Do differences in values influence disagreements in online discussions?

Meer, M. van der; Vossen, P.; Jonker, C.M.; Murukannaiah, P.K.; Bouamor, H.; Pino, J.; Bali, K.

Do Differences in Values Influence Disagreements in Online Discussions?

Michiel van der Meer
LIACS
Leiden University
m.t.van.der.meer@liacs.leidenuniv.nl

Piek Vossen
CLTL
Vrije Universiteit Amsterdam
p.t.j.m.vossen@vu.nl

Catholijn M. Jonker
Interactive Intelligence
TU Delft
c.m.jonker@tudelft.nl

Pradeep K. Murukannaiah
Interactive Intelligence
TU Delft
p.k.murukannaiah@tudelft.nl

Abstract

Disagreements are common in online discussions. Disagreement may foster collaboration and improve the quality of a discussion under some conditions. Although there exist methods for recognizing disagreement, a deeper understanding of factors that influence disagreement is lacking in the literature. We investigate a hypothesis that differences in *personal values* are indicative of disagreement in online discussions. We show how state-of-the-art models can be used for estimating values in online discussions and how the estimated values can be aggregated into value profiles. We evaluate the estimated value profiles based on human-annotated agreement labels. We find that the dissimilarity of value profiles correlates with disagreement in specific cases. We also find that including value information in agreement prediction improves performance.

1 Introduction

A large number of users participate in online deliberations on societal issues such as climate change (Beel et al., 2022) and vaccination hesitancy (Weinzierl and Harabagiu, 2022). Disagreement is an important aspect of a deliberation (Polletta and Gardner, 2018) since it can (1) drive novel ideas, (2) incentivize evaluation of the proposed ideas, (3) avoid echo chambers, and (4) cancel out individual biases (Klein, 2012). Discussions with disagreement help users understand the opposing viewpoints (Lin and Kim, 2023; Saveski et al., 2022). Further, discussions having adequate disagreement have been associated with a higher quality deliberation (Esterling et al., 2015).

Ensuring that participants express a sufficient level of disagreement in a discussion is hard. We do not know the nature of disagreement in online platforms (Stromer-Galley et al., 2020). Further, questions arise on how to control for disagreement to enhance reciprocity (Esau and Friess, 2022), and how too much exposure to opposing views drives polarization (Bail et al., 2018). Analysis methods for online discussions currently cannot accurately represent such diverse perspectives (Cabitza et al., 2023; van der Meer et al., 2022a), and measuring deliberative quality is an open challenge (Vecchi et al., 2021; Shortall et al., 2022).

We want to ensure that a discussion incorporates many perspectives and that those are actively communicated. For this reason, we turn to *value conflicts*, a potential root cause for disagreement. We consider the hypothesis that when users with conflicting values engage in a discussion, diverging views come up. Perspective and value clashes are at the heart of disagreement (Stromer-Galley and Muhlberger, 2009). In collaborative teams, value conflicts are linked to disagreement (Jehn, 1994). Specifically, values are said to be an effective way to make conflict explicit among participants in a discussion (Beck et al., 2019).

To evaluate our hypothesis, we construct value profiles based on user comments on Reddit, a social media platform. A value profile captures the relative importance a user ascribes to values. We employ ten values, e.g., stimulation, universalism, and security, from the well-known Schwartz theory of basic values (Schwartz, 2012). Then, we compare the similarities among profiles to the disagreement among users on different topics. This allows us to investigate the association between value conflict (low similarity) and disagreement. Figure 1 shows an overview of our approach.

We gather 11.4M comments from 19K users on Reddit to construct value profiles. We perform up to 200 tests with different settings to investigate our hypothesis. We further experiment with replacing estimated value profiles with self-reported ones. To do so, we collect 572 judgments from 26 annotators

Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 15986–16008

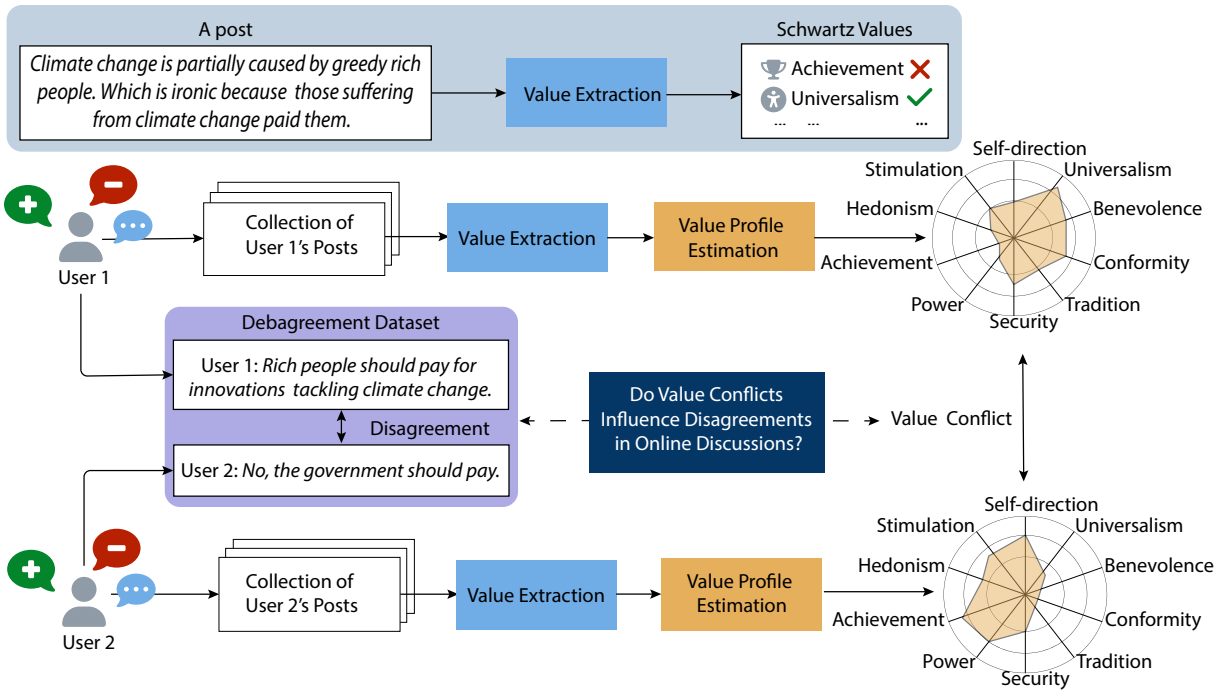


Figure 1: Setup of measuring value conflicts by means of Value Profile Estimation (VPE).

in combination with self-reported value profiles.

Selecting conversation partners based on their profile to manage value conflicts and influence the level of disagreement in a discussion could be a tool for moderators to balance conversations. To provide support for moderators, we investigate the impact of adding profile information to the agreement analysis task (Pougué-Biyong et al., 2021). Since the contextual implications of values are usually unknown, connecting user concerns to values (Alshomary et al., 2022) opens up human-machine collaboration opportunities for a more constructive conversation (Akata et al., 2020; Hadfi and Ito, 2022; Liscio et al., 2022b).

Contributions (1) We experiment with methods for value estimation from text to obtain value profiles from an online discourse (Reddit comments). (2) We investigate how value conflicts affect disagreement in discussions by showing that low-profile similarity can co-occur with disagreement under specific conditions for estimated and self-reported value profiles. (3) We make first steps in using the value-laden background information for predicting user disagreement and comparing it to other user-specific contextual information.

2 Related Work

Although there is existing work on analyzing agreement in online discussions, very few works focus on examining the reasons for disagreement. We review the existing work on agreement analysis, introduce two popular value theories, and outline previous research on value estimation.

2.1 (Dis)-agreement and discussion analysis

Detecting whether people agree or disagree with given statements is commonly framed as stance classification (e.g., ALDayel and Magdy, 2021). Recently, more effort has been put into exploring various aspects of the task (Hardalov et al., 2021; Allaway and McKeown, 2020; Liu et al., 2021). However, little work is done in adjusting the task to detect stances among users within online discussions. To this end, **agreement analysis** focuses on detecting (dis-)agreement in data that (1) represents realistic online discussions, (2) provides contextual information (post authors, timestamps, etc.), (3) contains diverse writing styles, (4) touches on multiple topics (Pougué-Biyong et al., 2021).

Existing work on agreement analysis is aimed at (1) identifying language that indicates disagreement (e.g., Niculae and Danescu-Niculescu-Mizil, 2016; Wojatzki et al., 2018; Fischer et al., 2022), (2) leveraging stylistic choices like sarcasm for detecting disagreement (Ghosh et al., 2021), (3) finding stance and target pairs, followed by the traditional stance classification (e.g., Chen et al., 2019; De Kock and Vlachos, 2021), and (4) mixing de-

tailed opinion information using e.g., logic of evaluation (Draws et al., 2022). Recently, adding social role context to textual comments was shown to have a positive impact on the agreement analysis task (Luo et al., 2023), which indicates the usefulness of background information. In this work, we focus on capturing the implicit motivations underlying opinions using *personal values*, which have been known to drive individual opinions and actions across cultures (Schwartz, 2012).

2.2 Value models

Values explain ideological beliefs underlying actions and opinions and may guide the design of applications (Friedman et al., 2013). Two leading value models have been used in NLP research: Moral Foundations (Graham et al., 2013) and the Schwartz Value model (Schwartz, 2012). Each of these models includes a set of general values.

The Moral Foundation Theory (MFT) includes five foundations, each a vice–virtue dichotomy (e.g., *harm–care*). However, MFT does not stipulate any relationship among the foundations. In contrast, the Schwartz model includes ten basic values organized as a circumplex (right-hand side of Figure 1), where similar values are placed close to each other. Further, Schwartz values can be grouped into four classes: *openness to change*, *conservation*, and *self-transcendence*, *self-enhancement*. Since the Schwartz model has more values and a structure among the values, it is better suited than MFT for comparing the value profiles of individuals. Thus, we employ Schwartz values in our work.

2.3 Value estimation

Most works based on representing an individual's value priorities (value profiles) use explicit preference elicitation, such as self-reporting and questionnaires (e.g., Boyd et al., 2021). However, a promising behavior-based approach focuses on analyzing textual motivations (Chen et al., 2014). To this end, dictionary-based approaches can be used for finding value mentions in texts (Ponizovskiy et al., 2020; Graham et al., 2009). Using such lexicons shows promising results in large-scale value estimation applications (Silva et al., 2021).

Recently, datasets annotated with personal values for training NLP methods have been released. In this paper, we use two recent datasets annotated with Schwartz values: (1) ValueNet (Qiu et al., 2021) is a dataset containing textual scenarios related to moral decision-making that have been annotated with relevant Schwartz values. (2) ValueArg (Kiesel et al., 2022) contains user-submitted arguments that relate to specific Schwartz values.

There are some datasets on MFT values, e.g., (Trager et al., 2022; Lourie et al., 2021; Hoover et al., 2020). These datasets include value annotations for messages but do not include a link between the messages and users. Thus, estimating value profiles from such datasets is not possible.

