-from-climate-change-in-their-community-and-around-the-u-s/ps_2023-10-25_climate-change-harms_00-02-png/. Accessed: 26-Aug-2024.

Flor Miriam Plaza-del-arco, Debora Nozza, and Dirk Hovy. 2023. Respectful or Toxic? Using Zero-Shot Learning with Language Models to Detect Hate Speech. In *Proceedings of the Workshop on Online Abuase and Harms (WOAH)*, pages 60–68, Toronto, Canada.

Martin Potthast, Benno Stein, Alberto Barrón-Cedeño, and Paolo Rosso. 2010. An Evaluation Framework for Plagiarism Detection. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, pages 997–1005, Beijing, China.

Ines Reinig, Maria Becker, Ines Rehbein, and Simone Ponzetto. 2024. A Survey on Modelling Morality for Text Analysis. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 4136–4155, Bangkok, Thailand.

Hannah Ritchie. 2024. More people care about climate change than you think. *Our World in Data*. https://archive.ourworldindata.org/20251209-133038/climate-change-support.html.

Hernando Rojas, Dhavan V. Shah, and Ronald J. Faber. 1996. For the Good of Others: Censorship and the Third-Person Effect. *International Journal of Public Opinion Research*, 8(2):163–186.

Milton Rokeach. 1973. *The Nature of Human Values*. Free Press.

Paul Rottger, Bertie Vidgen, Dirk Hovy, and Janet Pierrehumbert. 2022. Two Contrasting Data Annotation Paradigms for Subjective NLP Tasks. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL (HLT/NAACL)*, pages 175–190, Seattle, WA, USA.

Paul Röttger, Bertram Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts, and Janet B. Pierrehumbert. 2021. HateCheck: Functional Tests for Hate Speech Detection Models. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 41–58, Online.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 29971–30004, Hawaii, HI, USA.

Maarten Sap, Swapna Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2022. Annotators with Attitudes: How Annotator Beliefs And Identities Bias Toxic Language Detection. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL (HLT/NAACL)*, pages 5884–5906.

Shalom H. Schwartz. 1992. Universals in the Content and Structure of Values: Theoretical Advances and Empirical Tests in 20 Countries. *Advances in Experimental Social Psychology*, 25:1–65.

Qinlan Shen and Carolyn Rose. 2021. What Sounds "Right" to Me? Experiential Factors in the Perception of Political Ideology". In *Proceedings of the Conference on European Chapter of the Association for Computational Linguistics (EACL)*, pages 1762–1771, Online.

Mark Spranca, Elisa Minsk, and Jonathan Baron. 1991. Omission and commission in judgment and choice. *Journal of Experimental Social Psychology*, 27(1):76–105.

Judith Jarvis Thomson. 1976. Killing, Letting Die, and the Trolley Problem. *The Monist*, 59(2):204–217.

Hugo Touvron, Louis Martin, Kevin R. Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, D. Bikel, Lukas Blecher, Cristian Cantón Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, A. Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel M. Kloumann, A. Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, R. Subramanian, Xia Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zhengxu Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models.

Betty van Aken, Julian Risch, Ralf Krestel, and Alexander Löser. 2018. Challenges for Toxic Comment Classification: An In-Depth Error Analysis. In *Proceedings of the Workshop on Abusive Language Online (ALW)*, pages 33–42, Brussels, Belgium.

Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. 2021. Learning from the Worst: Dynamically Generated Datasets to Improve Online Hate Detection. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1667–1682, Online.

Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding Abuse: A Typology of Abusive Language Detection Subtasks. In *Proceedings of the ACL-Workshop on Abusive Language Online*, pages 78–84, Vancouver, BC, Canada.

Zeerak Waseem and Dirk Hovy. 2016. Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL – Student Research Workshop*, pages 88–93, San Diego, CA, USA.

Tharindu Weerasooriya, Sujan Dutta, Tharindu Ranasinghe, Marcos Zampieri, Christopher Homan, and Ashiqur KhudaBukhsh. 2023. Vicarious Offense and Noise Audit of Offensive Speech Classifiers: Unifying Human and Machine Disagreement on What is Offensive. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 11648–11668, Singapore.

Michael Wiegand, Elisabeth Eder, and Josef Ruppenhofer. 2022. Identifying Implicitly Abusive Remarks about Identity Groups using a Linguistically Informed Approach. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL (HLT/NAACL)*, pages 5600–5612, Seattle, WA, USA.

Michael Wiegand, Maja Geulig, and Josef Ruppenhofer. 2021a. Implicitly Abusive Comparisons – A New Dataset and Linguistic Analysis. In *Proceedings of the Conference on European Chapter of the Association for Computational Linguistics (EACL)*, pages 358–368, Online.

Michael Wiegand, Josef Ruppenhofer, and Elisabeth Eder. 2021b. Implicitly Abusive Language – What does it actually look like and why are we not getting there? In *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL (HLT/NAACL)*, pages 576–587, Online.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. Qwen3 Technical Report. *Preprint*, arXiv:2505.09388.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Predicting the Type and Target of Offensive Posts in Social Media. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL (HLT/NAACL)*, pages 1415–1420, Minneapolis, MN, USA.

Jinguang Zhang. 2023. Perceived offensiveness to the self, not that to others, is a robust positive predictor of support of censoring sexual, alcoholic, and violent media content. *Frontiers in Psychology*, 14.

Haoti Zhong, Hao Li, Anna Cinzia Squicciarini, Sarah Michele Rajtmajer, Christopher Griffin, David J. Miller, and Cornelia Caragea. 2016. Content-Driven Detection of Cyberbullying on the Instagram Social Network. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pages 3952–3958, New York City, NY, USA.

Deyu Zhou, Jianan Wang, Linhai Zhang, and Yulan He. 2021. Implicit Sentiment Analysis with Event-Centered Text Representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6884–6893, Online and Punta Cana, Dominican Republic.

Xuhui Zhou, Hao Zhu, Akhila Yerukola, Thomas Davidson, Jena D. Hwang, Swabha Swayamdipta, and Maarten Sap. 2023. COBRA Frames: Contextual Reasoning about Effects and Harms of Offensive Statements. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6294–6315, Toronto, Canada.

