Universiteit
Leiden
The Netherlands

Opinion diversity through hybrid intelligence

Meer, M.T. van der

Citation
Meer, M. T. van der. (2025, March 26). *Opinion diversity through hybrid intelligence*. *SIKS Dissertation Series*. Retrieved from https://hdl.handle.net/1887/4209024

Version:	Publisher's Version
License:	Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden
Downloaded from:	https://hdl.handle.net/1887/4209024

Note: To cite this publication please use the final published version (if applicable).

V

Appendices

A

An Empirical Analysis of Diversity in Argument Summarization

A.1 Detailed Experimental Setup

We describe our experimental setup, starting with the data we use for conducting our analysis. We follow with a detailed description of each approach and finally present a description of the metrics used.

A.1.1 Data

Dataset	Num. arguments	Num. Key Points	Num. claims	Avg. arguments per claim	Avg. arguments per KP
ArgKP	10717	277	31	245	20
PVE	269	185	3	67	4
Perspectrum	10927	3804	905	12	3

Table A.1: Quantitative statistics of the datasets used in the experiments.

We provide some quantitative statistics on the three datasets used in our work in Table A.1. In addition, we show some qualitative examples of the content in our datasets in Table A.3. Since Perspectrum and ArgKP listed the same debate platforms as sources, we investigate the overlap between the claims and arguments between pairs of datasets. In terms of claims, there is no direct overlap between any two datasets. To rule out that the same arguments were scraped from the debate platforms, we also measure n-gram overlap [78]. We show the overlap in unigrams, bigrams, and trigrams in Table A.2. The overlap scores report the ratio of n-grams from one dataset that is found in the other.

For PVE, since the key point analysis was performed using a mixture of crowd and AI techniques, we take only the correctly matched key point–motivation pairs. That is, we take only those pairs that were deemed matching according to the final evaluation performed.

A

		Target		
		ArgKP	PVE	Perspectrum
Source	ArgKP	–	0.40/0.08/0.01	0.70/0.21/0.14
	PVE	0.41/0.16/0.06	–	0.66/0.24/0.10
	Perspectrum	0.17/0.04/0.02	0.22/0.03/0.01	–

Table A.2: Maximum uni-/bi-/trigram overlap between datasets.

Dataset	Claim	Key Point	Argument
ArgKP	We should subsidize journalism	Journalism is important to information-spreading/accountability.	Journalism should be subsidized because democracy can only function if the electorate is well informed.
PVE	Young people may come together in small groups	Young people are at low risk of getting infected with COVID-19 and therefore can benefit from gathering together with limited risk and potential profit.	Risks of contamination or transfer have so far been found to be much smaller.
Perspectrum	The threat of Climate Change is exaggerated	Overwhelming scientific consensus says human activity is primarily responsible for global climate change.	The biggest collection of specialist scientists in the world says that the world's climate is changing as a result of human activity. The scientific community almost unanimously agrees that man-caused global warming is a severe threat, and the evidence is stacking.

Table A.3: Qualtitative examples of claims, key points, and arguments across our dataset.

A.1.2 Per-approach Specifics

See Table A.4 for the language models used in each approach. We further outline any details depending on the approach used.

Debater The Debater API allows multiple parameters when running the KPA analysis. We manually tuned the parameters separately for KPG and KPM. For both tasks, we started with the most permissive configuration to optimize for recall first, and gradually made parameters more strict to improve precision without lowering recall scores. Once recall scores started dropping, we fixed the parameters. The final configuration is shown in Table A.5.

ChatGPT We strive to make our results as reproducible as possible, but due to the nature of the OpenAI API results may be specific to model availability. We conducted the experiments between July and August 2023, using the `gpt-3.5-turbo` and `gpt-3.5-turbo-16k`

Approach name	Model
ChatGPT	`gpt-3.5-turbo-16k`
ChatGPT (closed book)	`gpt-3.5-turbo`
Debater	*closed-source*
SMatchToPR (base)	`RoBERTa-base`
SMatchToPR (large)	`RoBERTa-large`

Table A.4: Models used for each KPA approach. Model choice is independent of subtask.