Applications include dialogues about moral scenarios (Qiu et al., 2021), review texts (Obie et al., 2021), and value-laden arguments (Kobbe et al., 2020; Alshomary et al., 2022). However, both the annotation and extraction of values remain difficult, with specific questions relating to the granularity of the value labels (Kiesel et al., 2022), their transfer to new domains (Liscio et al., 2022a), and how classifiers understand morality in language (Liscio et al., 2023). Moreover, large variances exist between the frequency of values across domain (Kennedy et al., 2021), and even the relevance of values differs depending on the domain (Bouman et al., 2018; Liscio et al., 2021). However, users can still be represented inside each domain by examining relative frequencies inside value profiles, as stipulated by Schwartz (2012).

3 Method

Figure 1 shows an overview of our method. We collect posts from users in online discussions. Using a trained value estimation model, we aggregate predictions over the collection to form a value profile. Then, to evaluate our hypothesis, we compare the value profiles for users known to be in disagreement based on an existing dataset. Our code[1] and data (van der Meer et al., 2023) is available online.

3.1 Data

We use **Debagreement** (Pougué-Biyong et al., 2021) as the dataset containing (dis-)agreement labels. This dataset contains user-submitted post pairs in English from five topics (Table 1), with post pairs annotated as {agree, neutral, disagree} by at least three crowd annotators.

We gather additional posts through the Reddit API using the usernames available in the Debagreement dataset. For each user still active, we collect up to 1000 most recent posts, which can be in any subreddit. The resulting posts range from Septem-

[1] https://github.com/m0re4u/value-disagreement

ber 2015 to April 2022. Subreddits host content on a variety of topics, not all of which encourage users to provide opinions based on their values. We are interested in finding preferences among values with respect to widespread societal issues, such as climate change. Thus, we filter out posts that are not likely to be of relevance to such issues. We (1) exclude Not Safe For Work and entertainment-related subreddits, removing 1.4M posts, (2) filter out noisy low-frequency subreddits (those with less than 50 collected posts), removing an additional 850K posts, and (3) retain only English text posts, removing 377K posts. Table 1 shows the amount of data collected after filtering.

Subcorpus	# users	# found	# comments
BREXIT	722	543	372K
CLIMATE	4580	3778	2.2M
BLM	2516	2121	1.1M
DEMOCRATS	6925	5646	3.8M
REPUBLICAN	8832	6839	3.9M

Table 1: List of subcorpora gathered in Debagreement.

3.2 Value Extraction

We formulate the value estimation task as recognizing whether a comment is related to a value by means of binary classification per value, matching the setup of Qiu et al. (2021). Our training data comprises general texts annotated for the presence of values across multiple domains. We combine data from two sources.

(1) **ValueNet** (Qiu et al., 2021): We collapse non-neutral labels (1 and -1) into a single positive class and take the neutral labels (0) as a negative class. A non-neutral utility means that annotators considered the value to be relevant to the scenario, whereas the neutral class indicates that the value plays no apparent role.

(2) **ValueArg** (Kiesel et al., 2022): Their annotation scheme uses an updated (20) Schwartz values (Schwartz et al., 2012), which we map back to the original 10 Schwartz values to allow joint training with the ValueNet dataset.

We train all models with 10 seeds on random splits of learning data into train and validation sets to observe training stability. For both datasets, we split data into predefined learning (training and validation) and evaluation (test) sets. We ensure that all ten values occur equally frequently in the evaluation set. Each text sample is presented to our model ten times, once for each value by prepending a value-specific token. We describe the additional hyperparameters in the Appendix.

3.3 Value Profile Estimation

Using a trained model, we construct a value profile v per user by summing over value estimations of all individual messages. We assume relative frequencies of value mentions to be indicative of value preference similar to Siebert et al. (2022).

To measure value conflicts, we introduce a lower limit l on the total value mentions in each profile, i.e., requiring that each user has at least l posts related to at least one value. Further, we normalize profile mention count by dividing it by the total number of value mentions per user. After this preprocessing, we compute the similarity $\mathcal{S}$ between two value profiles v and w in multiple ways.

Kendall τ We sort value mentions by frequency and assign a rank label to each value. Kendall's rank correlation metric τ is a robust measure of correlation (Croux and Dehon, 2010), and considers the ranks of all pairs of values. If a pair of values is ranked differently in v than in w, the pair is considered discordant. Low scores indicate value conflict.

$$\mathcal{S}^{\tau}(v,w) = 1 - \frac{2 \times (\text{\# discordant pairs})}{\binom{n}{2}} \quad (1)$$

Manhattan Distance (MD) We compute the absolute difference between two profiles. High scores indicate value conflict.

$$\mathcal{S}^{MD}(v,w) = \sum_{i=1}^{n} |v_i - w_i| \quad (2)$$

Cosine (CO) We compute traditional cosine similarity, low scores indicate conflict.

$$\mathcal{S}^{CO}(v,w) = \frac{v \cdot w}{||v||\ ||w||} \quad (3)$$

Weighted-cosine (WC) We compute a weighted cosine similarity that weighs similarities between values using the Schwartz Value Circumplex Model. For computing the similarity between value v_i and v_j, we use a similarity matrix $\mathcal{B}$ constructed using a normal distribution with $\sigma = 1$ centered on each value. Low scores indicate conflict.

$$\mathcal{S}^{WC}(v,w) = \frac{\sum_{i=1}^{n} \mathcal{B}_i v_i w_i}{\sqrt{\sum_{i=1}^{n} \mathcal{B}_i v_i^2}\sqrt{\sum_{i=1}^{n} \mathcal{B}_i w_i^2}} \quad (4)$$

4 Experiments and Results

We train models for value extraction and use those models to estimate value profiles. We check the consistency of our results with previous work, investigate differences in value profiles of disagreeing users, and perform qualitative analyses.

Method	Training	Test		
		ValueNet	ValueArg	Both
All-ones	–	0.40	0.11	0.26
Value Dict.	–	0.45	0.64	0.57
(Kiesel et al., 2022)*	ValueArg	0.15	0.37	0.28
(Qiu et al., 2021)*	ValueNet	0.59	0.52	0.57
BERT_{VE}	ValueNet	**0.66**	0.57	0.65
	ValueArg	0.46	0.76	0.67
	Both	0.63	**0.81**	**0.79**
$\text{RoBERTa}_{\text{VE}}$	ValueNet	0.62	0.59	0.63
	ValueArg	0.46	0.76	0.67
	Both	0.63	0.78	0.78

Table 2: Macro-averaged F_1 scores of the value estimation approaches on the value datasets. Methods marked with * are adapted for our comparison.

4.1 Training Models for Value Estimation

We experiment with two popular BERT-based models, BERT (Devlin et al., 2019) and RoBERTa (Liu and Lapata, 2019), for value estimation. Further, we employ multiple baselines: (1) always predict all values for a comment ("All-ones"), (2) predict values based on mentions of value words from the **Schwartz Value Dictionary** (Ponizovskiy et al., 2020), (3) the multi-label approach from Kiesel et al. (2022), which uses an expanded label set, and (4) the utility model from Qiu et al. (2021). The latter two baselines are BERT-based models. For Kiesel et al. (2022), we use their multi-label setup to make predictions and map to the 10 Schwartz values at inference time (*humility* and *face* are not mapped to any value). Similarly, we map the rounded ternary utility labels from Qiu et al. (2021) into binary value relevance labels at inference.

Table 2 shows the F_1 scores for the value extraction methods for different combinations of training and test datasets. We outperform all our baselines, including those from previous work. BERT_{VE} and $\text{RoBERTa}_{\text{VE}}$ yield similar F_1 scores, and they perform best when trained on both datasets. We use our best-performing BERT_{VE} model, trained on *both* datasets, to construct the value profiles in the rest of the experiments.

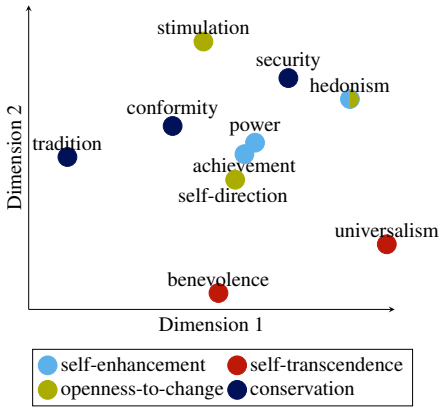


Figure 2: Visualization of the covariance between values in estimated profiles.

4.2 Value Profile Estimation

Table 3 shows the top two frequent values in each domain. We observe that the distribution of values is specific to discussion contexts. For example, although stimulation is a common and frequent value, it is not the most frequent value in the BREXIT subcorpus. We aggregate the values extracted for each user into their value profile. Table 3 (last column) shows the mean pairwise τ distance (Equation 1) among the value profiles in each domain. We observe that the BLM subcorpus has the most diversity among the five subcorpora.

Subcorpus	Top Two Values	Avg. τ
BREXIT	Security, Stimulation	0.260
CLIMATE	Stimulation, Security	0.308
BLM	Self-direction, Stimulation	0.343
DEMOCRATS	Stimulation, Self-direction	0.319
REPUBLICAN	Stimulation, Security	0.315

Table 3: Frequent values, and the mean similarity among value profiles in each domain.

Next, to qualitatively assess the estimated value profiles, we normalize profiles (by the total number of value mentions) and compute covariance between profiles. Then, we perform metric Multi-Dimensional Scaling (MDS) of the covariance matrix similar to Ponizovskiy et al. (2020). Figure 2 shows a visualization of the first two dimensions after MDS. We observe that values that are close to each other in the Schwartz circumplex (Schwartz, 2012), e.g., achievement and power, also tend to be closer in the MDS visualization.

4.3 Value Conflicts and Disagreement

We aim to analyze whether value conflicts influence disagreement in online discussions, using measurements of similarity between value profiles. We evaluate the following alternative hypothesis ($\mathbf{H_a}$) against a null hypothesis ($\mathbf{H_0}$).

$\mathbf{H_0}$ The mean value profile similarity score between user pairs that disagree is equal to the mean value profile similarity score between user pairs that agree.

$\mathbf{H_a}$ The mean value profile similarity score between user pairs that disagree is lower than the mean value profile similarity score between user pairs that agree.

We report the Bayes' Factor (BF_{10}) [2] to assess the relative increase in odds for assuming the alternative over the null hypothesis after observing data (Azer et al., 2020). BF_{10} scores in $[3^{-1}, 3]$ are considered to indicate evidence for neither hypothesis, whereas more extreme values favor one hypothesis over the other, allowing us to make conclusions in either direction (Kass and Raftery, 1995).

We perform two experiments. First, we test the hypothesis for profiles constructed using the Value Profile Estimation (VPE) method. In the second experiment, we replace one of the profiles in each pair with a self-reported profile and agreement label. Thus, the second experiment removes some of the noise stemming from the VPE method.