Appendix Overview

This appendix provides more detailed information regarding certain aspects of our research for which there was not sufficient space in the main paper.

A Hyperparameters of Statistical Models

For all statistical models we used in this research we **refrained from heavy tuning of hyperparameters**. This is due to the fact that several experiments were evaluated in a cross-dataset setting, i.e. the training and test data originated from different datasets. As a consequence, tuning hyperparameters would only be possible by using some development data from the source domain. This, however, would mean that the resulting models would be tuned for the wrong domain. By running the tools with frequently used (default) settings of hyperparameters, we hope to produce models that are overall more robust across different domains (i.e. different datasets) than models fine-tuned on the wrong domain. Thus, we follow the strategy that was proposed for the large-scale cross-dataset evaluation reported in Wiegand et al. (2022).

A.1 Computing Infrastructure and Running Time

Our experiments were carried out on two servers:

- server 1: Lenovo ThinkSystem SR665; 1TB RAM; 2x32 Core AMD CPU that is equipped with one GPU (NVIDIA RTX A40, 48GB RAM)
- server 2: Quanton CS-221G-TRAN10-G12; 256GB RAM; 1 Intel Xeon Silver 4310 that is equipped with two GPUs (both: NVIDIA RTX A40, 48GB RAM)

We estimate a total computational budget of 80 GPU hours.

A.2 PerspectiveAPI

In our evaluation, we also included *PerspectiveAPI*[14] as one baseline. This tool runs on unrestricted text and, from the publicly available classifiers, it is currently considered the state of the art for the general detection of offensive language (Röttger et al., 2021). The tool predicts several subtypes of offensive language. They are:

- toxicity
- severe_toxicity
- insult
- identity_attack
- threat
- sexually_explicit
- profanity

For each of these categories, we examine how well it correlates with each label assignment of

[14] www.perspectiveapi.com

our 4 value systems. We expected that the different value systems might have different Perspective-categories they correlate most strongly with. However, in our preliminary experiments we found for all 4 value systems, the category *insult* produced best classification performance. This meant that we used this category in our evaluation (i.e. Tables 7 & 9).

A.3 Logistic Regression

For logistic regression, we used the implementation within **LIBLINEAR** (Fan et al., 2008) **with L1 regularization**. The advantage of logistic regression is that it is a robust classifier which does **not require any hyperparameter tuning**.

A.4 DeBERTa

For classification, we fine-tuned DeBERTa (more specifically `deberta-large`) using the implementation for text classification within the FLAIR framework (version 12) (Akbik et al., 2019). This language model belongs to the more recent models of larger size (11.5 billion parameters). It has been reported to achieve high classification performance in different NLP tasks.

In order **not to overfit the model**, we chose the hyperparameter settings from Wiegand et al. (2022):

- learning rate=3e-5
- mini batch size=16
- maximal epochs=5

That work addressed similar data as our work (i.e. implicitly offensive language detection). Since that work utilized RoBERTa (Liu et al., 2019) as its transformer model, we hope that those hyperparameters are not biased towards DeBERTa.

A.5 GPT-4/GPT-5

From GPT-models, we used GPT-4 (`gpt-4.0`) and GPT-5 (`gpt-5`).

For GPT-4, we mainly used the default settings of the hyperparameters:

- temperature=1
- top_p=1
- frequency_penalty=0
- presence_penalty=0

The only parameter for which we chose a setting different to the default settings in all of our experiments is the maximum number of tokens (*max_tokens*). Our aim was to ensure the completion was sufficiently long to solve the two tasks of our prompt (4). In our exploratory experiments, we determined that the following parameter setting achieved that goal:

- max_tokens=60

Please note that we only conducted a few experiments varying the parameter values at coarse-grained intervals. We basically chose a value in which the response covered complete sentences to answer the prompt. If we had chosen a smaller number of tokens, the response would have often resulted in incomplete sentences. We did not explicitly optimize this parameter on classification performance since this would have overfit our settings to our dataset. Leaving this parameter at its default setting (i.e. *max_tokens*=256) was no option either, as such responses would have been unnecessarily verbose, not to mention the unnecessary financial costs.

For GPT-5, we used the default settings; we did not specify a maximum number of tokens, as this is no longer available as a hyperparameter.

A.6 Open-Weight Models

All open-weight models were accessed through the API provided by Replicate.[15] We primarily used the platform's **default settings**. Similar to our approach with GPT-4, we adjusted only the *max_tokens* hyperparameter, i.e. we set it to 60 tokens. This adjustment was made for the same reasons as discussed above for GPT-4. For each language model type, we selected the most advanced variant available on the platform.

In the following, we specify the individual settings:

A.6.1 DeepSeek V3

From DeepSeek language models, we utilized `deepseek-ai/deepseek-v3`. As hyperparameters, we chose:

- temperature=0.6
- top_p=1
- presence_penalty=0
- frequency_penalty=0

A.6.2 LLaMA 4

From LLaMA language models, we utilized `meta/llama-4-maverick-instruct`. As hyperparameters, we chose:

- temperature=0.6
- top_p=1
- presence_penalty=0
- frequency_penalty=0

[15] https://replicate.com

A.6.3 Qwen3

From Qwen language models, we utilized `qwen/qwen3-235b-a22b-instruct-2507`. As hyperparameters, we chose:

- temperature=0.1
- top_p=1
- presence_penalty=0
- frequency_penalty=0

B Topics Occurring in Our Dataset

Table 10 lists all topics contained in our dataset. It also provides the percentage of offensiveness per value system and group (i.e. favouring group 'T+' and disfavouring group 'T−'). The percentage represents the proportion of sentences within a group that were perceived as offensive for a particular topic. For example, 80% in 'T−' under *Liberal* indicates that 80% of the disfavouring sentences are perceived as offensive by crowdworkers identifying as liberal.

The topics of our dataset were chosen manually by the co-authors. However, they also sought feedback from members of the 4 value systems via crowdsourcing. The topics should encompass a broad spectrum, with each one being **controversial to varying degrees**, such as *eating meat* versus *caning children*. Assuming the range of possible topics adheres to a power-law distribution, we also aimed to maintain a **balance between frequently discussed topics** (e.g. *consuming alcohol*) **and less frequently discussed topics** (e.g. *speaking in a dialect in front of people from a different region*) to accommodate the long tail.