Subtask	Parameter	Value
KPG	mapping_policy	*LOOSE*
	kp_granularity	*FINE*
	kp_relative_aq_threshold	0.5
	kp_min_len	0
	kp_max_len	100
	kp_min_kp_quality	0.5
KPM	min_matches_per_kp	0
	mapping_policy	*LOOSE*

Table A.5: API Configuration for Debater approach.

models. We provide a template for the prompts below, in Prompts 1, 2, and 3. Open-book ChatGPT for KPG uses up to $B_{KPG} = 600, 100, 100$ for ArgKP, PVE, Perspectrum respectively. ChatGPT uses a batch size of $B_{KPM} = 10$ when making match predictions for KPM. Interpreting the responses was done by prompting the model to output valid JSON, and writing a script that parses the generated response. Invalid JSON responses are considered errors on the model's side, resulting in an empty string for KPG and a 'no-match' label for KPM. In order to cut down on costs, we subsampled the test set for Perspectrum, taking a random 15% of the claims in order to drive down the costs further.

Prompt 1: ChatGPT closed book, KPG prompt

Give me a JSON object of key arguments for and against the claim: {**claim**}. Make sure the reasons start with addressing the main point. Indicate per reason whether it supports (pro) or opposes (con) the claim. Rank all reasons from most to least popular. Make sure you generate a valid JSON object. The object should contain a list of dicts containing fields: 'reason' (str), 'popularity' (int), and 'stance' (str).

A

Prompt 2: ChatGPT open book, KPG prompt

Extract key arguments for and against the claim: **{claim}**. You need to extract the key arguments from the comments listed here: **{up to B_{KPG} arguments}** Give me a JSON object of key arguments for and against the claim. Make sure the reasons start with addressing the main point. Indicate per reason whether it supports (pro) or opposes (con) the claim. Rank all reasons from most to least popular. Make sure you generate a valid JSON object. The object should contain a list of dicts containing fields: 'reason' (str), 'popularity' (int), and 'stance' (str).

Prompt 3: ChatGPT open book, KPM prompt

For the claim of **{claim}**, indicate for each of the following argument/key point pairs whether the argument matches the key point. Return a JSON object with just a "match" boolean per argument/key point pair.
ID: **{pair id}** Argument: **{argument}** Key point: **{key point}** (*up to B_{KPM} times*) …

SMatchToPR We preprocess the PERSPECTRUM dataset analogously to the ARGKP dataset. We train the SMatchToPR model using contrastive loss for 10 epochs and a batch size of 32. The training has a warmup phase of the first 10% of data. The base and large variants use the same parameters. See Table A.6 for the hyperparameters when executing KPG and KPM. The computing infrastructure used contained two RTX3090 Ti GPUs. Training the RoBERTa large variant takes around 30 minutes.

Parameter	Value
PR d	0.2
PR min quality score	0.8
PR min match score	0.8
PR min length	5
PR max length	20
filter min match score	0.5
filter min result length	5
filter timeout	1000

Table A.6: Hyperparameters for SMatchToPR approach. Parameters are independent of subtask.

A.1.3 Evaluation metrics

For Key Point Generation, we resort to measuring lexical overlap and semantic similarity. To make our results reproducible we provide further details on the configuration of the ROUGE metrics [150]. Our evaluation uses the `sacrerouge` package that wraps the original ROUGE implementation[1]. The full evaluation parameters can be seen in Table A.7.

Furthermore, we use two learned metrics (BLEURT and BARTScore) to report the semantic similarity of generated key points and reference key points. For BLEURT, we use

[1] https://github.com/danieldeutsch/sacrerouge

Parameter	Value
Porter Stemmer	*yes*
Confidence Interval	95
Bootstrap samples	1000
α	0.5
Counting unit	*sentence*

Table A.7: Configuration parameters for the ROUGE evaluation of KPG.

the publicly available `BLEURT-20` model, which is a RemBERT [76] model trained on an augmented version of the WMT shared task data [257]. BARTScore uses a BART model trained on ParaBank2 [174].

A.2 Additional results

We present two additional results: we provide fine-grained ROUGE results for KPG, and provide examples of key points generated by ChatGPT.