4.3.1 Profiles from VPE

We split Debagreement based on *agree* and *disagree* labels (and drop all pairs with a neutral label), obtaining respectively G^+ and G^-. For each group, we compute the profile similarity scores using each method mentioned in Section 3.2. We do this per subreddit and observe the differences in score distributions. The alternative hypothesis is defined as the mean similarity scores in G^- being lower[3] than the mean for G^+:

$$\theta_G = \frac{1}{|G|} \sum_{\{p,c\} \in G} \mathcal{S}(p, c) \quad (5)$$

$$H_0 : \theta_{G^-} = \theta_{G^+} \quad (6)$$

$$H_a : \theta_{G^-} < \theta_{G^+} \quad (7)$$

We report the BF_{10} for all combinations of similarity methods and parameters. We run 100 tests, considering 5 subreddits, 4 similarity scores, and 5 value profile thresholds $l = \{1, 10, 50, 200, 500\}$. Figure 3 provides an overview of the BF_{10} scores.

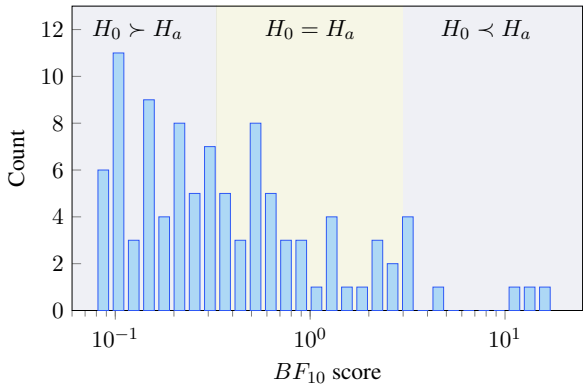


Figure 3: BF_{10} scores obtained for the combinations of data, value estimation methods, and scoring metrics.

First, we observe that a majority of the combinations show stronger support for accepting the null hypothesis over the alternative hypothesis (i.e., most scores fall inside the leftmost blue bin). This indicates that value conflicts may not be directly correlated to disagreement in many cases. Possibly, other content-related factors play a stronger role in these discussions. However, there are some tests that still show evidence for rejecting the null hypothesis ($BF_{10} > 3$).

Thus, given specific settings and domains, we can trace disagreement between users to value conflicts. Table 4 shows the tests where $BF_{10} > 3$. In all cases, the filter l was 10 or more, stipulating that populated value profiles are required for measuring value conflicts reliably. We observe that BLM, the subcorpus with the highest profile diversity (Table 3), is frequent among these positive cases. Thus, having diverse profiles increases the likelihood of finding a link between values and disagreement. One positive test result is observed for the BREXIT subcorpus for a high profile threshold (500). Brexit includes the smallest number of user profiles; the high profile threshold further removes several profiles. Thus, the positive result for BREXIT, based on a low number of profile comparisons, may not be reliable.

4.3.2 Mixing with Self-reported Profiles

Given that we use a novel method for estimating value profiles, we compare the results from the previous experiment with one that uses self-reported value profiles. Self-reported profiles mitigate the noise stemming from the value estimation step. The

[2] BF hypothesis tests are sensitive to the choice of prior. We use the implementation of `pingouin` (Vallat, 2018), which includes a Jeffreys-Zellner-Siow prior, an objective prior for two-sample cases (Rouder et al., 2009)

[3] Higher for the **MD** metric, which flips the sign in Eqn. 7.

BF_{10}	Subreddit	Similarity score	Profile threshold
17.451	BLM	CO	10
12.485	BLM	WC	10
10.504	BLM	τ	250
4.223	BLM	MD	10
3.442	BREXIT	WC	500
3.159	BLM	WC	50

Table 4: The six tests between two VPE-constructed profiles with $BF_{10} > 3$.

setup is identical to Section 4.3.1, but now we compute similarities between an estimated profile and a self-reported profile, obtained from a value survey.

We run a user study to obtain (1) self-reports of value profiles using an established value survey (PVQ-21, Schwartz, 2021), and (2) agreement labels on posts in Debagreement. We obtained an IRB approval (exempt status) for our study.

We collected annotations from 26 Prolific (prolific.co) users. We selected five task instances for each subreddit from Debagreement posts with populated value profiles, rendering testing on multiple profile thresholds unnecessary. We removed three task instances, which obtained a majority of neutral and not-enough-information judgments, leaving 22 rated instances. Thus, our analyses include a total of 572 judgments.

The results are shown in Figure 4. We observe that deciding between the two hypotheses is not possible, in a majority of cases, as most evidence attributed both as equally likely. However, it is interesting to notice that using self-reported value profiles shifts the majority of results from favoring the null hypothesis to the undecidable range. In combination with the results from the previous section, this indicates that VPE methods need careful evaluation with respect to self-reported profiles as both may contain errors stemming from different sources and may have complementary merits. VPE suffers from errors made by the value estimation model but has the potential to use large amounts of data. In contrast, although self-reports yield a profile directly, they may be prone to biases.

Two tests still show evidence in favor of accepting H_a (see Table 5). They are on two task instances in the same domain, DEMOCRATS, and are measured for the τ and MD metrics. Here, our results differ from the previous experiment, and

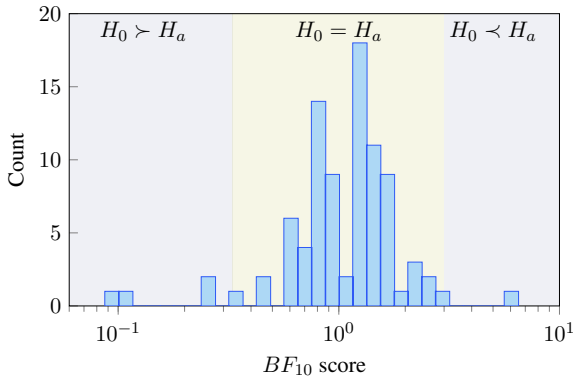


Figure 4: BF_{10} scores for all similarity scores and task instances comparing VPE and self-reported profiles.

different subreddits result in high BF_{10} scores. In this case, one user's value profile is constructed using self-reports, which are obtained without reference to discussions (i.e. not estimated from posts on Reddit). This may cause other factors to influence the diversity of profiles stemming from the PVQ. Furthermore, the task instances contained a call for action (e.g., *Please just vote [..]* and *The gloves should come off [..]*). The values embedded in the call to action may be one of the reasons why annotators felt inclined to disagree or agree.

BF_{10}	Subreddit	Similarity score
6.490	DEMOCRATS	τ
3.066	DEMOCRATS	MD
2.543	BREXIT	MD
2.407	BREXIT	CO
2.230	CLIMATE	CO

Table 5: The top-five BF_{10} scores, when comparing a VPE-constructed profile and a self-reported profile.

4.3.3 Qualitative Assessment

To better understand when value conflicts influence disagreement, we perform a qualitative analysis of some instances (comment pairs) from the dataset that follow our hypothesis and some that do not (Figure 11 in Appendix B shows such examples).

We identify five trends in misaligned instances. (1) **Not enough information** in a value profile (i.e., low-frequency value mentions). This means that the user posted little value-laden content or that the value extraction method erroneously ignored some value-laden comments. (2) **No apparent value-based reasoning** involved in the comments, e.g., factual answers to a question.

(3) **(Dis-)agreement** happens **on a content level** since profiles do not dictate individual utterances. This occurs when users disagree that a decision is "for good," but fail to motivate their motivations for what is "good." (4) The **target** of disagreement can be **partial**, whereas value conflicts are measured between two users. (5) In a few cases, the label given in Debagreement is **faulty** (e.g., annotators misinterpreting sarcasm or the text is vague).

4.4 Use Case: Predicting (Dis-)agreement

We assume that users' value profiles (in addition to the content of users' posts) play a role in predicting the agreement between users. We adopt the setup from Pougué-Biyong et al. (2021), where an agreement label is predicted between parent p and child comments c. We add extra information to p and c using four methods.

Random noise (ϵ) Random noise to test for spurious correlations.

User centroids (z) Centroids of all posts from a single user by constructing TF-IDF vectors for each post and then taking an average.

Explicit user features (u) Nine features commonly extracted for representing users on Reddit (e.g., (Jhaver et al., 2019; Chew et al., 2021)) to add extra contextual information.

Value profile (v) Value estimation on user posts to extract an explicit value profile for the ten Schwartz values.

We create embeddings (TF-IDF or BERT) for p and c and concatenate them to the user-specific context (Gu and Budhkar, 2021). We standardize the user-specific context information to avoid raw values having a large impact, similar to the value profiles (v). When training with user profiles, we subsample Debagreement to include only those (p, c) pairs in which we have background data for both p and c. This leaves 65% of the data (28K samples). We train our classifier on an 80/10/10 split, retaining the most recent 20% as validation and test sets to reflect a real-world training scenario on historical data (Søgaard et al., 2021).

Figure 5 shows the results. Classifiers using TF-IDF embeddings fail to use the information effectively. BERT outperforms both our baselines, in line with the results for (Pougué-Biyong et al., 2021). In this setting, none of the additional information causes major changes in performance, but we see an improvement using the value profiles and centroids. Compared to other work, using

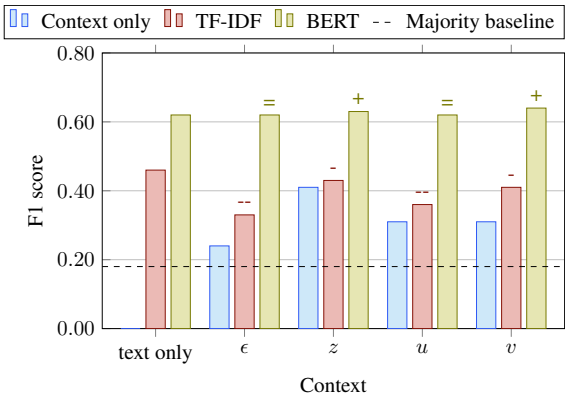


Figure 5: F_1 scores when adding extra context information. Symbols above bars show changes with respect to text-only: $--$ for $\Delta F_1 < -0.1$; $-$ for $-0.1 < \Delta F_1 < 0$; $=$ for $\Delta F_1 = 0$; and $+$ for $\Delta F_1 > 0$.

user-specific information is surprisingly difficult (Al Khatib et al., 2020). Further inspection for BERT indicates that the *neutral* class is hard to predict, as information from the value profiles may not be relevant. Mixing background information using, e.g., GNNs (Luo et al., 2023) may make more effective use of the profile information.

5 Conclusion

Our results on the role of value conflicts in disagreements are mixed. On the one hand, we mostly note negative evidence of a correlation between profile similarity and disagreeing users when using the VPE methods. When using self-reported profiles, the negative evidence reduces and results become inconclusive for a majority of the cases. This suggests that the nature of the profiles differs, and further investigation is necessary.

On the other hand, we observe that value conflicts were found to lead to disagreements in specific cases. When values are likely to be relevant and diverse, we find evidence for a correlation between value conflict and disagreement. While value conflicts may not be directly related to disagreement, they do signal diversity with respect to the underlying motivations of participants.