Though we tried to achieve a balanced distribution of topics, we did not exhaustively add obvious counterexamples. For instance, two of our topics *products by Nike* and *shopping at Walmart* were included because Nike and Walmart are heavily criticized for their labor practices and environmental footprint. We refrained from including companies that are free from these criticisms, as we had no indication from our initial feedback from members of the 4 value systems that these topics would result in sentences perceived as offensive by members of any value system. However, this design choice does not imply that our set of sentences for each topic lacks important variants or is politically, ideologically or conceptually restricted. By having two groups, namely, those favouring and those disfavouring (c.f. Table 1), a balanced consideration is ensured for proponents and opponents of a particular topic, such as *guns*, where gun rights advocates are represented by the sentences of the favouring group, whereas gun control advocates are represented by sentences of the disfavouring group.

topic	Christian		Conservative		Ecological		Liberal	
	T+	T−	T+	T−	T+	T−	T+	T−
100oz steak challenge (eating competition for 1 person)	0.0	0.0	0.0	16.7	50.0	0.0	50.0	16.7
10-pint challenge (beer drinking competition)	100.0	16.7	100.0	0.0	33.3	0.0	0.0	0.0
abortion	100.0	0.0	100.0	0.0	33.3	50.0	0.0	66.7
animal circuses	0.0	0.0	0.0	0.0	66.7	0.0	50.0	0.0
animal hunting	0.0	0.0	0.0	33.3	100.0	16.7	83.3	0.0
animal testing	66.7	0.0	83.3	0.0	100.0	0.0	100.0	0.0
antidepressants	16.7	0.0	0.0	16.7	16.7	33.3	0.0	66.7
antipsychotics	33.3	0.0	33.3	0.0	16.7	33.3	16.7	66.7
asking about someone's heritage	0.0	0.0	0.0	16.7	0.0	16.7	50.0	16.7
asking for someone's gender	16.7	16.7	33.3	33.3	0.0	16.7	50.0	0.0
asking personal questions	0.0	0.0	50.0	33.3	16.7	0.0	16.7	0.0
assisted suicide	100.0	0.0	100.0	0.0	83.3	33.3	25.0	50.0
BASE jumping	0.0	0.0	16.7	0.0	0.0	0.0	0.0	16.7
blackfacing	66.7	0.0	100.0	0.0	100.0	0.0	100.0	0.0
blood transfusions	0.0	50.0	0.0	33.3	0.0	33.3	0.0	66.7
Botox injections	100.0	16.7	33.3	0.0	50.0	33.3	16.7	0.0
breastfeeding in public	0.0	16.7	50.0	66.7	0.0	50.0	0.0	66.7
Breitbart News	0.0	0.0	0.0	33.3	16.7	0.0	83.3	0.0
bubblegum pop	16.7	0.0	0.0	0.0	0.0	16.7	0.0	33.3
bullfighting	100.0	0.0	16.7	0.0	100.0	0.0	83.3	0.0
bungee jumping	0.0	0.0	0.0	0.0	0.0	16.7	0.0	0.0
buying best brands	0.0	50.0	0.0	16.7	0.0	33.3	0.0	33.3
caning children	100.0	0.0	100.0	0.0	100.0	0.0	100.0	0.0
capital punishment	16.7	0.0	0.0	16.7	83.3	0.0	100.0	0.0
car/motor races	0.0	16.7	0.0	16.7	0.0	0.0	0.0	33.3
casino gambling	100.0	0.0	16.7	0.0	16.7	16.7	16.7	0.0
charity gambling	100.0	0.0	16.7	16.7	33.3	0.0	0.0	16.7
chemical castration	83.3	0.0	83.3	0.0	50.0	0.0	100.0	0.0
circumcision	0.0	60.0	0.0	40.0	40.0	20.0	0.0	40.0
climate change	0.0	16.7	0.0	0.0	16.7	83.3	16.7	83.3
CNN	0.0	33.3	50.0	0.0	0.0	16.7	0.0	33.3
cockfighting	100.0	0.0	100.0	0.0	100.0	0.0	100.0	0.0
compulsory vaccination	33.3	0.0	83.3	0.0	0.0	16.7	0.0	66.7
conserving energy	0.0	33.3	0.0	50.0	0.0	66.7	0.0	100.0
consuming alcohol	100.0	0.0	50.0	33.3	50.0	0.0	16.7	16.7
consuming alcohol in public	100.0	0.0	83.3	0.0	50.0	0.0	33.3	33.3
consuming marijuana for medical purposes	50.0	16.7	16.7	16.7	0.0	33.3	0.0	0.0
consuming marijuana for recreational purposes	83.3	0.0	83.3	0.0	0.0	16.7	0.0	33.3
consuming marijuana for stress relief	83.3	16.7	66.7	0.0	0.0	50.0	0.0	50.0
consumption of marijuana for spiritual purposes	100.0	0.0	50.0	0.0	0.0	50.0	0.0	50.0
contraception	66.7	16.7	16.7	66.7	0.0	66.7	0.0	66.7
conversion therapy	33.3	16.7	33.3	33.3	66.7	16.7	83.3	16.7
corporal punishment on adults	50.0	0.0	66.7	16.7	100.0	0.0	100.0	0.0
corporal punishment on children	83.3	0.0	83.3	0.0	100.0	0.0	100.0	0.0
creationism	0.0	33.3	16.7	60.0	16.7	33.3	83.3	16.7
critical race theory	33.3	16.7	100.0	0.0	0.0	83.3	0.0	100.0
crying in public	0.0	16.7	16.7	33.3	16.7	66.7	16.7	66.7
dating people outside one's faith	0.0	0.0	50.0	33.3	0.0	33.3	0.0	66.7
death metal	83.3	0.0	33.3	0.0	0.0	33.3	0.0	33.3
dialysis	0.0	33.3	0.0	50.0	0.0	66.7	0.0	66.7
direct one's life according to the horoscope	66.7	0.0	50.0	0.0	0.0	16.7	33.3	0.0
displaying acts of physical intimacy in public	100.0	16.7	100.0	0.0	33.3	33.3	33.3	50.0
displaying emotions in public	0.0	33.3	0.0	50.0	0.0	50.0	0.0	66.7
displaying one's wealth	0.0	0.0	33.3	0.0	50.0	16.7	0.0	16.7
display of blood	66.7	0.0	60.0	0.0	66.7	0.0	66.7	33.3
diverse sexual practices	100.0	16.7	100.0	33.3	16.7	33.3	66.7	50.0
divorce	83.3	16.7	66.7	16.7	33.3	16.7	16.7	33.3
dollar menu items in fast-food restaurants	0.0	0.0	0.0	33.3	0.0	33.3	0.0	16.7
dressing casually in public	0.0	0.0	0.0	33.3	0.0	33.3	0.0	50.0
dressing dogs in clothes	0.0	16.7	0.0	0.0	0.0	16.7	0.0	0.0
dressing lavishly	0.0	0.0	0.0	16.7	0.0	16.7	0.0	16.7
drinking competitions	100.0	0.0	100.0	0.0	33.3	0.0	16.7	0.0
driving big luxury cars	0.0	0.0	0.0	0.0	66.7	0.0	0.0	0.0
dumpster diving	33.3	16.7	50.0	0.0	16.7	33.3	0.0	16.7
early childhood enrichment	0.0	66.7	0.0	50.0	0.0	66.7	0.0	83.3
ear stretching	60.0	0.0	40.0	0.0	0.0	33.3	0.0	33.3

