Opinion diversity through hybrid intelligence

Meer, M.T. van der

6

Do Differences in Values Influence Disagreements in Online Discussions?

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.*

Michiel van der Meer, Piek Vossen, Catholijn M. Jonker, and Pradeep K. Murukannaiah. 2023. Do Differences in Values Influence Disagreements in Online Discussions? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15986–16008, Singapore. Association for Computational Linguistics.

6.1 Introduction

A large number of users participate in online deliberations on societal issues such as climate change [45] and vaccination hesitancy [428]. Disagreement is an important aspect of a deliberation [303] since it can (1) drive novel ideas, (2) incentivize evaluation of the proposed ideas, (3) avoid echo chambers, and (4) cancel out individual biases [204]. Discussions with disagreement help users understand the opposing viewpoints [234, 335]. Further, discussions having adequate disagreement have been associated with a higher quality deliberation [119]. 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 [372]. Further, questions arise on how to control for disagreement to enhance reciprocity [117], and how too much exposure to opposing views drives polarization [27]. Analysis methods for online discussions currently cannot accurately represent such diverse perspectives [61, 398], and measuring deliberative quality is an open challenge [352, 414].

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 [371]. In collaborative teams, value conflicts are linked to disagreement [182]. Specifically, values are said to be an effective way to make conflict explicit among participants in a discussion [41]. 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 [344]. 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 6.1 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 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 [305]. Since the contextual implications of values are usually unknown, connecting user concerns to values [11] opens up human-machine collaboration opportunities for a more constructive conversation [5, 158, 238].

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.

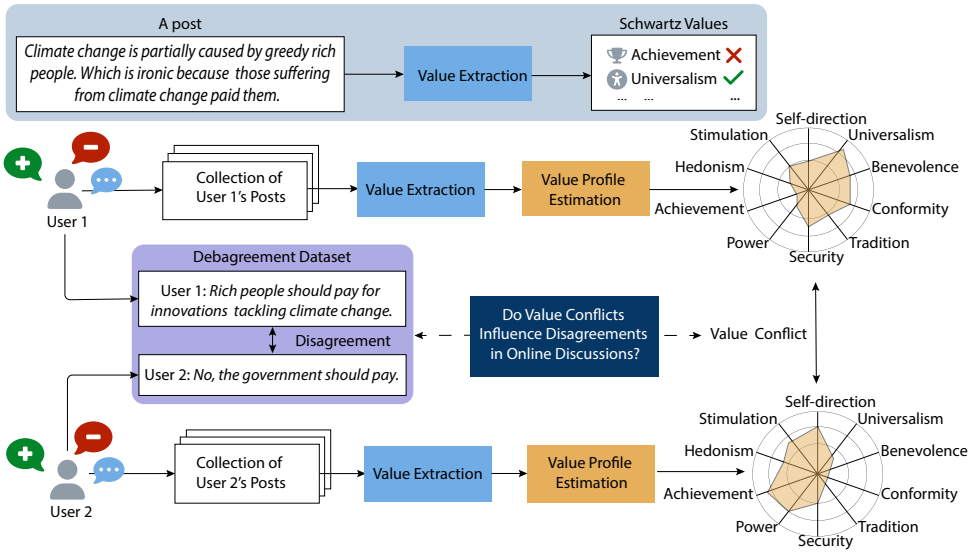

Figure 6.1: Setup of measuring value conflicts by means of Value Profile Estimation (VPE).

6

6.2 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.

6.2.1 (Dis)-agreement and discussion analysis

Detecting whether people agree or disagree with given statements is commonly framed as stance classification [e.g., 7]. Recently, more effort has been put into exploring various aspects of the task [9, 161, 246]. 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 [305].

Existing work on agreement analysis is aimed at (1) identifying language that indicates disagreement [e.g., 126, 284, 434], (2) leveraging stylistic choices like sarcasm for detecting disagreement [139], (3) finding stance and target pairs, followed by the traditional stance classification [e.g., 71, 96], and (4) mixing detailed opinion information using e.g., logic of evaluation [104]. Recently, adding social role context to textual comments was shown to have a positive impact on the agreement analysis task [255], 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 [344].

6.2.2 Value models

Values explain ideological beliefs underlying actions and opinions and may guide the design of applications [130]. Two leading value models have been used in NLP research: Moral Foundations [143] and the Schwartz Value model [344]. 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 6.1), 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.

6.2.3 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., 57]. However, a promising behavior-based approach focuses on analyzing textual motivations [70]. To this end, dictionary-based approaches can be used for finding value mentions in texts [141, 304]. Using such lexicons shows promising results in large-scale value estimation applications [356].