Using value profiles in combination with BERT performs marginally better than a text-only baseline in predicting agreement. Yet, VPE can be valuable for characterizing and enhancing diversity in discussions. Further, making participants value-aware could enhance the discussion quality.

Constructing profiles from behavioral cues, such as written opinions, is noisy. For future work, we hope to see the creation of resources that allow

end-to-end evaluation by combining text posts with a consistent set of users that allows aggregation to ground truth profiles or self-reported profiles. However, gathering such profile information outside controlled lab settings is highly complex. Future experiments may incorporate more judgments and provide stronger evidence for one hypothesis. These can be retrofitted with our results through Bayesian updating (Moerbeek, 2021).

Limitations

We outline four limitations of our work related to the experimental setup and the interpretation of results that are specific to the modeling of value conflicts in online discussions.

First, the value extraction methods we employ (see Table 2) may have unknown errors. Our work is not focused on optimizing value extraction, which is an emerging research direction (Kiesel et al., 2023). Adding more annotated Reddit data would allow us to judge the performance of value extraction models better. A future direction is to employ other training paradigms like Multi-task Learning (e.g., Fang et al., 2019) or techniques for mixing in general-purpose language models (e.g., van der Meer et al., 2022b).

Second, we obtain the self-reported value profiles with the PVQ-21 questionnaire (see Section 4.4). Since we run the questionnaire before starting an annotation experiment to obtain agreement labels, there may be ordering bias in the obtained labels. The experiments could be enhanced by swapping the order of PVQ-21 and the annotation tasks to estimate the effect of answering the questionnaire on the agreement labels.

Third, the reporting of our results is limited to the Bayes Factor (BF). Further, most of our results fall inside the neutral category ("cannot decide between H_0 and H_a"). We require more data to decide which of the hypotheses is more likely. An estimation of the posterior odds of the hypotheses e.g., in the form of *Highest Density Intervals* (HDI) might yield more insights, and would involve deciding on a *region of practical equivalence* (ROPE), as well as picking a thus far unknown prior distribution over the values for $\mathcal{S}$ in our two hypotheses (Kruschke, 2018). However, BF and HDI interpretations can be seen as complementary, respectively quantifying evidence or beliefs (van Ravenzwaaij and Wagenmakers, 2022).

Lastly, our qualitative findings are derived from examining online interactions with limited context. To obtain a more complete picture, both the values and the interpretation of the author's role in discussions should be verified by the authors themselves. However, running such experiments in controlled lab settings is beyond the scope of our work since we focus on disagreements in online discussions.

Ethics Statement

First, the dataset used to model online discussions, Debagreement, was sourced from online interactions between users on Reddit. Research conducted on Reddit data is biased to a WEIRD (Western, Educated, Industrialized, Rich, Democratic) demographic, and results may not generalize to a broader set of users (Proferes et al., 2021). However, our method outlines which data is required for performing the same analysis given the availability of richer data, not necessarily stemming from Reddit. Second, models for predicting values may be wrong, they may lead to harmful outcomes for particular groups or populations (Mehrabi et al., 2021). In any application, the incorporation of control mechanisms (i.e., providing users a way to influence the construction of their own value profile) is a requirement for making sure the value profiling is conducted in a transparent and accountable manner. Broadly, this work should further be situated in a system containing checks and balances, making sure any output stemming from automated classification is verified by human agents before having an effect on actual users.

Acknowledgements

This research was funded by the Netherlands Organisation for Scientific Research (NWO) through the Hybrid Intelligence Centre via the Zwaartekracht grant (024.004.022). We would like to thank the ARR reviewers for their feedback.

References

Zeynep Akata, Dan Balliet, Maarten de Rijke, Frank Dignum, Virginia Dignum, Guszti Eiben, Antske Fokkens, Davide Grossi, Koen Hindriks, Holger Hoos, Hayley Hung, Catholijn Jonker, Christof Monz, Mark Neerincx, Frans Oliehoek, Henry Prakken, Stefan Schlobach, Linda van der Gaag, Frank van Harmelen, Herke van Hoof, Birna van Riemsdijk, Aimee van Wynsberghe, Rineke Verbrugge, Bart Verheij, Piek Vossen, and Max Welling. 2020. A research agenda for hybrid intelligence: Augmenting human intellect with collaborative,

adaptive, responsible, and explainable artificial intelligence. *Computer*, 53(8):18–28.

Khalid Al Khatib, Michael Völske, Shahbaz Syed, Nikolay Kolyada, and Benno Stein. 2020. Exploiting personal characteristics of debaters for predicting persuasiveness. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7067–7072.

Abeer ALDayel and Walid Magdy. 2021. Stance detection on social media: State of the art and trends. *Information Processing & Management*, 58(4):102597.

Emily Allaway and Kathleen McKeown. 2020. Zero-Shot Stance Detection: A Dataset and Model using Generalized Topic Representations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8913–8931, Online. Association for Computational Linguistics.

Milad Alshomary, Roxanne El Baff, Timon Gurcke, and Henning Wachsmuth. 2022. The moral debater: A study on the computational generation of morally framed arguments. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8782–8797, Dublin, Ireland. Association for Computational Linguistics.

Erfan Sadeqi Azer, Daniel Khashabi, Ashish Sabharwal, and Dan Roth. 2020. Not all claims are created equal: Choosing the right statistical approach to assess hypotheses. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5715–5725.

Christopher A. Bail, Lisa P. Argyle, Taylor W. Brown, John P. Bumpus, Haohan Chen, M. B. Fallin Hunzaker, Jaemin Lee, Marcus Mann, Friedolin Merhout, and Alexander Volfovsky. 2018. Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences*, 115(37):9216–9221.

Jordan Beck, Bikalpa Neupane, and John M. Carroll. 2019. Managing conflict in online debate communities. *First Monday*, 24(7).

Jacob Beel, Tong Xiang, Sandeep Soni, and Diyi Yang. 2022. Linguistic characterization of divisive topics online: Case studies on contentiousness in abortion, climate change, and gun control. *Proceedings of the International AAAI Conference on Web and Social Media*, 16(1):32–42.

Thijs Bouman, Linda Steg, and Henk A. L. Kiers. 2018. Measuring values in environmental research: A test of an environmental portrait value questionnaire. *Frontiers in Psychology*, 9.

Ryan Boyd, Steven Wilson, James Pennebaker, Michal Kosinski, David Stillwell, and Rada Mihalcea. 2021. Values in words: Using language to evaluate and understand personal values. *Proceedings of the International AAAI Conference on Web and Social Media*, 9(1):31–40.

Federico Cabitza, Andrea Campagner, and Valerio Basile. 2023. Toward a perspectivist turn in ground truthing for predictive computing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(6):6860–6868.

Jilin Chen, Gary Hsieh, Jalal U. Mahmud, and Jeffrey Nichols. 2014. Understanding individuals' personal values from social media word use. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*, CSCW '14, page 405–414, New York, NY, USA. Association for Computing Machinery.

Sihao Chen, Daniel Khashabi, Wenpeng Yin, Chris Callison-Burch, and Dan Roth. 2019. Seeing things from a different angle: Discovering diverse perspectives about claims. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 542–557.

Robert Chew, Caroline Kery, Laura Baum, Thomas Bukowski, Annice Kim, and Mario Navarro. 2021. Predicting age groups of reddit users based on posting behavior and metadata: Classification model development and validation. *JMIR Public Health Surveill*, 7(3):e25807.

Christophe Croux and Catherine Dehon. 2010. Influence functions of the spearman and kendall correlation measures. *Statistical Methods & Applications*, 19(4):497–515.

Christine De Kock and Andreas Vlachos. 2021. I beg to differ: A study of constructive disagreement in online conversations. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2017–2027.

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

Tim Draws, Oana Inel, Nava Tintarev, Christian Baden, and Benjamin Timmermans. 2022. Comprehensive viewpoint representations for a deeper understanding of user interactions with debated topics. In *ACM SIGIR Conference on Human Information Interaction and Retrieval*, pages 135–145.

K Esau and Dennis Friess. 2022. What creates listening online? exploring reciprocity in online political discussions with relational content analysis. *Journal of Deliberative Democracy*, 18(1):1–16.

Kevin M Esterling, Archon Fung, and Taeku Lee. 2015. How much disagreement is good for democratic deliberation? *Political Communication*, 32(4):529–551.

Wei Fang, Moin Nadeem, Mitra Mohtarami, and James Glass. 2019. Neural multi-task learning for stance prediction. In *Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER)*, pages 13–19, Hong Kong, China. Association for Computational Linguistics.

Ken Fischer, Justin Reedy, Cameron Piercy, and Rashmi Thapaliya. 2022. A typology of reasoning in deliberative processes: A study of the 2010 oregon citizens' initiative review. *Journal of Deliberative Democracy*, 18(2).

Batya Friedman, Peter H. Kahn, Alan Borning, and Alina Huldtgren. 2013. *Value Sensitive Design and Information Systems*, pages 55–95. Springer Netherlands, Dordrecht.

Debanjan Ghosh, Ritvik Shrivastava, and Smaranda Muresan. 2021. "laughing at you or with you": The role of sarcasm in shaping the disagreement space. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1998–2010, Online. Association for Computational Linguistics.

Jesse Graham, Jonathan Haidt, Sena Koleva, Matt Motyl, Ravi Iyer, Sean P. Wojcik, and Peter H. Ditto. 2013. Moral foundations theory: The pragmatic validity of moral pluralism. In Patricia Devine and Ashby Plant, editors, *Advances in Experimental Social Psychology*, volume 47, pages 55–130. Academic Press.

Jesse Graham, Jonathan Haidt, and Brian A Nosek. 2009. Liberals and conservatives rely on different sets of moral foundations. *Journal of personality and social psychology*, 96(5):1029–1046.

Ken Gu and Akshay Budhkar. 2021. A package for learning on tabular and text data with transformers. In *Proceedings of the Third Workshop on Multimodal Artificial Intelligence*, pages 69–73, Mexico City, Mexico. Association for Computational Linguistics.

Rafik Hadfi and Takayuki Ito. 2022. Augmented democratic deliberation: Can conversational agents boost deliberation in social media? In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*, pages 1794–1798.

Momchil Hardalov, Arnav Arora, Preslav Nakov, and Isabelle Augenstein. 2021. Cross-domain label-adaptive stance detection. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9011–9028, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Joe Hoover, Gwenyth Portillo-Wightman, Leigh Yeh, Shreya Havaldar, Aida Mostafazadeh Davani, Ying Lin, Brendan Kennedy, Mohammad Atari, Zahra Kamel, Madelyn Mendlen, Gabriela Moreno, Christina Park, Tingyee E. Chang, Jenna Chin, Christian Leong, 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.

Karen A. Jehn. 1994. Enhancing effectiveness: An investigation of advantages and disadvantages of value-based intragroup conflict. *International Journal of Conflict Management*, 5(3):223–238.

Shagun Jhaver, Amy Bruckman, and Eric Gilbert. 2019. Does transparency in moderation really matter? user behavior after content removal explanations on reddit. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW).