A.2.1 Detailed ROUGE scores for Key Point Generation

Earlier, we provided aggregated F_1 scores for the KPG evaluation. Here, we also show Precision and Recall scores in Table A.8. We see that the models that perform best in terms of F_1 score are consistently scoring well in terms of precision and recall across all datasets. For instance, open-book ChatGPT performs best on ArgKP in terms of F_1 (see Table 2.3), achieving consistently high precision and recall scores. Other approaches may score higher on individual metrics (e.g. SMatchToPR large scores higher in terms of ROUGE-1 recall), but this pattern is not consistent across all metric types.

A.2.2 Additional BERTScores for Key Point Generation

Next to BLEURT and BARTScore, we report BERTScore [448] for the approaches in the KPG evaluation, to examine the relation between the various learned metrics. See Table A.9 for an overview.

A.2.3 Long-tail experiment for KPG

We perform the long-tail analysis for Key Point Generation, adopting the same cutoff parameter f from the KPM analysis. Figure A.1 shows the results when including a fraction of key points f, starting from the least frequent (i.e. the key points with the lowest amount of arguments matched to them). The figure shows that for a low fraction of data, all approaches perform considerably worse. Note that due to the evaluation setup in Li et al. [231], scores may be lower due to a smaller pool of key points. Since we report averages of the maximum scoring match between any given generated and reference key points, this smaller pool may lead to overall lower scores. We still report these results to show the impact of making the evaluation set smaller, next to focusing on infrequent opinions.

Data	Approach	Precision			Recall		
		R-1	**R-2**	**R-L**	**R-1**	**R-2**	**R-L**
ArgKP	ChatGPT	29.1	**10.6**	25.6	45.2	**16.1**	41.2
	ChatGPT (closed book)	**30.8**	6.8	**26.9**	32.0	8.6	27.3
	Debater	25.3	5.5	23.1	28.2	5.3	23.4
	SMatchToPR (base)	24.5	9.3	23.2	44.5	11.2	41.5
	SMatchToPR (large)	22.0	6.4	19.4	**53.0**	13.0	**47.5**
PVE	ChatGPT	25.1	6.4	21.1	19.1	3.9	15.8
	ChatGPT (closed book)	30.1	**9.8**	22.6	**26.4**	**8.1**	**21.6**
	Debater	**33.3**	0.0	**33.3**	13.3	7.1	13.3
	SMatchToPR (base)	28.8	5.6	22.6	18.0	2.9	14.4
	SMatchToPR (large)	27.8	5.6	22.6	18.0	2.9	14.4
Perspectrum	ChatGPT	17.5	4.7	14.8	**35.0**	**10.2**	**30.5**
	ChatGPT (closed book)	14.8	3.1	12.8	25.4	6.3	22.7
	Debater	8.6	0.4	7.6	25.5	6.3	22.7
	SMatchToPR (base)	18.8	5.5	15.9	32.0	9.2	27.8
	SMatchToPR (large)	**19.0**	**5.7**	**16.1**	32.3	9.8	28.3

Table A.8: ROUGE Precision and Recall scores for the Key Point Generation task.

A.2.4 ChatGPT generated key points for PVE

See Table A.10. A cursory search for the content of the open-book key points shows the key points are directly taken from arguments in PVE. While ChatGPT performs conditioned language generation, it behaves like extractive summarization when using the open-book approach for the arguments in PVE. This leads to potentially incomplete or subjective key points. For the closed-book approach, we observe that ChatGPT generates independent and objective key points.

Data	Approach	BERTScore Precision	Recall	F_1
ArgKP	ChatGPT	0.412	0.470	0.422
	ChatGPT (closed book)	0.322	0.336	0.324
	Debater	0.406	0.367	0.379
	SMatchToPR (base)	0.362	0.463	0.394
	SMatchToPR (large)	0.361	0.482	0.402
PVE	ChatGPT	0.184	0.157	0.153
	ChatGPT (closed book)	0.386	0.280	0.324
	Debater	0.523	0.146	0.301
	SMatchToPR (base)	0.339	0.210	0.257
	SMatchToPR (large)	0.339	0.210	0.257
Perspectrum	ChatGPT	0.208	0.308	0.252
	ChatGPT (closed book)	0.244	0.274	0.243
	Debater	0.228	0.274	0.246
	SMatchToPR (base)	0.231	0.297	0.258
	SMatchToPR (large)	0.235	0.296	0.260

Table A.9: BERTScore Precision, Recall, and F_1 scores for the Key Point Generation task.