Continued on next page

topic	Christian		Conservative		Ecological		Liberal	
	T+	T−	T+	T−	T+	T−	T+	T−
eating at Chick-fil-A	0.0	50.0	0.0	50.0	16.7	0.0	50.0	0.0
eating competitions	0.0	0.0	0.0	0.0	50.0	0.0	0.0	16.7
eating foie gras	16.7	20.0	16.7	0.0	16.7	0.0	66.7	0.0
eating horse meat	50.0	0.0	83.3	0.0	100.0	0.0	83.3	0.0
eating meat	0.0	0.0	0.0	16.7	50.0	0.0	0.0	16.7
eating pork	0.0	33.3	0.0	16.7	33.3	16.7	16.7	33.3
eating raw meat	66.7	0.0	50.0	0.0	66.7	0.0	33.3	33.3
equal income	0.0	83.3	0.0	83.3	0.0	100.0	0.0	100.0
euthanasia	66.7	0.0	100.0	0.0	50.0	0.0	16.7	66.7
evolution	66.7	0.0	33.3	0.0	0.0	83.3	0.0	100.0
face lifting	33.3	0.0	16.7	0.0	66.7	0.0	33.3	0.0
factory farming	0.0	16.7	0.0	0.0	50.0	0.0	66.7	0.0
fast fashion	0.0	0.0	0.0	0.0	50.0	0.0	16.7	16.7
fast fashion that gets worn out after just a very few washes	0.0	16.7	0.0	0.0	100.0	0.0	66.7	0.0
fast food	0.0	16.7	0.0	16.7	0.0	0.0	16.7	16.7
fine dining	0.0	0.0	0.0	0.0	0.0	16.7	0.0	16.7
fossil fuels	0.0	0.0	33.3	33.3	66.7	0.0	83.3	0.0
Fox News	0.0	0.0	0.0	33.3	33.3	16.7	100.0	0.0
gambling	100.0	0.0	83.3	16.7	16.7	0.0	16.7	0.0
gender affirming treatment	100.0	0.0	100.0	0.0	0.0	66.7	0.0	66.7
gender-specific bathrooms	0.0	50.0	0.0	83.3	16.7	50.0	33.3	33.3
gender-specific locker rooms	0.0	33.3	16.7	100.0	16.7	33.3	33.3	16.7
genetically modified food	16.7	0.0	16.7	16.7	66.7	0.0	33.3	16.7
genetic engineering	83.3	0.0	16.7	0.0	0.0	33.3	0.0	50.0
getting drunk	100.0	0.0	100.0	0.0	33.3	0.0	50.0	33.3
going by car for short distances that could also be easily managed on foot	0.0	16.7	33.3	16.7	83.3	0.0	83.3	16.7
guns	33.3	33.3	0.0	50.0	50.0	0.0	66.7	0.0
having a private chef	0.0	0.0	0.0	33.3	0.0	16.7	0.0	33.3
having children	0.0	80.0	0.0	80.0	0.0	20.0	0.0	60.0
having dreadlocks in spite of having no relation to the cultures in which they are traditionally worn	33.3	16.7	83.3	33.3	83.3	16.7	100.0	16.7
having many children	0.0	40.0	0.0	80.0	16.7	40.0	0.0	40.0
having personal staff	0.0	33.3	0.0	16.7	0.0	16.7	0.0	16.7
homeopathy	0.0	0.0	0.0	0.0	0.0	33.3	0.0	33.3
hormone replacement therapy for transgender individuals	100.0	0.0	100.0	16.7	0.0	33.3	0.0	66.7
hormone replacement therapy for transgender individuals before reaching the age of 18	100.0	16.7	100.0	16.7	33.3	50.0	0.0	60.0
horror movies featuring demonic activities	66.7	0.0	100.0	0.0	0.0	0.0	16.7	33.3
horse races	83.3	0.0	50.0	0.0	50.0	0.0	0.0	0.0
intelligent design	0.0	100.0	0.0	16.7	0.0	0.0	66.7	16.7
invasive surveillance technology	16.7	0.0	83.3	0.0	100.0	0.0	66.7	0.0
in-vitro fertilization	0.0	16.7	0.0	0.0	16.7	33.3	16.7	66.7
jokes on death	100.0	0.0	66.7	0.0	50.0	16.7	100.0	16.7
jokes on illness	100.0	0.0	100.0	0.0	100.0	0.0	100.0	0.0
jokes on religion	100.0	0.0	100.0	0.0	100.0	0.0	83.3	0.0
jokes on sex	100.0	0.0	83.3	0.0	16.7	16.7	16.7	33.3
knowing the words of the national anthem	0.0	83.3	0.0	83.3	0.0	50.0	0.0	50.0
letting dogs sleep in their owner's bed	0.0	0.0	16.7	0.0	0.0	33.3	0.0	16.7
marry one's romantic partner	0.0	66.7	0.0	83.3	0.0	33.3	0.0	50.0
martial arts	0.0	16.7	0.0	33.3	0.0	16.7	0.0	33.3
meditation	0.0	50.0	0.0	0.0	0.0	0.0	0.0	66.7
national pride	0.0	83.3	16.7	66.7	0.0	33.3	0.0	50.0
neutering cats	0.0	50.0	0.0	33.3	0.0	83.3	0.0	66.7
New York Times	0.0	0.0	33.3	0.0	0.0	16.7	0.0	50.0
non-restorative breast implants	100.0	0.0	50.0	33.3	16.7	16.7	50.0	33.3
non-restorative cosmetic surgery	80.0	16.7	0.0	16.7	20.0	33.3	20.0	33.3
non-restorative cosmetic surgery before reaching the age of 18	100.0	0.0	100.0	0.0	83.3	20.0	16.7	0.0
nudism	100.0	0.0	100.0	0.0	50.0	0.0	16.7	33.3
nudism in public	100.0	16.7	100.0	16.7	83.3	0.0	100.0	16.7
offshore banking	0.0	0.0	33.3	0.0	33.3	0.0	66.7	0.0
one-night stands	100.0	0.0	100.0	0.0	33.3	0.0	16.7	16.7
open relationships	100.0	0.0	100.0	0.0	33.3	16.7	0.0	66.7