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain. Association for Computational Linguistics.

Robert E. Kass and Adrian E. Raftery. 1995. Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795.

Brendan Kennedy, Mohammad Atari, Aida Mostafazadeh Davani, Joe Hoover, Ali Omrani, Jesse Graham, and Morteza Dehghani. 2021. Moral concerns are differentially observable in language. *Cognition*, 212:104696.

Johannes Kiesel, Milad Alshomary, Nicolas Handke, Xiaoni Cai, Henning Wachsmuth, and Benno Stein. 2022. Identifying the human values behind arguments. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4459–4471.

Johannes Kiesel, Milad Alshomary, Nailia Mirzakhmedova, Maximilian Heinrich, Nicolas Handke, Henning Wachsmuth, and Benno Stein. 2023. SemEval-2023 task 4: ValueEval: Identification of human values behind arguments. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2287–2303, Toronto, Canada. Association for Computational Linguistics.

Mark Klein. 2012. Enabling large-scale deliberation using attention-mediation metrics. *Computer Supported Cooperative Work (CSCW)*, 21(4-5):449–473.

Jonathan Kobbe, Ines Rehbein, Ioana Hulpuș, and Heiner Stuckenschmidt. 2020. Exploring morality in argumentation. In *Proceedings of the 7th Workshop on Argument Mining*, pages 30–40, Online. Association for Computational Linguistics.

John K. Kruschke. 2018. Rejecting or accepting parameter values in bayesian estimation. *Advances in Methods and Practices in Psychological Science*, 1(2):270–280.

Han Lin and Yonghwan Kim. 2023. Learning from disagreement on social media: The mediating role of like-minded and cross-cutting discussion and the moderating role of fact-checking. *Computers in Human Behavior*, 139:107558.

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 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14113–14132, Toronto, Canada. Association for Computational Linguistics.

Enrico Liscio, Alin E Dondera, Andrei Geadau, Catholijn M Jonker, and Pradeep K Murukannaiah. 2022a. Cross-domain classification of moral values. In *Findings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics. NAACL*, volume 22, pages 1–13.

Enrico Liscio, Michiel van der Meer, Luciano C. Siebert, Catholijn M. Jonker, Niek Mouter, and Pradeep K. Murukannaiah. 2021. Axies: Identifying and evaluating context-specific values. In *Proceedings of the 20th Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '21, pages 799–808, London.

Enrico Liscio, Michiel van der Meer, Luciano C. Siebert, Catholijn M. Jonker, and Pradeep K. Murukannaiah. 2022b. What values should an agent align with? An empirical comparison of general and context-specific values. *Autonomous Agents and Multi-Agent Systems*, 36(1):23.

Rui Liu, Zheng Lin, Yutong Tan, and Weiping Wang. 2021. Enhancing zero-shot and few-shot stance detection with commonsense knowledge graph. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3152–3157, Online. Association for Computational Linguistics.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. *arXiv preprint arXiv:1908.08345*.

Nicholas Lourie, Ronan Le Bras, and Yejin Choi. 2021. Scruples: A corpus of community ethical judgments on 32,000 real-life anecdotes. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15):13470–13479.

Yun Luo, Zihan Liu, Stan Z. Li, and Yue Zhang. 2023. Improving (dis)agreement detection with inductive social relation information from comment-reply interactions. In *Proceedings of the ACM Web Conference 2023*, WWW '23, page 1584–1593, New York, NY, USA. Association for Computing Machinery.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Comput. Surv.*, 54(6).

Mirjam Moerbeek. 2021. Bayesian updating: increasing sample size during the course of a study. *BMC Medical Research Methodology*, 21(1):137.

Ines Montani and Matthew Honnibal. 2022. Prodigy: A modern and scriptable annotation tool for creating training data for machine learning models.

Vlad Niculae and Cristian Danescu-Niculescu-Mizil. 2016. Conversational markers of constructive discussions. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 568–578.

Humphrey O. Obie, Waqar Hussain, Xin Xia, John Grundy, Li Li, Burak Turhan, Jon Whittle, and Mojtaba Shahin. 2021. A first look at human values-violation in app reviews. In *2021 IEEE/ACM 43rd International Conference on Software Engineering: Software Engineering in Society (ICSE-SEIS)*, pages 29–38.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(85):2825–2830.

Francesca Polletta and Beth Gardner. 2018. The Forms of Deliberative Communication. In *The Oxford Handbook of Deliberative Democracy*. Oxford University Press.

Vladimir Ponizovskiy, Murat Ardag, Lusine Grigoryan, Ryan Boyd, Henrik Dobewall, and Peter Holtz. 2020. Development and validation of the personal values dictionary: A theory–driven tool for investigating references to basic human values in text. *European Journal of Personality*, 34(5):885–902.

John Pougué-Biyong, Valentina Semenova, Alexandre Matton, Rachel Han, Aerin Kim, Renaud Lambiotte, and Doyne Farmer. 2021. Debagreement: A comment-reply dataset for (dis) agreement detection in online debates. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Nicholas Proferes, Naiyan Jones, Sarah Gilbert, Casey Fiesler, and Michael Zimmer. 2021. Studying Reddit: A systematic overview of disciplines, approaches, methods, and ethics. *Social Media + Society*, 7(2):20563051211019004.

Liang Qiu, Yizhou Zhao, Jinchao Li, Pan Lu, Baolin Peng, Jianfeng Gao, and Song-Chun Zhu. 2021. Valuenet: A new dataset for human value driven dialogue system. *arXiv preprint arXiv:2112.06346*.

Jeffrey N. Rouder, Paul L. Speckman, Dongchu Sun, Richard D. Morey, and Geoffrey Iverson. 2009. Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, 16(2):225–237.

Martin Saveski, Nabeel Gillani, Ann Yuan, Prashanth Vijayaraghavan, and Deb Roy. 2022. Perspective-taking to reduce affective polarization on social media. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, pages 885–895.

Shalom Schwartz. 2021. A repository of schwartz value scales with instructions and an introduction. *Online Readings in Psychology and Culture*, 2.

Shalom H Schwartz. 2012. An overview of the schwartz theory of basic values. *Online readings in Psychology and Culture*, 2(1):2307–0919.

Shalom H Schwartz, Jan Cieciuch, Michele Vecchione, Eldad Davidov, Ronald Fischer, Constanze Beierlein, Alice Ramos, Markku Verkasalo, Jan-Erik Lönnqvist, Kursad Demirutku, et al. 2012. Refining the theory of basic individual values. *Journal of personality and social psychology*, 103(4):663.

Ruth Shortall, Anatol Itten, Michiel van der Meer, Pradeep Murukannaiah, and Catholijn Jonker. 2022. Reason against the machine? Future directions for mass online deliberation. *Frontiers in Political Science*.

Luciano C Siebert, Enrico Liscio, Pradeep K Murukannaiah, Lionel Kaptein, Shannon Spruit, Jeroen van den Hoven, and Catholijn Jonker. 2022. Estimating value preferences in a hybrid participatory system. In *Proceedings of the first International Conference on Hybrid Human-Artificial Intelligence (HHAI 2022)*, pages 1–14, Amsterdam, the Netherlands. IOS Press.

Amila Silva, Pei-Chi Lo, and Ee Peng Lim. 2021. On predicting personal values of social media users using community-specific language features and personal value correlation. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, pages 680–690.

Anders Søgaard, Sebastian Ebert, Jasmijn Bastings, and Katja Filippova. 2021. We need to talk about random splits. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1823–1832, Online. Association for Computational Linguistics.

Jennifer Stromer-Galley, Lauren Bryant, and Bruce Bimber. 2020. Context and medium matter: Expressing disagreements online and face-to-face in political deliberations. *Journal of Deliberative Democracy*, 11(1).

Jennifer Stromer-Galley and Peter Muhlberger. 2009. Agreement and disagreement in group deliberation: Effects on deliberation satisfaction, future engagement, and decision legitimacy. *Political Communication*, 26(2):173–192.

Jackson Trager, Alireza S Ziabari, Aida Mostafazadeh Davani, Preni Golazazian, Farzan Karimi-Malekabadi, Ali Omrani, Zhihe Li, Brendan Kennedy, Nils Karl Reimer, Melissa Reyes, et al. 2022. The moral foundations reddit corpus. *arXiv preprint arXiv:2208.05545*.

Raphael Vallat. 2018. Pingouin: statistics in python. *Journal of Open Source Software*, 3(31):1026.

Michiel van der Meer, Enrico Liscio, Catholijn M. Jonker, Aske Plaat, Piek Vossen, and Pradeep K. Murukannaiah. 2022a. HyEnA: A Hybrid Method for Extracting Arguments from Opinions. In *Proceedings of the first International Conference on Hybrid Human-Artificial Intelligence (HHAI 2022)*, pages 1–15, Amsterdam, the Netherlands. IOS Press.

Michiel van der Meer, Myrthe Reuver, Urja Khurana, Lea Krause, and Selene Baez Santamaria. 2022b. Will it blend? Mixing training paradigms & prompting for argument quality prediction. In *Proceedings of the 9th Workshop on Argument Mining*, pages 95–103, Online and in Gyeongju, Republic of Korea. International Conference on Computational Linguistics.

Michiel van der Meer, Piek Vossen, Catholijn M. Jonker, and Pradeep K. Murukannaiah. 2023. Do differences in values influence disagreements in online discussions? Supplementary material.

Don van Ravenzwaaij and Eric-Jan Wagenmakers. 2022. Advantages masquerading as "issues" in bayesian hypothesis testing: A commentary on tendeiro and kiers (2019). *Psychological Methods*, 27(3):451–465.

Eva Maria Vecchi, Neele Falk, Iman Jundi, and Gabriella Lapesa. 2021. Towards argument mining for social good: A survey. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1338–1352.

Maxwell A Weinzierl and Sanda M Harabagiu. 2022. From hesitancy framings to vaccine hesitancy profiles: A journey of stance, ontological commitments and moral foundations. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, pages 1087–1097.

Michael Wojatzki, Torsten Zesch, Saif Mohammad, and Svetlana Kiritchenko. 2018. Agree or disagree: Predicting judgments on nuanced assertions. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 214–224.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

A Methodological details

A.1 Training Value extraction methods

For training our Transformer-based NLP models, we turned to the Huggingface `transformers` Python package (Wolf et al., 2020). See Table 6 for the hyperparameters used for training value extraction models. All computational experiments were run on machines containing up to 2x 3090 Nvidia RTX GPUs. Training a single value extraction model takes around 3 hours. Running VPE on background data takes significantly longer due to the number of inferences made, up to 7 days of computation.

Hyperparameter	Value
`train epochs`	10
`learning rate`	$5e-05$
`model`	bert-base-uncased
`batch size`	256

Table 6: Hyperparameters used for training models for value extraction

Filtering Reddit data We construct value profiles from the data scraped from Reddit, from which we filter posts not likely to be of relevance to discussing widespread societal issues. We remove posts from (1) NSFW subreddits[4], (2) gaming subreddits[5], (3) image-related subreddits[6], (4) user subreddits, all subreddits starting with "`u_`", (5) non-English posts (as detected using the FastText (Joulin et al., 2017) Language Identification model[7]), (6) and subreddits for which we could extract less than 50 posts.