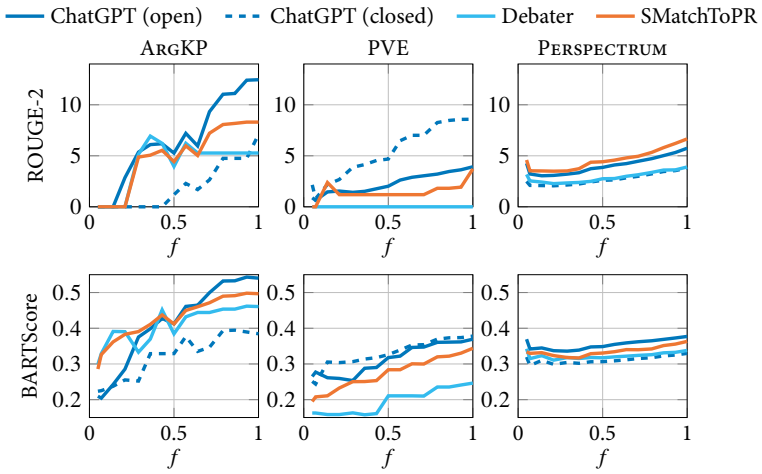


Figure A.1: KPG performance when limiting data usage to a fraction f, starting with the long tail first.

Claim	Stance	KP (open-book)	KP (closed-book)
All restrictions are lifted for persons who are immune	con	The coronavirus is an assassin, let's really learn more about this first	There may still be unknown long-term effects of the virus, even in those who have recovered.
Re-open hospitality and entertainment industry	pro	Economy needs to start running again	Reopening the hospitality and entertainment industry will help stimulate the economy and create job opportunities.
Young people may come together in small groups	con	The spread will then come back in all its intensity.	Small group gatherings may pose a risk of spreading contagious diseases.

Table A.10: Examples of generated key points from the open-book and closed-book ChatGPT approach.

B

Will It Blend? Mixing Training Paradigms & Prompting for Argument Quality Prediction

B.1 Hyperparameters

GPT-3 Prompt We used the model `text-davinci-002` with a temperature of 0 and no penalties on frequency and presence. We experimented with various prompt designs (e.g. dynamic or longer examples, more/fewer examples, joint prompting of novelty and validity) but manual inspection showed the best results for the setup described in Chapter 3 (i.e. separate prompts, static prompt style).

Transformers We report the hyperparameters for each approach in Table B.1 that differ from the default. In all Transformer models, we used the AdamW optimizer [252].

Model	LR	epochs	g.acc.
CLTeamL-2	1e-05	9	1
CLTeamL-3 (novelty)	1e-05	9	1
CLTeamL-4	5e-06	6	4
CLTeamL-5 (novelty)	5e-06	6	4

Table B.1: Hyperparameters for our approaches that involve gradient-based learning.

SVM The best performing model on the validation set is one with a C parameter of 0.09 for validity and 4.7 for novelty. The text representation concatenates the two texts, in a TF-IDF and stemmed (with the SnowBall stemmer as implemented in NLTK) representation.

	Prec.	Rec.	F1	Support
non-valid	0.732	0.636	0.681	179
valid	0.780	0.847	0.812	341
non-novel	0.563	0.806	0.663	421
novel	0.424	0.186	0.259	99

Table B.2: Performance statistics for approach *CLTeamL-1*.

	Prec.	Rec.	F1	Support
non-valid	0.364	0.806	0.502	93
valid	0.943	0.693	0.799	427
non-novel	0.901	0.646	0.753	410
novel	0.358	0.736	0.482	110

Table B.3: Performance statistics for approach *CLTeamL-2*.

B.2 Additional results

For every analysis, we show the results for approaches *CLTeamL-1* and *CLTeamL-2*, which can be combined into *CLTeamL-3* by merging their results (take validity and novelty, respectively for *1* and *2*).