Recently, datasets annotated with personal values for training NLP methods have been released. In this chapter, we use two recent datasets annotated with Schwartz values: (1) ValueNet [311] is a dataset containing textual scenarios related to moral decision-making that have been annotated with relevant Schwartz values. (2) ValueArg [201] contains user-submitted arguments that relate to specific Schwartz values. There are some datasets on MFT values, e.g., [169, 253, 388]. 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 [311], review texts [288], and value-laden arguments [11, 207]. However, both the annotation and extraction of values remain difficult, with specific questions relating to the granularity of the value labels [201], their transfer to new domains [237], and how classifiers understand morality in language [239]. Moreover, large variances exist between the frequency of values across domain [196], and even the relevance of values differs depending on the domain [55, 235]. However, users can still be represented inside each domain by examining relative frequencies inside value profiles, as stipulated by Schwartz [344].

6.3 Method

Figure 6.1 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 [401] is available online.

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

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 6.1: List of subcorpora gathered in Debagreement.

6.3.1 Data

We use **Debagreement** [305] as the dataset containing (dis-)agreement labels. This dataset contains user-submitted post pairs in English from five topics (Table 6.1), 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 September 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 6.1 shows the amount of data collected after filtering.

6.3.2 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. [311]. Our training data comprises general texts annotated for the presence of values across multiple domains. We combine data from two sources.

(1) **ValueNet** [311]: 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** [201]: Their annotation scheme uses an updated (20) Schwartz values [345], 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.

6.3.3 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. [354].

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 [85], 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.

$$\mathcal{S}^{\tau}(v,w) = 1 - \frac{2 \times (\#\text{ discordant pairs})}{\binom{n}{2}} \tag{6.1}$$

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

$$\mathcal{S}^{MD}(v,w) = \sum_{i=1}^{n} |v_i - w_i| \tag{6.2}$$

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

$$\mathcal{S}^{CO}(v,w) = \frac{v \cdot w}{||v||\ ||w||} \tag{6.3}$$

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.

$$\mathcal{S}^{WC}(v,w) = \frac{\sum_{i=1}^{n} \mathcal{B}_i v_i w_i}{\sqrt{\sum_{i=1}^{n} \mathcal{B}_i v_i^2}\ \sqrt{\sum_{i=1}^{n} \mathcal{B}_i w_i^2}} \tag{6.4}$$

6.4 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.

6.4.1 Training Models for Value Estimation

We experiment with two popular BERT-based models, BERT [100] and RoBERTa [247], for value estimation. Further, we employ multiple baselines: (1) always predict all values for a

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. [201]*	ValueArg	0.15	0.37	0.28
Qiu et al. [311]*	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 6.2: Macro-averaged F_1 scores of the value estimation approaches on the value datasets. Methods marked with * are adapted for our comparison.

comment ("All-ones") to examine label imbalance, (2) predict values based on mentions of value words from the **Schwartz Value Dictionary** [304], (3) the multi-label approach from Kiesel et al. [201], which uses an expanded label set, and (4) the utility model from Qiu et al. [311]. The latter two baselines are BERT-based models. For Kiesel et al. [201], 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. [311] into binary value relevance labels at inference.

Table 6.2 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.

6.4.2 Value Profile Estimation

Table 6.3 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 6.3 (last column) shows the mean pairwise τ distance (Equation 6.1) among the value profiles in each domain. We observe that the BLM subcorpus has the most diversity among the five subcorpora.

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. [304]. Figure 6.2 shows a visualization of the first two dimensions after MDS. We observe that values that are close to each other in the Schwartz circumplex [344], e.g., achievement and power, also tend to be closer in the MDS visualization.

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

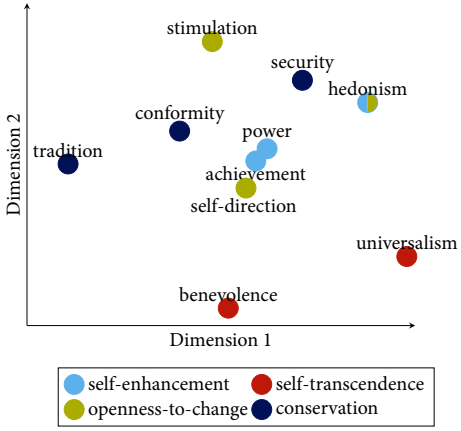

Figure 6.2: Visualization of the covariance between values in estimated profiles.

6.4.3 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 [23]. 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 [195].

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.