Using Value Dictionary for VPE We use the following pipeline for constructing value profiles using the **Schwartz Value Dictionary**.

1. Load words from Ponizovskiy et al. (2020). Some values have more words in the dictionary, and thus we introduce a weighting scheme to normalize over the number of words, such that a value v inside the profile with relatively few dictionary words has a higher weight w_v.
2. Replace URLs with a special `[URL]` token.
3. Apply lemmatization to all comments from a single user.
4. Classify individual comments for values. If a comment contains at least one term from the VD, classify the comment as being relevant for that value.
5. Aggregate over all comments.
6. Apply weighting $z = count(v) \times w_v$.
7. Apply normalization over the profile so it sums to 1.

A.2 Annotator experiment

We separated our annotator experiment into two phases: (1) the filling in of the PVQ-21, and (2) providing judgments on posts from Debagreement. The first phase was performed through Qualtrics questionnaire software. We provide screenshots of all steps (informed consent, annotation instructions) below. The second phase is hosted on Prodigy (Montani and Honnibal, 2022).

- **Informed consent** See Figure 6. Shown to users before starting the experiment outlining the data protection and disclaimers of any risks.
- **Value Survey** See Figure 7. Users fill in 21 items on a Likert scale.
- **Annotation instructions** See Figure 8.
- **Annotation interface** See Figure 9. Users were asked to fill in 25 task instances (five per subcorpus) on the annotation platform.

Annotators were recruited from the Prolific (`prolific.co`) crowd worker platform. All participants were paid at least the recommended £9/h wage, and on average spent 20 minutes on the two tasks combined. This payment is considered an ethical reward according to Prolific.

Transforming survey responses into profiles We adopt the suggestions from Schwartz (2012) for constructing a numerical value profile that reflects preferences among values. We create the following pipeline:

[4] `https://www.reddit.com/r/ListOfSubreddits/wiki/nsfw`

[5] `https://www.reddit.com/r/gaming/wiki/faq`

[6] `https://www.reddit.com/r/ListOfSubreddits/wiki/sfwporn`

[7] `https://fasttext.cc/docs/en/language-identification.html`

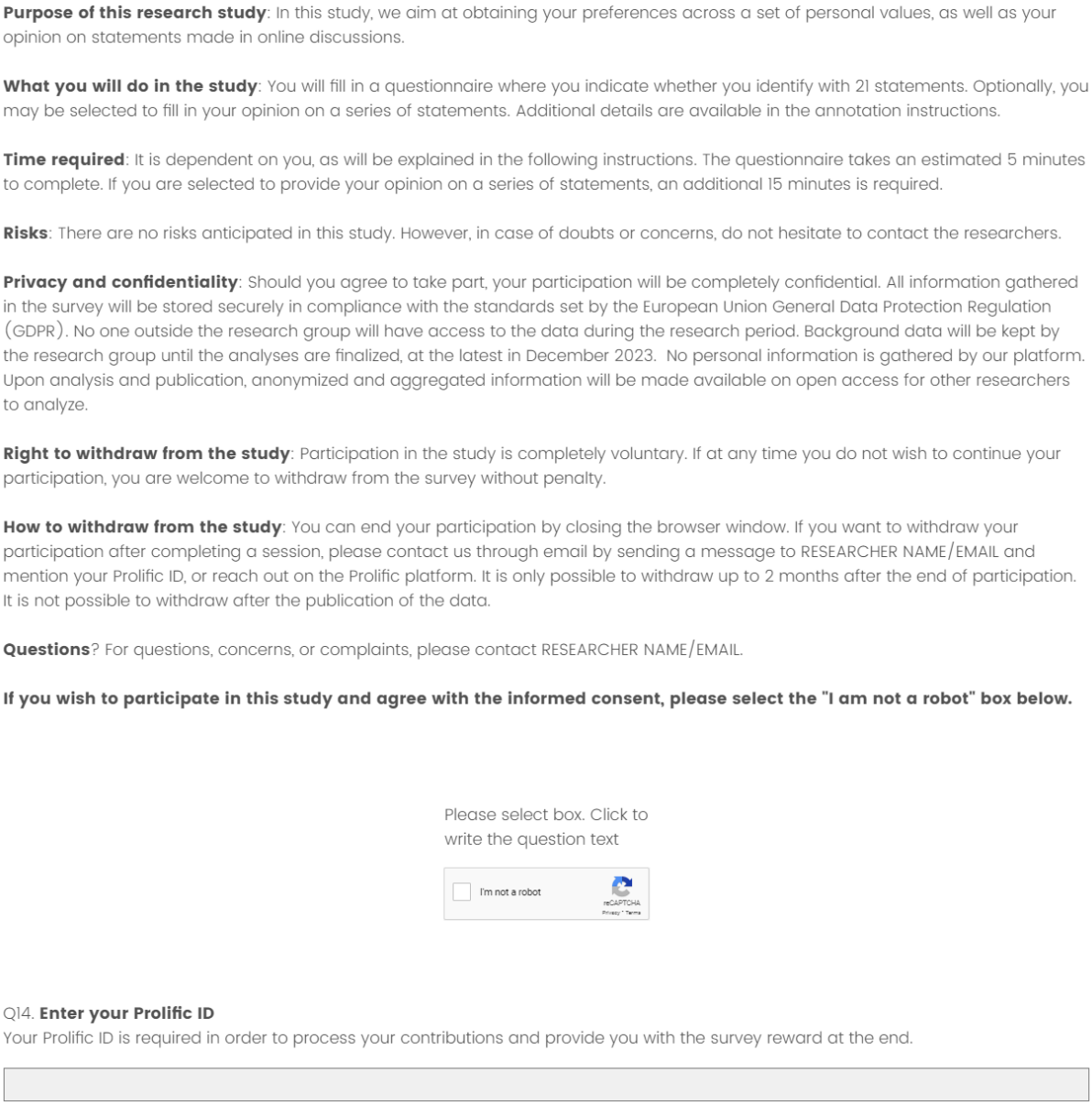


Figure 6: Informed consent shown to users before starting the experiment.

1. Gather Likert-scale answers on all 21 items.
2. Check if two attention check items were correctly answered. Participants were asked to fill in a given score. Disregard participant results otherwise.
3. Compute Mean Rating for each participant (MRAT).
4. Subtract the mean score from all other scores to obtain centered response scores.
5. Normalize the profile by dividing by the sum of all scores.

A.3 Training agreement analysis models

Training models for agreement analysis takes around 4 hours for the BERT models on the subsampled Debagreement dataset. See Table 7 for the hyperparameters used. Debagreement may be reused under the CC BY 4.0 license. For the implementation of the TF-IDF, we used the `sklearn` (Pedregosa et al., 2011) Python package. All training involving TF-IDF embeddings takes under 1 hour.

We constructed three types of extra user information for the agreement analysis task:

Random noise We sample a vector of size 768 from a random uniform distribution over $[0, 1)$.

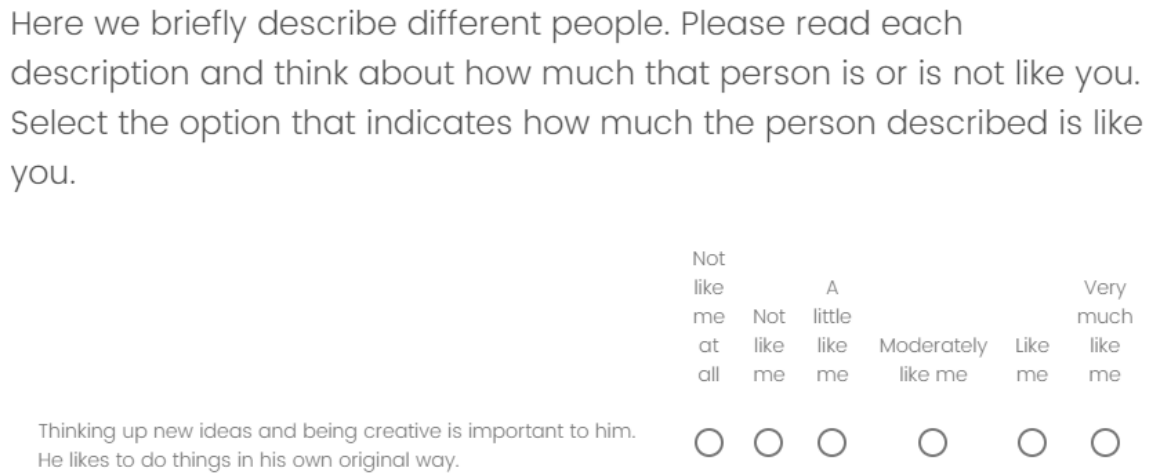


Figure 7: Screenshot of the PVQ-21 survey.

Hyperparameter	Value
`train epochs`	7
`learning rate`	$5e-05$
`model`	bert-base-uncased
`batch size`	64

Table 7: Hyperparameters used for training models for agreement analysis

User centroids We stem the posts from users that contain at least one value term according to the value dictionary and transform comments to TF-IDF vectors. We restrict the vocabulary to the 768 most frequent terms. We then compute the average over all vectors for a single user.

Explicit user features We construct user feature vectors for Reddit users through the Reddit PRAW API. See Table 8 for the features used.

Feature	Explanation
comment_karma	Total amount of upvotes minus downvotes on comments.
link_karma	Total amount of upvotes minus downvotes on link submissions.
date_created	Timestamp of account creation.
gold_status	Whether the user is a gold member.
mod_status	Whether the user is a mod of any subreddit.
employee_status	Whether the user is an employee of Reddit.
num_gilded	Number of gilded items.
num_comments	Number of comments posted by user.
num_links	Number of links submitted by user.

Table 8: Features used to represent a user from Reddit

B Additional Results

B.1 Value Extraction

For a complete overview of the performance of the value extraction models, including the standard deviation over 10 random seeds for the VE models, see Table 9.

B.2 Value Survey

Demographics We received a total of 27 responses, one of which was ignored because of a failed attention check. Different ages were represented in our sample (M=28.0, SD=8.7), and annotators originated from Europe (18 annotators), South Africa (8 annotators), the UK (1), and the US (1). About half (13) were registered students.

Reliability Since the PVQ has two questions for each personal value, we are able to compute internal consistency using Cronbach α per value. See the results in Table 10. We observe a wide range of reliability scores, of which only conformity reaches above a score of 0.7. Most interestingly, we see that tradition is of very low reliability, possibly due to the demographic of some of our participants (students). Three task instances received mostly neutral or not-enough-information labels, and were disregarded in our analysis.