Continued on next page

topic	Christian		Conservative		Ecological		Liberal	
	T+	T−	T+	T−	T+	T−	T+	T−
organizing birthday parties for one's dog that includes presents and inviting other people's dogs	0.0	0.0	0.0	16.7	0.0	16.7	0.0	0.0
organizing social time for one's dog (e.g. by inviting other dogs to play with them)	0.0	33.3	0.0	16.7	0.0	16.7	0.0	16.7
organ transplantation	0.0	16.7	16.7	33.3	16.7	33.3	0.0	66.7
overeating	100.0	0.0	66.7	0.0	83.3	16.7	66.7	0.0
paying high annual salaries of up to 40 million dollars (after taxes) to excellent football players	16.7	0.0	33.3	33.3	16.7	0.0	50.0	0.0
paying top managers of big international companies an annual salary of 60 million dollars (after taxes)	0.0	0.0	0.0	0.0	50.0	16.7	100.0	0.0
piercings	33.3	0.0	0.0	16.7	0.0	33.3	0.0	50.0
playing poker	66.7	0.0	16.7	0.0	0.0	0.0	0.0	16.7
playing the lottery	100.0	0.0	16.7	16.7	0.0	0.0	0.0	0.0
pledge to allegiance	0.0	83.3	0.0	100.0	0.0	50.0	0.0	33.3
polyamory	100.0	16.7	100.0	0.0	33.3	50.0	33.3	50.0
polygamy	100.0	0.0	100.0	16.7	33.3	0.0	66.7	0.0
pornography	100.0	0.0	100.0	0.0	50.0	0.0	50.0	16.7
praying	0.0	100.0	0.0	100.0	0.0	66.7	0.0	50.0
pre-marital sex	100.0	0.0	100.0	0.0	33.3	33.3	33.3	66.7
private prisons	0.0	0.0	0.0	16.7	50.0	0.0	100.0	0.0
products by Nike	0.0	33.3	0.0	16.7	0.0	0.0	0.0	16.7
prostitution	100.0	16.7	100.0	0.0	100.0	0.0	83.3	66.7
public transport	0.0	33.3	0.0	16.7	0.0	66.7	0.0	50.0
public transport subsidized by one's employer	0.0	0.0	0.0	50.0	0.0	16.7	0.0	33.3
pugs and French bulldogs	0.0	16.7	0.0	33.3	0.0	0.0	0.0	16.7
transgender people using locker rooms that align with their gender identity rather than their biological sex	100.0	50.0	100.0	50.0	0.0	66.7	0.0	66.7
read one's horoscope	50.0	0.0	66.7	0.0	0.0	16.7	0.0	16.7
reality TV	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
redefinition of the traditional family	33.3	0.0	100.0	16.7	0.0	50.0	0.0	83.3
religious elements in public affairs/buildings	0.0	100.0	0.0	100.0	33.3	16.7	100.0	16.7
respecting pronouns	16.7	16.7	50.0	16.7	0.0	83.3	0.0	100.0
same-sex marriage	100.0	20.0	83.3	33.3	0.0	66.7	0.0	66.7
self promotion	0.0	0.0	0.0	16.7	33.3	16.7	16.7	16.7
sex education	16.7	16.7	50.0	0.0	0.0	66.7	0.0	66.7
shopping at H&M	0.0	0.0	0.0	0.0	0.0	0.0	33.3	0.0
shopping at Walmart	0.0	16.7	0.0	16.7	0.0	0.0	0.0	16.7
skiing	0.0	0.0	0.0	16.7	0.0	0.0	0.0	16.7
smoking	100.0	0.0	83.3	0.0	50.0	0.0	16.7	0.0
speaking in a dialect in front of people who come from a different region	0.0	50.0	0.0	50.0	0.0	50.0	0.0	33.3
sperm donation	100.0	0.0	16.7	33.3	0.0	33.3	0.0	66.7
standing up for the national anthem	16.7	100.0	0.0	100.0	0.0	66.7	0.0	66.7
stem cell research	66.7	0.0	60.0	0.0	0.0	16.7	0.0	66.7
strip clubs	100.0	0.0	100.0	0.0	33.3	16.7	16.7	33.3
swinging	100.0	0.0	100.0	0.0	33.3	0.0	20.0	50.0
taking selfies	0.0	16.7	0.0	16.7	16.7	33.3	16.7	33.3
taking the Bible literally	0.0	50.0	0.0	100.0	33.3	0.0	50.0	0.0
talking about illnesses	33.3	0.0	0.0	16.7	0.0	16.7	0.0	16.7
talking about one's emotions	0.0	33.3	0.0	33.3	0.0	50.0	0.0	66.7
talking about sex	100.0	0.0	66.7	33.3	16.7	33.3	0.0	50.0
talking about toileting	83.3	16.7	83.3	16.7	16.7	16.7	33.3	0.0
tattoos	100.0	0.0	0.0	20.0	0.0	60.0	0.0	60.0
teaching children about LGBTQ	100.0	0.0	100.0	16.7	0.0	83.3	0.0	100.0
teaching children Chinese at the age of 3 (as part of early childhood enrichment)	0.0	16.7	0.0	16.7	0.0	50.0	0.0	83.3
the concept of the 2 (traditional) genders	0.0	66.7	0.0	83.3	33.3	16.7	50.0	33.3
the local police	0.0	100.0	0.0	100.0	0.0	50.0	0.0	66.7
throwing away food	100.0	0.0	83.3	0.0	100.0	0.0	83.3	0.0
transgender people using bathrooms that align with their gender identity rather than their biological sex	100.0	20.0	100.0	33.3	0.0	83.3	0.0	80.0
trans story hour	83.3	0.0	100.0	0.0	0.0	50.0	0.0	83.3
travelling by airplane	0.0	0.0	0.0	0.0	16.7	0.0	0.0	16.7