B.2.1 Per-label Performance

See Tables B.2 and B.3.

B.2.2 Label confusion

See Tables 3.4 and B.4.

B.2.3 Seed Variance

While the results for the task were obtained using a single model, we investigate training stability over multiple seeds. We show the results and variance from five different seeds for our best-performing MTL model. The results can be seen in Figure B.1. Training is relatively stable, but individual models may have small performance differences on the test set.

		Predicted	
		-	+
True	-	131	75
	+	48	266

(a) GPT-3

		Predicted	
		-	+
True	-	75	131
	+	18	296

(b) MTL

Table B.4: Confusion matrices for the validity labels.

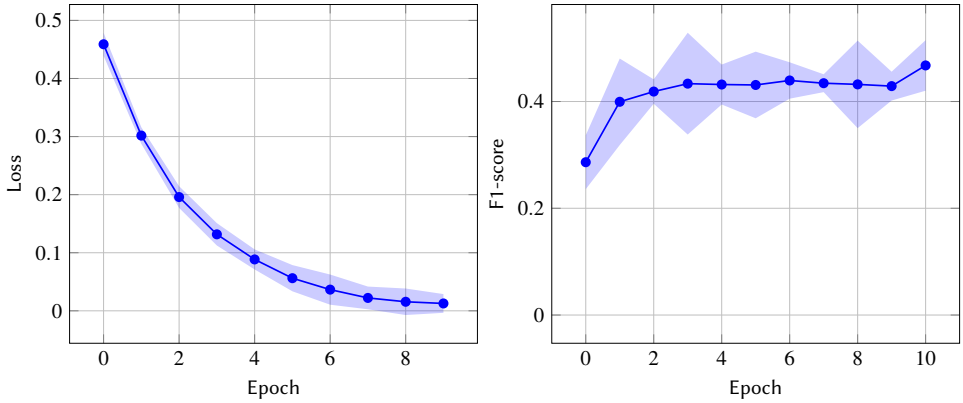


Figure B.1: Training loss and combined F1 score for multiple training runs of *CLTeamL-2* with different seeds.

B.2.4 Topics

The three most error-prone topics were different for approaches. Notable is that "Vegetarianism" is an error-prone topic across tasks and approaches.

GPT-3 - Validity "Was the Iraq War Worth it?" (unseen) with 44.8% errors, "Year Round School" (unseen), 39.7% errors, and "Withdrawing from Iraq" (unseen), 38.1% errors.

GPT-3 - Novelty "Yucca Mountain nuclear waste" (62.5% error rate), "Vegetarianism" (60% error rate), "Wiretapping in the U.S. (59.2% error rate).

MTL - Validity "Zero Tolerance Law" (42.1%), "Vegetarianism" (40% error rate) and "Yucca Mountain nuclear waste" (37.5% error rate).

MTL - Novelty "Withdrawing from Iraq" (44.7% error rate), "Vegetarianism" (44% error rate), "Wiretapping in the United States" (44% error rate)

Topics not in dev, only in test 'Video games', "Zero tolerance law', "Was the War in Iraq worth it?', "Withdrawing from Iraq', "Year-round school', "Veal', "Water privatization'.

C

A Hybrid Intelligence Method for Argument Mining

C.1 Experiment Protocol & Description

In order to reproduce the experiments performed in this research, we provide a complete overview of the guidelines, preliminaries, data, and technical artifacts created. This overview contains additional information about how the experiments were conducted. The texts presented to the annotators, such as the informed consent, the annotation introduction, and instructions are provided in the supplementary material as well. In addition, we provide details on the average run times per experiment, as well as any other auxiliary details here.

C.1.1 Preliminaries

Before starting the experiments, annotators were required to familiarize themselves with the annotation procedure and web interface. Upon entering the web platform, they were provided with an informed consent form and instructions for their task. The instructions consist of a short introduction to the context of the task, followed by detailed instructions about the components they would be annotating (opinions, arguments, topics, etc.). In addition, they were provided example annotations, both in writing and by means of a video.