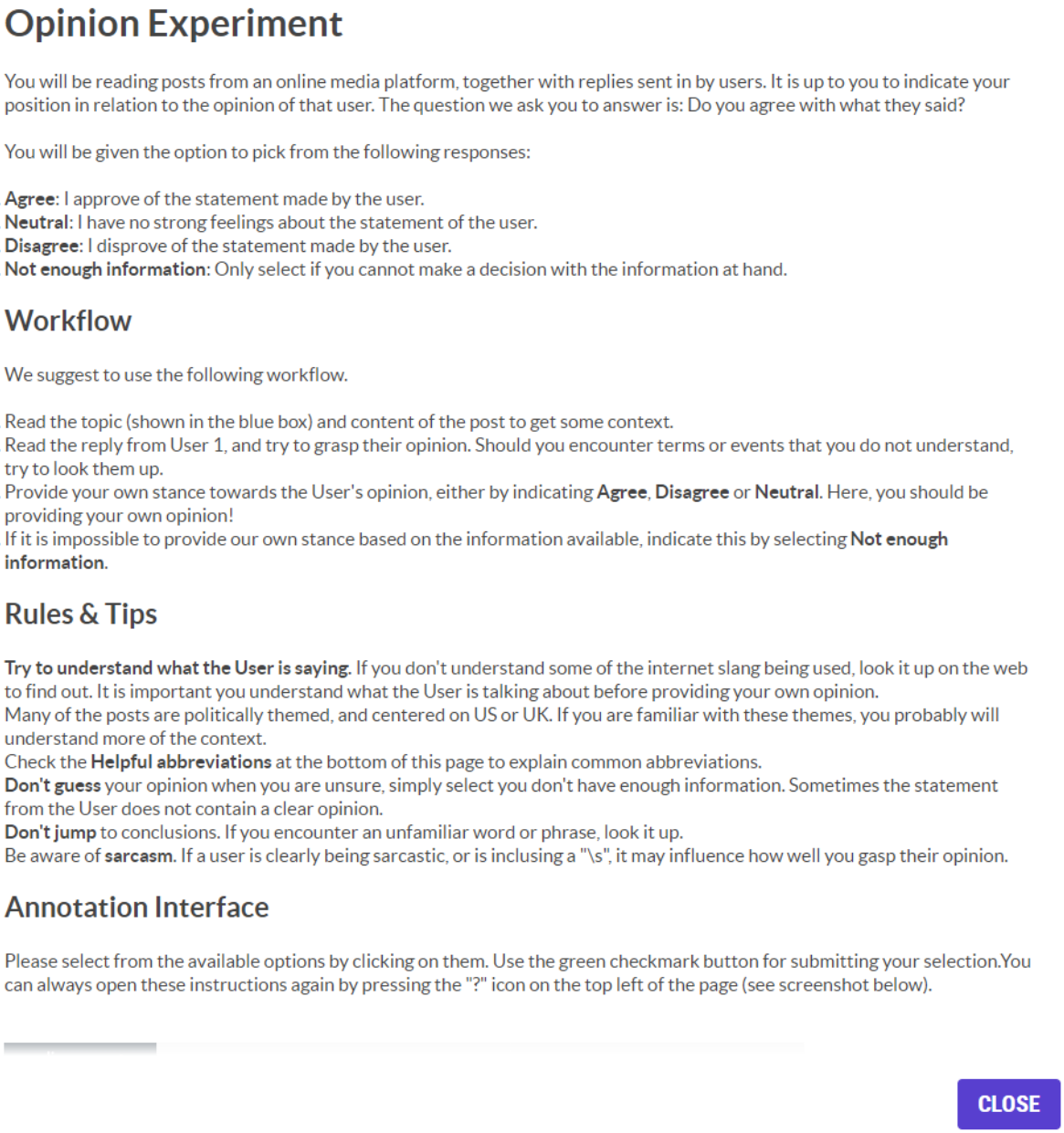


Figure 8: Instructions shown to users for the annotation experiment.

B.3 Qualitative Examples of Value Conflicts and (Dis-)agreement

We perform a qualitative analysis of some instances (comment pairs) from the dataset that follow our hypothesis and some that do not to gain a better understanding of when value conflicts influence disagreement. Table 11 shows examples of the types of pairs we analyze.

B.4 Decomposition of BF_{10} results

We create overviews of the different tests performed in Sections 4.3.1 and 4.3.2. We decompose the aggregated scores into three separate figures, each showing how a single variable (either subreddit, similarity score, or profile threshold) impacts the obtained results. We show the decomposition for the BF_{10} scores obtained for comparisons between two VPE-estimated profiles in Figures 10 and for the comparison between VPE and self-reports in Figure 11. In the latter case, since we picked samples from Debagreement with authors with populated value profiles, we do not need to test over multiple profile thresholds.

We show the highest and lowest BF_{10} scores and the test parameters in Tables 12 and 13 between two VPE profiles, and in Tables 14 and 15 for the

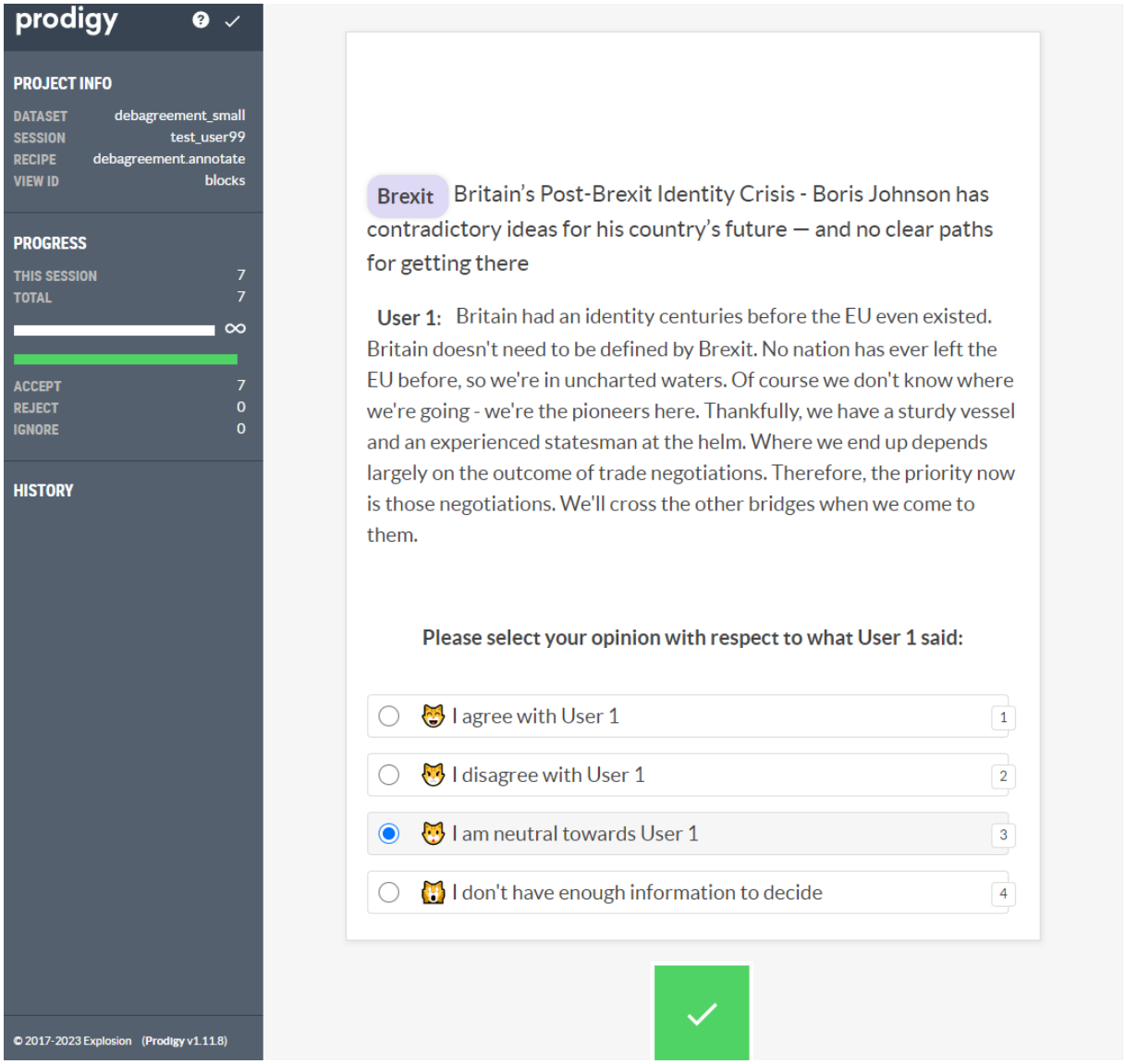


Figure 9: Annotation interface.

Method	Training data	P(VN)	R(VN)	F1(VN)	P(VA)	R(VA)	F1(VA)	F1
All-ones	–	0.34	0.50	0.40	0.11	0.50	0.18	0.26
VD	–	0.56	0.55	0.45	0.64	0.58	0.59	0.57
(Kiesel et al., 2022)*	*VA*	0.20	0.21	0.15	0.47	0.34	0.37	0.28
(Qiu et al., 2021)*	*VN*	0.64	0.65	0.59	0.53	0.52	0.52	0.57
BERT	*VN*	$0.66_{\pm 0.00}$	$0.68_{\pm 0.00}$	$0.66_{\pm 0.00}$	$0.57_{\pm 0.02}$	$0.60_{\pm 0.02}$	$0.57_{\pm 0.03}$	$0.65_{\pm 0.02}$
	VA	$0.57_{\pm 0.00}$	$0.56_{\pm 0.00}$	$0.46_{\pm 0.00}$	$0.79_{\pm 0.02}$	$0.74_{\pm 0.01}$	$0.76_{\pm 0.01}$	$0.67_{\pm 0.01}$
	Both	$0.63_{\pm 0.00}$	$0.64_{\pm 0.00}$	$0.63_{\pm 0.00}$	$0.84_{\pm 0.02}$	$0.79_{\pm 0.00}$	$0.81_{\pm 0.01}$	$0.79_{\pm 0.00}$
RoBERTa	*VN*	$0.61_{\pm 0.15}$	$0.66_{\pm 0.05}$	$0.62_{\pm 0.12}$	$0.58_{\pm 0.02}$	$0.61_{\pm 0.02}$	$0.59_{\pm 0.02}$	$0.63_{\pm 0.03}$
	VA	$0.57_{\pm 0.00}$	$0.56_{\pm 0.00}$	$0.46_{\pm 0.00}$	$0.79_{\pm 0.02}$	$0.74_{\pm 0.01}$	$0.76_{\pm 0.01}$	$0.67_{\pm 0.01}$
	Both	$0.63_{\pm 0.00}$	$0.64_{\pm 0.00}$	$0.63_{\pm 0.00}$	$0.83_{\pm 0.02}$	$0.78_{\pm 0.01}$	$0.80_{\pm 0.01}$	$0.78_{\pm 0.00}$

Table 9: Macro-averaged performance of the value estimation approaches on the value datasets, showing averages and standard deviation for our own models over 10 different seeds. *VN* denotes ValueNet, *VA* denotes ValueArg. Methods marked with * are trained on a different objective than our VE task.

experiments comparing VPE and self-reported profiles.