Continued on next page

topic	Christian		Conservative		Ecological		Liberal	
	T+	T−	T+	T−	T+	T−	T+	T−
travelling by airplane if there is an alternative (e.g. bus or train)	0.0	16.7	0.0	16.7	100.0	0.0	100.0	0.0
travelling by car	0.0	33.3	0.0	16.7	16.7	16.7	0.0	16.7
travesty	100.0	0.0	100.0	0.0	16.7	66.7	0.0	66.7
trophy hunting	16.7	16.7	50.0	16.7	100.0	0.0	100.0	0.0
unconditional basic income	0.0	0.0	50.0	16.7	0.0	0.0	0.0	50.0
urinating in public	100.0	0.0	100.0	0.0	100.0	0.0	100.0	0.0
using air-conditioning	0.0	66.7	0.0	0.0	0.0	0.0	16.7	16.7
using paper plates and cups	0.0	16.7	16.7	33.3	50.0	0.0	66.7	0.0
using single-use plastic bags	0.0	0.0	0.0	0.0	100.0	0.0	83.3	0.0
using the ocean as disposal site for garbage	100.0	0.0	100.0	0.0	83.3	0.0	100.0	0.0
veganism	0.0	16.7	0.0	16.7	0.0	50.0	0.0	50.0
violent movies	100.0	0.0	50.0	0.0	0.0	16.7	50.0	16.7
violent video games	100.0	0.0	100.0	0.0	16.7	0.0	33.3	16.7
visiting a sauna	0.0	50.0	0.0	0.0	0.0	0.0	0.0	33.3
walking one's dog in the dog stroller	0.0	0.0	0.0	16.7	0.0	16.7	0.0	16.7
waterboarding for interrogation	16.7	0.0	83.3	33.3	100.0	16.7	100.0	0.0
wearing a mariachi costume for a fancy-dress party	16.7	0.0	16.7	16.7	50.0	0.0	83.3	0.0
wearing a Native American costume for a fancy-dress party	66.7	0.0	33.3	33.3	100.0	0.0	100.0	0.0
wearing clothes made of synthetic fabrics	0.0	0.0	0.0	16.7	33.3	0.0	0.0	16.7
wearing flip-flops	0.0	33.3	0.0	0.0	0.0	16.7	0.0	16.7
wearing flip-flops in the office	50.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
wearing fur	0.0	0.0	0.0	33.3	100.0	0.0	83.3	0.0
wearing religious symbols as fashion accessories	83.3	0.0	16.7	50.0	33.3	50.0	50.0	33.3
wearing revealing clothes	100.0	0.0	83.3	0.0	33.3	50.0	50.0	66.7
wearing silk	0.0	16.7	0.0	33.3	16.7	0.0	0.0	0.0
wearing sweatpants	0.0	50.0	0.0	50.0	0.0	0.0	0.0	33.3
wearing sweatpants in the office	50.0	0.0	16.7	0.0	0.0	0.0	16.7	0.0
working on Sunday	83.3	0.0	66.7	0.0	0.0	0.0	16.7	33.3
worshiping celebrities	100.0	0.0	100.0	0.0	16.7	0.0	33.3	0.0
wrestling	0.0	16.7	0.0	16.7	0.0	16.7	0.0	16.7
Yoga	0.0	50.0	0.0	16.7	0.0	16.7	0.0	50.0

Table 10: Topics of the dataset with percentages of offensiveness per value system and group; 'T+' signifies the offensiveness of *favouring* sentences, 'T−' signifies the offensiveness of *disfavouring* sentences (e.g. 80% in 'T−' under *Liberal* indicates that 80% of the disfavouring sentences are perceived as offensive by crowdworkers identifying as liberal); darker colors denote higher offensiveness.

value systems		annotations	
		offensiveness	sentiment
Christian	Conservative	**0.587**	**0.776**
Christian	Ecological	0.205	0.410
Christian	Liberal	0.140	0.259
Conservative	Ecological	0.249	0.402
Conservative	Liberal	0.193	0.275
Ecological	Liberal	**0.572**	**0.824**

Table 11: Correlation (Spearman's ρ) between value systems.

C Correlation between Value Systems

In Table 11, we present the correlation scores (based on Spearman's ρ) between pairs of value systems for two annotations: offensive language detection (sentence-level annotation, §2.1) and sentiment classification (topic-level annotation, §2.2). For both annotations, the correlation is notably stronger between Christians and conservative people, as well as between ecologically-conscious and liberal people, compared to any other pair of value systems.

D Providing Statistics for Distinctiveness of the Individual Value Systems

In this section, we want to provide evidence for the distinctiveness between the concepts that underlie our 4 chosen value systems.

The value systems of conservatives and liberals represent two opposing ideologies, and therefore, we expect no overlap between them.

While Christians are typically considered conservative, according to Pew Research Center (2014a), a considerable number of Christians also identify as liberals. Thus, Christians do not constitute a proper subset of conservatives.

Similarly, while liberal and ecologically-conscious individuals may share some beliefs, ecologically-conscious people are not a proper subset of liberals. According to Pew Research Center (2023), both liberals and conservatives have a majority who recognize the need for action (the sum of people who are prepared to make either *minor sacrifice* or *major sacrifice*) when it comes to addressing climate change. This suggests that there is also a significant ecological consciousness among conservatives.