After having seen all these, annotators were asked to fill in a short exercise annotation. This exercise consisted of 3 or 4 items, applicable to a hypothetical policy option, each with a predefined correct answer. Annotators were required to get the answers correct but had unlimited tries to perform the exercise. Completing the exercise enabled the actual annotation task, which in all cases was upper-bounded by a fixed number of items. Annotators were paid 7,50 per hour which is considered an ethical monetary reward on Prolific.

C.1.2 Phase 1: Argument Annotation

This first phase of HyEnA consists of three stages. We provide some additional details per stage. For the interpretation of the results, we refer to Chapter 4.

Argument Annotation Five annotators were given one hour to explore 51 opinions from the corpus for a single option. On average, they took 44, 31, and 43 minutes respectively for the options of YOUNG, IMMUNE and REOPEN.

Topic Generation Two experts worked to generate a short list of topics from the 15 most frequent BERTopic generated topics, with the short list containing only coherent and unique topics. Two experts worked for 23 minutes on average to rate all topics across all three options.

Topic Assignment In the topic assignment, each argument from the **argument annotation** stage had to be provided with a manual topic assignment. Topics are assigned by five overlapping annotators. For YOUNG, IMMUNE and REOPEN, they took 26, 30, and 33 minutes respectively on average.

C.1.3 Phase 2: Argument Consolidation

The arguments were consolidated by 99, 57, and 87 annotators for the options of YOUNG, IMMUNE and REOPEN respectively. The median completion time was 20, 20 and 18 minutes. In the Multi Path algorithm in use by POWER multiple annotators are able to work in parallel, supported by our annotation platform.

C.1.4 Comparison to Automated Baseline

Lastly, in the comparison between HyEnA and ArgKP, annotators rated a fixed number of opinions and arguments. For the option YOUNG, 28 annotators took 23 minutes on average. For both IMMUNE and REOPEN, both options saw 21 annotators, which took 25 and 23 minutes on average respectively. In this task, the annotators were asked to assess the match between arguments and opinions, where *matching* is defined as "an argument capturing the gist of the opinion, or directly supports a point made in the opinion."

C.1.5 Annotation platform

We run the HyEnA experiments by employing workers from Prolific (`www.prolific.co`). The support our experiments, we created our own web platform for the phases in HyEnA. The platform allows annotators to work in parallel and is equipped with control mechanisms for conducting the experiments. Furthermore, we run an evaluation study on the Prodigy annotation platform (`https://prodi.gy/`).

Where possible, computations are performed offline, which is possible for all phases with the exception of the Parallel Pairwise Annotation method, POWER. For this phase, we precomputed the dependency graph G, and extracted the disjoint paths containing the pairs to be annotated. Following the annotator's decisions, we then make automated judgements over sections of these paths. We add screenshots of the pages as presented to the annotators in the `screenshots/` directory.

The ArgKP baseline was run using two RTX 3090 Ti GPUs, which took around 30 hours per opinion corpus. For HyEnA, the opinion corpus was transformed into embeddings using the same device within 4 hours. Training the BERTopic models took less than an hour. All web-based experiments were hosted on a single server with 16GB RAM, without access to a GPU.

C.2 Method Details

C.2.1 Parallel Pairwise Annotation Algorithm

To accommodate annotators performing asynchronous annotation, we take an incremental procedure for pairwise annotation. As soon as a pair has seen three annotations, the automatic labeling procedure is run, and the next pair to be annotated in the same path is opened up for annotation. When all pairs are (either manually or automatically) labeled, the algorithm is complete. See Algorithm 3 for computational description of the parallel pairwise annotation algorithm [67]. Since the paths are annotated through a binary traversal method, we can also obtain an upper bound of number of annotations required, which is the number of paths $|P|$ multiplied by the maximum number of annotations required for the longest path g, $P \times \lceil \log_2(|\, g \,|) \rceil$.

Algorithm 3: Parallel Pairwise annotation

Input: Dependency graph $G = \{V, E\}$
Output: Labeled vertices V
$B =$ create bipartite graph (G)
$Y =$ find maximal matching (B)
$P =$ find disjoint paths (Y)
while *!fully labeled(G)* **do**
  for $p \in P$ **do**
    $v =$ find middle(p)
    label vertex(v) ; ▷ N humans
  end
  automatically label paths(P, label)
end

C.2.2 Hyperparameters

HyEnA

An overview of hyperparameters for HyEnA is given in Table C.2.