[2] BF hypothesis tests are sensitive to the choice of prior. We use the implementation of `pingouin` [395], which includes a Jeffreys-Zellner-Siow prior, an objective prior for two-sample cases [324]

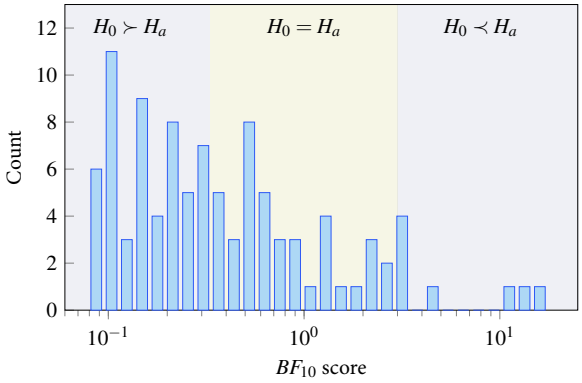

Figure 6.3: BF_{10} scores obtained for the combinations of data, value estimation methods, and scoring metrics.

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 6.3.2. 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^+:

$$\theta_G = \frac{1}{|G|} \sum_{\{p,c\} \in G} \mathcal{S}(p,c) \tag{6.5}$$

$$H_0 : \theta_{G^-} = \theta_{G^+} \tag{6.6}$$

$$H_a : \theta_{G^-} < \theta_{G^+} \tag{6.7}$$

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 6.3 provides an overview of the BF_{10} scores.

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 6.4 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 6.3), 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

[3]Higher for the **MD** metric, which flips the sign in Eqn. 6.7.

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

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.

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 setup is identical to Section 6.4.3, 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, 343], 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 6.4. 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 6.5). 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 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 [..]*

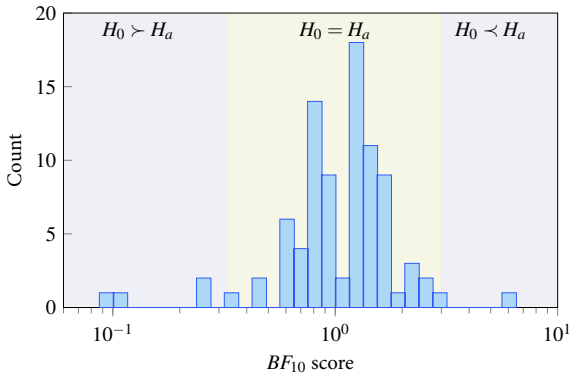

Figure 6.4: BF_{10} scores for all similarity scores and task instances comparing VPE and self-reported profiles.

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

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.

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 E.6 in Appendix E.2 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).

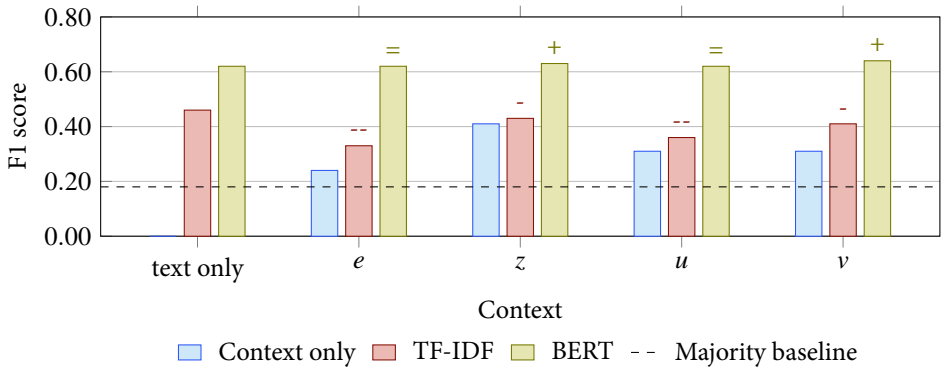

Figure 6.5: F_1 scores when adding extra context information. Symbols above bars show changes with respect to text-only: $--$ for $\Delta F_1 < -0.1$; $-$ for $-0.1 < \Delta F_1 < 0$; $=$ for $\Delta F_1 = 0$; and $+$ for $\Delta F_1 > 0$.

6.4.4 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. [305], 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., [74, 184]) 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 [151]. 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 [361].

Figure 6.5 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 [305]. 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 user-specific information is surprisingly difficult [6]. 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 [255] may make more effective use of the profile information.

6.5 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 [268].

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 [202]. 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., 122] or techniques for mixing in general-purpose language models [e.g., 399].

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 [211]. However, BF and HDI interpretations can be seen as complementary, respectively

quantifying evidence or beliefs [410].

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.

Ethical Considerations

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 [308]. 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 [265]. 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.