Theoretically, there could be some overlap between ecologically-conscious people and Christians. However, according to Pew Research Center (2014b), there is a nearly equal split among Christians between those who support environmental regulation and those who do not. From this, we can conclude that a significant proportion of Christians are not particularly ecologically conscious.

E How Crowdworkers were Recruited

All crowdworkers, irrespective of the particular group of value system they belong to, had to be **native speakers of English** and be **located within the USA**. The latter was necessary since the demographic categories to establish members of particular value systems were based on the US system. Such categories may not be applicable to other countries. We focused on crowdworkers from the USA since, on Prolific, there is a sufficiently large pool of crowdworkers for all the 4 value systems from this country that we want to examine in our research.

In order to be eligible for any of our crowdsourcing surveys, each crowdworker also had to pass a **pre-screening** test in which we asked them **how easily they are generally offended**. The crowdworkers had to rate themselves on a Likert-scale where 1 corresponds to *not at all* and 5 to *very easily*. We only admitted any crowdworker who rated themselves as 3 or higher. This pre-screening test was necessary since in our exploratory surveys in which we hired crowdworkers without previously passing that test, we often obtained responses by crowdworkers in which all or virtually all sentences were rated as *not offensive*. Some of these crowdworkers also confirmed in their voluntary feedback that they considered themselves to be resistant against any type offensive language. No matter whether these surveys have been answered truthfully or not by these crowdworkers, such responses are of little use to our current research.

In the following, we detail the specific recruitment process for the crowdworkers representing the 4 individual value systems:

Christians. We are interested in *believing Christians*, i.e. crowdworkers who do not just identify with the Christian tradition due to upbringing or societal context but who actively believe or practice the faith. We approximated this by asking for Christian denominations that especially emphasize active participation. They are:

- The Church of Jesus Christ of Latter-Day Saints,
- Pentecostals/Apostolic, *and*
- Jehovah's Witnesses

There is no direct method available on Prolific to verify whether a crowdworker is a believing Christian. Our decision to use specific Christian denominations as a proxy may seem somewhat strict, as there are undoubtedly members of other Christian denominations who could also be considered believing Christians. However, including these other groups would have introduced too many *nominal* Christians, which could have significantly distorted our results. For certain Christian denominations, such as The Church of Jesus Christ of Latter-day Saints, it is statistically proven that they are more rigorous in practicing their faith. For example, weekly service attendance is approximately twice as high as that of Catholics or mainline Protestants (Gallup, 2024). We included Pentecostals/Apostolics and Jehovah's Witnesses because they closely resemble The Church of Jesus Christ of Latter-day Saints in terms of religious commitment, distinctiveness in beliefs and lifestyle.

Conservative People. We are focusing on individuals who align with modern conservative values. For this group, we recruited people who:

- identify themselves as politically conservative people,
- have a leaning towards the Republican Party (USA), *and*
- have voted for the Republican candidate in the last 2 previous US presidential elections.

Liberal People. We are focusing on individuals who align with modern liberal values. For this group, we recruited people who:

- identify themselves as politically liberal people,
- have a leaning towards the Democratic Party (USA), *and*
- have voted for the Democratic candidate in the last 2 previous US presidential elections.

Ecologically-Conscious People. For this group, we recruited people who:

- believe that climate change exists *and*
- rate their concern about environmental issues with a 5 on a Likert-scale (where 1 corresponds to *not at all concerned* and 5 to *very concerned*).

Some of the above criteria may appear repetitive or too restrictive. However, a minimalist approach in which we used as few criteria as possible was found not sufficiently reliable. We observed that several crowdworkers use demographic categories to describe themselves fairly flippantly or inconsistently. (This could also be confirmed by the voluntary feedback that many crowdworkers provided at the end of a survey in which they reported some of their key values which were irreconcilable with the actual demographic categories that they previously labelled themselves with.) Although the above choice of criteria restricts the set of possible candidate of crowdworkers to participate in our crowdsourcing surveys, that pool turned out to be much more reliable.

As we noted in the main paper in §1, value systems can sometimes overlap. However, we found that the overlap of crowdworkers between pairs of value systems was minimal (i.e. less than 5%). Given this minimal overlap and the fact that each crowdworker contributed to only a small fraction of the overall dataset, we consider the risk of this overlap distorting the resulting annotations to be negligible.

F Interannotator Agreement

In this section, we provide details on the agreement between annotators of both the sentence-level dataset (§F.1), where the task is offensive language detection, and the underlying topics (§F.2), where the task is sentiment classification.

F.1 Agreement on Offensiveness

Table 12 presents the agreement between two majority votes (measured by Cohen's κ) for a random sample of 200 sentences from each value system on the task of offensive language detection. Each majority vote was determined based on the annotations of 5 different crowdworkers who identified with the respective value system. Thus, for each value system, 10 crowdworkers annotated the same set of 200 sentences. The agreement scores range from 0.607 (for *Conservative*) to 0.783 (for *Ecological*). All these scores can be interpreted as indicating **substantial** agreement (McHugh, 2012).

F.2 Agreement on Sentiment Scores

Table 13 displays the Spearman rank-order correlation coefficient ρ of 2 averages over 5 separate crowdworkers identifying with the same value system on the task of sentiment scoring. The correlation was measured on the entire set of 212 topics of

value system	Christian	Conservative	Ecological	Liberal
Cohen's κ	0.607	0.644	0.783	0.663

Table 12: Interannotator agreement measured on a random sample of 200 sentences between 2 majority votes each derived from the judgments of 5 different crowdworkers.

value system	Christian	Conservative	Ecological	Liberal
Spearman's ρ	0.850	0.754	0.855	0.774

Table 13: Interannotator agreement measured on the entire set of topics (212 topics) between 2 averages of sentiment scores each derived from the judgments of 5 different crowdworkers.

the dataset. The scores range from 0.754 (*Conservative*) to 0.855 (*Ecological*). All scores represent a **strong** agreement.

G Detailed Annotation Instructions

Figure 4 displays the complete annotation instructions for the task of offensiveness detection as presented to the crowdworkers.

H Offensiveness within (Topic-Related) Sentiment Categories across Different Sentence Types

Table 14 shows the distribution of offensiveness within (topic-related) sentiment categories across the different sentence types. This table complements Table 5 in that the distribution of each individual value system is listed rather than the average over all groups.