ArgKP

Table C.3 shows the hyperparameters for the ArgKP baseline. The hyperparameters for the ArgKP baseline were picked such that they are balanced between the ones used for the Argument dataset [33], but also would increase (up to ∼10%) the ratio of comments picked as key point candidates. While this is lower than the recommended 20%, we avoided relaxing the heuristic hyperparameters to prevent picking overly specific arguments as candidates. In Figure C.1, we show the ratio of number of candidates extracted out of all opinions depending on the hyperparameters.

Running ArgKP does not come cheap. The number of comparisons required to be made (forward passes through the matching model) is $\mathcal{O}(NM)$ where N is the number of candidates and M the number of opinions. Table C.1 shows the number of comparisons made by the model in use in our experiments.

Option	Stance	# Opinions	# Candidates	# Comparisons
YOUNG	pro	8804	1307	12M
YOUNG	con	4596	463	2M
IMMUNE	pro	1760	369	649K
IMMUNE	con	8807	657	6M
REOPEN	pro	7027	690	5M
REOPEN	con	5787	457	3M

Table C.1: Quantative descriptive information for running ArgKP.

Parameter	Option	Value	Description
M_{SBERT}	all	`paraphrase-MiniLM-L6-v2`	Model used to transform opinions and arguments into a numerical representation.
$\mathcal{T}$	all	`paraphrase-MiniLM-L6-v2`	Model in use by BERTopic.
f	all	5	Number of farthest opinions to sample using FFT.
clustering method	YOUNG	`louvain`	Clustering method used to extract argument clusters per option.
	IMMUNE	`louvain`	
	REOPEN	`spectral`	
r	YOUNG	0.449	Resolution parameter for Louvain clustering.
r	IMMUNE	0.449	Resolution parameter for Louvain clustering.
k	REOPEN	18	Number of desired clusters for spectral clustering.

Table C.2: Hyperparameters used by HyEnA.

Parameter	Value	Baseline Values	Description
min_words	1	1	Minimum number of words in an opinion to be considered a key point candidate.
max_words	15	10, 12	Maximum number of words in an opinion to be considered a key point candidate.
Q	0.5	0.4, 0.5, 0.7	Minimum argument quality according to a model trained on the ArgQ dataset [144].
θ	0.9	0.856, 0.999	Threshold value for match scores for (1) assigning opinions to key point candidates and (2) merging similar key point candidates.

Table C.3: Hyperparameters for the ArgKP baseline used in the comparison against HyEnA. We also show the originally proposed baseline values from Bar-Haim et al. [33]. Parameters are the same across options.

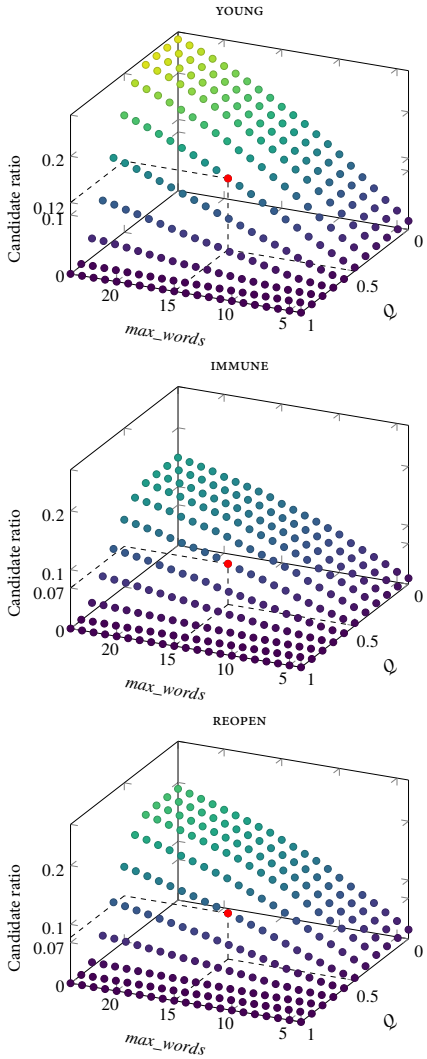


Figure C.1: Hyperparameter sweep for ArgKP (*max_words* and *Q*) and its impact on the ratio of candidates picked. The indicated red dot shows the chosen parameter settings.