Value	α	95% CI
conformity	0.717	(0.514,0.835)
tradition	0.051	(-0.627,0.447)
benevolence	0.336	(-0.138,0.613)
universalism	0.407	(-0.016,0.654)
self-direction	0.641	(0.384,0.790)
stimulation	0.589	(0.295,0.760)
hedonism	0.618	(0.345,0.777)
achievement	0.504	(0.149,0.711)
power	0.371	(-0.078,0.633)
security	0.388	(-0.050,0.643)

Table 10: Internal consistency scores (Cronbach's α) for the values in the PVQ-21 questionnaire.

B.5 Kendall τ vs. Spearman ρ

We include a comparative overview of the tests that use the Kendall τ and add the BF_{10} scores for the same tests conducted with Spearman ρ. See Figure 12. We see that generally, the ρ scores are similarly distributed as the τ scores. Two tests that for τ fall into the undecidable range, for ρ favor the null hypothesis H_0. We attribute this to the size of our value profiles: since we have only 10 entries, ties are likely, and Spearman ρ does not explicitly account for them.

B.6 Agreement Analysis

For additional results (**P**recision, **R**ecall, F_1 scores, accuracy, and the change w.r.t. a text-only baseline), see Table 16.

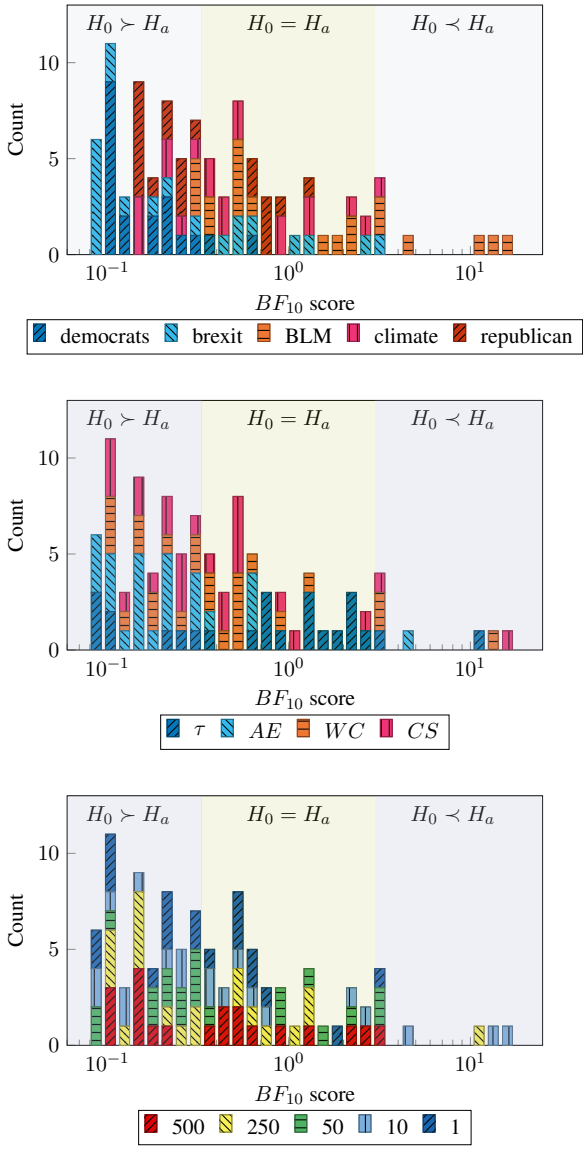
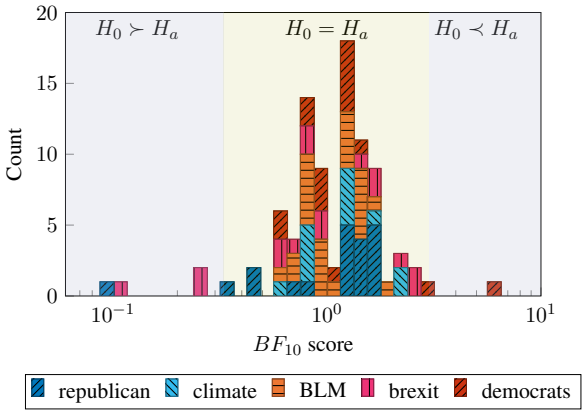


Figure 10: BF_{10} scores when testing between two VPE-constructed profiles, obtained for all combinations of subreddits (top figure), similarity scores (middle figure) and profile thresholds (bottom figure).

	Disagree	Agree
No Value Conflict	This is NOT a public statue. It's a privately owned statue on private property.. the government has zero right to take it down. Not so sure. A crime on private property is still a crime, and defending racism is a crime.	Climate justice has waited too long to be served. The time is now! Guys, get out there and support people, politicians, businesses, companies, and local stores who support climate justice and sustained efforts to promote sustainability and eco-friendliness alike!!
Value Conflict	The EU moves very slowly.. Don't blame the UK if the EU is so slow. So you're saying the EU should make the UK its priority? Why should the UK have priority over another issue?	Brexit is a symptom, not a problem in itself. Don't just make the symptom go away, treat the many underlying problems first I agree, but you have a parliament that took control from May then did the dumbest thing it could do by not voting for any of the proposals.

Table 11: Confusion matrix of qualitative examples of the match between value conflict and (dis-)agreement.



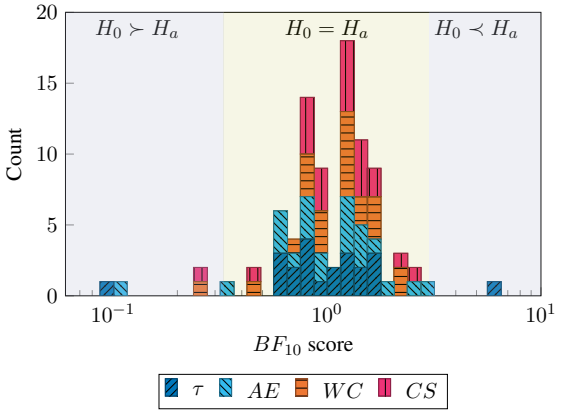


Figure 11: BF_{10} scores when testing between a VPE-constructed profile and a self-reported profile, split into different subreddits (top figure) and different similarity scores (bottom figure).

BF_{10}	Subreddit	Similarity score	Profile threshold
17.451	BLM	CO	10
12.485	BLM	WC	10
10.504	BLM	τ	250
4.223	BLM	MD	10
3.442	Brexit	WC	500

Table 12: The five tests between two VPE-constructed profiles with the highest BF_{10} scores.

BF_{10}	Subreddit	Similarity score	Profile threshold
0.079	Brexit	MD	50
0.081	Brexit	τ	50
0.083	Brexit	τ	10
0.085	Brexit	τ	1
0.086	Brexit	MD	10

Table 13: The five tests between two VPE-constructed profiles with the lowest BF_{10} scores.

BF_{10}	**Subreddit**	**Similarity score**
6.490	democrats	τ
3.066	democrats	MD
2.543	Brexit	MD
2.407	Brexit	CO
2.230	climate	CO

Table 14: The five tests between a VPE-constructed profile and a self-reported profile with the highest BF_{10} scores.

BF_{10}	**Subreddit**	**Similarity score**
0.087	republican	τ
0.108	Brexit	MD
0.247	Brexit	CO
0.273	Brexit	WC
0.359	repulican	MD

Table 15: The five tests between a VPE-constructed profile and a self-reported profile with the highest BF_{10} scores.

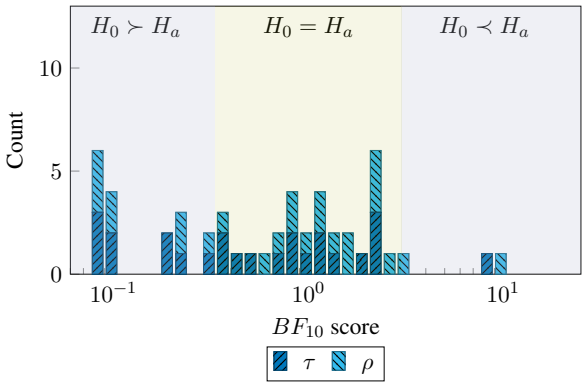


Figure 12: BF_{10} scores when testing between two VPE-constructed profiles, obtained for the similarity scores Kendall τ and Spearman ρ.

Model	**P**	**R**	**F1**	**Acc.**	**Δ F1**
Majority	0.12	0.33	0.18	0.37	
Only context (ϵ)	$0.21_{\pm 0.10}$	$0.34_{\pm 0.01}$	$0.24_{\pm 0.07}$	$0.36_{\pm 0.00}$	
Only context (z)	$0.42_{\pm 0.00}$	$0.41_{\pm 0.00}$	$0.41_{\pm 0.00}$	$0.43_{\pm 0.00}$	
Only context (u)	$0.33_{\pm 0.01}$	$0.35_{\pm 0.00}$	$0.31_{\pm 0.00}$	$0.38_{\pm 0.00}$	
Only context (v)	$0.27_{\pm 0.00}$	$0.37_{\pm 0.00}$	$0.31_{\pm 0.00}$	$0.40_{\pm 0.00}$	
TF-IDF + Logistic Regression	$0.48_{\pm 0.01}$	$0.47_{\pm 0.02}$	$0.46_{\pm 0.03}$	$0.48_{\pm 0.01}$	–
+ ϵ	$0.38_{\pm 0.01}$	$0.37_{\pm 0.01}$	$0.33_{\pm 0.05}$	$0.36_{\pm 0.03}$	-0.12
+ z	$0.51_{\pm 0.02}$	$0.47_{\pm 0.04}$	$0.43_{\pm 0.09}$	$0.45_{\pm 0.06}$	-0.03
+ u	$0.37_{\pm 0.00}$	$0.36_{\pm 0.00}$	$0.36_{\pm 0.01}$	$0.36_{\pm 0.01}$	-0.12
+ v	$0.51_{\pm 0.01}$	$0.45_{\pm 0.02}$	$0.41_{\pm 0.05}$	$0.45_{\pm 0.04}$	-0.04
BERT(-base-uncased)	$0.62_{\pm 0.00}$	$0.62_{\pm 0.01}$	$0.62_{\pm 0.01}$	$0.63_{\pm 0.01}$	–
+ ϵ	$0.63_{\pm 0.00}$	$0.62_{\pm 0.00}$	$0.62_{\pm 0.00}$	$0.64_{\pm 0.00}$	0.00
+ z	$0.63_{\pm 0.00}$	$0.63_{\pm 0.00}$	$0.63_{\pm 0.00}$	$0.63_{\pm 0.00}$	0.01
+ u	$0.62_{\pm 0.00}$	$0.62_{\pm 0.01}$	$0.62_{\pm 0.01}$	$0.63_{\pm 0.00}$	0.00
+ v	$0.64_{\pm 0.01}$	$0.64_{\pm 0.01}$	$0.64_{\pm 0.01}$	$0.65_{\pm 0.01}$	0.02

Table 16: Performance of the agreement classification on a subset of Debagreement (sentence pairs for which both users were available on Reddit).