I Distribution of Offensive Sentences across Different Sentiment Scores

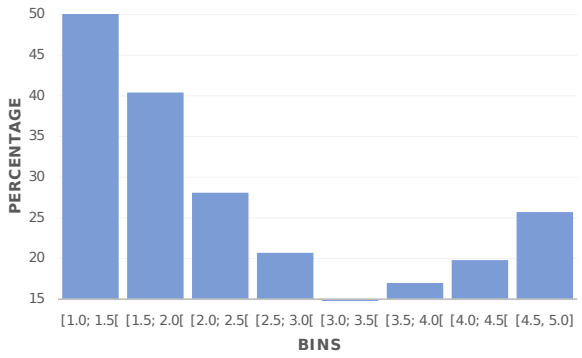
Figure 5 shows the distribution of offensive sentences within bins of a particular range of sentiment scores. This figure complements Figure 3 in that it presents the distribution for each value systems individually. Overall, the distribution of the individual values share notable similarities. The proportion of offensive language is notably stronger in bins representing very negative and very positive sentiment, where negative sentiment typically has an even higher proportion than positive sentiment.

> In the following, you are asked to annotate a set of utterances. **Imagine that these utterances were made by some other people about a good friend of yours. Your friend and you share similar beliefs, views and values.** Please also imagine that you coincidentally listen to what these people are saying about your friend and that they are not aware that you hear what they say about your friend. That said, these utterances may not be meant in a positive way. We ask you to decide for every individual utterance whether you think it is potentially **offensive or not** to `<value_system>`.

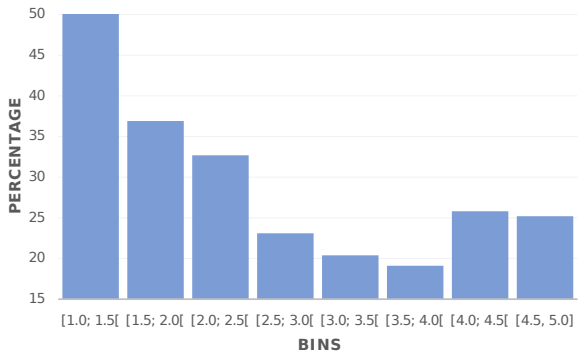
Figure 4: Annotation instructions for offensiveness detection.

		Christian			Conservative			Ecological			Liberal		
	sentence type	**pos.**	**neut.**	**neg.**	**pos.**	**neut.**	**neg.**	**pos.**	**neut.**	**neg.**	**pos.**	**neut.**	**neg.**
favouring	approval	1.3	13.0	**16.7**	1.9	8.3	**16.0**	5.3	8.4	**14.8**	1.5	5.6	**14.9**
	approval::intensive	2.6	12.4	**15.8**	3.2	12.5	**17.0**	3.2	13.3	**16.3**	1.1	10.7	**14.5**
	practice	2.6	10.7	**16.4**	2.6	9.4	**14.9**	0.5	8.0	**13.3**	1.1	4.6	**13.7**
	practice::intensive	4.6	11.9	**15.5**	5.8	12.1	**15.3**	4.2	13.3	**16.3**	3.7	12.7	**14.9**
	let_others_participate	2.0	11.9	**15.3**	3.2	8.7	**12.4**	1.1	5.3	**14.3**	0.7	4.6	**11.8**
	encourage_others_to_participate	2.0	16.4	**16.4**	5.8	12.1	**16.3**	3.2	14.2	**16.3**	3.0	12.2	**16.9**
disfavouring	disapproval	**17.0**	4.0	0.3	**12.9**	6.8	0.7	**16.9**	3.6	1.0	**18.2**	5.1	1.2
	disapproval::intensive	**7.8**	4.5	2.3	**9.0**	8.3	2.1	**13.8**	7.6	2.0	**17.4**	8.1	2.4
	no_practice	**7.2**	0.0	0.3	**5.8**	1.1	0.4	**4.2**	0.9	0.5	**3.7**	1.5	0.4
	no_practice::intensive	**10.5**	0.0	0.9	**8.4**	1.1	0.7	**4.8**	1.8	0.5	**3.0**	0.5	0.4
	discourage_others_from_participating	**20.9**	6.8	0.3	**19.4**	6.0	0.7	**19.1**	8.0	1.5	**21.1**	8.6	3.9
	prevent_others_from_participating	**21.6**	8.5	0.0	**21.9**	13.6	3.6	**28.6**	15.6	3.5	25.6	**25.9**	5.1

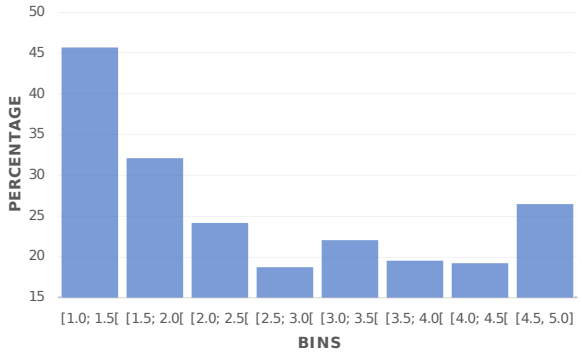
Table 14: Percentage of offensiveness within (topic-related) sentiment categories across different sentence types for each individual value system.



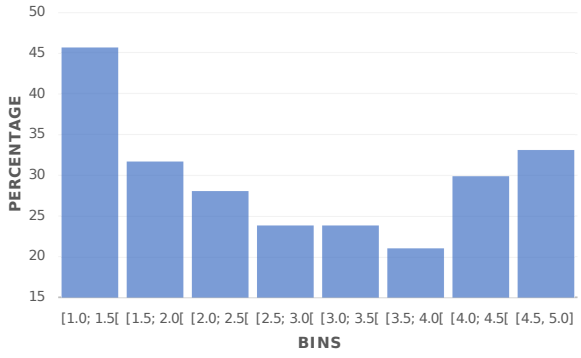
(a) Christians



(b) Conservative People



(c) Ecologically-Conscious People



(d) Liberal People

Figure 5: Percentage of offensive sentences within bins of a particular range of sentiment scores: distribution for each value system.