C.3 Detailed Results

C.3.1 Unclear Translation Actions

In the argument annotation phase of HyEnA, when extracting arguments from opinions, annotators had the option to skip the opinion if they could not extract any argument from the opinion. Since opinions were automatically translated by the Azure translation service, we also made it optional to indicate that the reason for skipping the argument was because of an unclear translation. Out of 51 actions, annotators indicated mistranslations in 6, 7, and 2 opinions on average for YOUNG, IMMUNE, and REOPEN respectively. This shows that the machine translation caused only some noise, and the majority of the skipped opinions were skipped because of different reasons (e.g. no argument was present in them).

C.3.2 Clustering Arguments

$E = 1$ vs $E = 0$ for single member clusters We also experiment with setting $E = 0$ for argument clusters of size 1 (i.e., clusters containing only a single key argument), as opposed to $E = 1$. The results are displayed in Figure C.2, overlaid over the previous results where $E = 1$ for single-member clusters (Figure 4.6 in Chapter 4). As expected, the error is low when a large number of clusters are obtained by each method (low r, high k). The optimal parameter setting chosen in our approach corresponds to the tipping point where E switches between low E to high E.

C.3.3 Key Arguments

The key arguments extracted by HyEnA are shown in Tables C.4, C.5 and C.6. The results for the ArgKP automated baseline are shown in Tables C.7, C.8 and C.9. Tables C.10, C.11 and C.12 show the results from the manual expert-driven baseline.

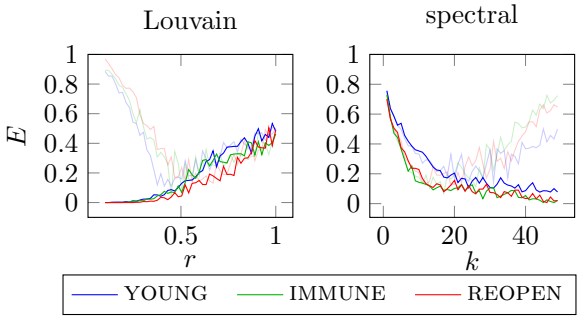


Figure C.2: Parameter tuning for argument clustering with $E = 0$ for argument clusters of size 1. Results are overlaid on Figure 4.6.

Option	ID	Stance	Argument cluster
YOUNG	0	pro	⟨ Social contact is essential for development, It will be positive for support and acceptance, possitive for the psychological health of children, Young people have already suffered enough and got deprived of so many things like parties, holidays, sports. They are missing out on the best time of their lives, Young people's mental health will improve, Removes a lot of annoyance among the elderly, The lifting of this measure significantly reduces loneliness, while having minimal effects, Young people show more cooperation and thinking along when the way they live is taken into account, co they don't have to maintain distance ⟩
	1	pro	⟨ Going back to normality, Second wave, Following research results, this should be possible ⟩
	2	con	⟨ There's a limit to the restrictions, More measures lifted is good, As long as it can still be controlled ⟩
	3	pro	⟨ No risk of contamination , Young people have fewer contamination risks, It's not dangerous for the young people, The group is not at risk at dying of covid, Limited risk, large profit for that group, They're less likely to be contagious, and they're already together anyway. , Young people less infects ⟩
	4	con	⟨ Maintaining distance between your friends and family is easier than being locked down and deprived of the change to make a living ⟩
	5	con	⟨ Joggers don't maintain the distance and the effects of such behaviour are very small and negligible , Maintaining distance while exercising with each other is very difficult, It is dangerous for young people's health to don't keep the distance ⟩
	6	con	⟨ Risk of contamination, The infections will increase, The chances of the second peak of corona virus is too high, The risks are too large, The numbers of the infected have peaked following the holidays, Does not solve the risk of contamination, Unnecessary risk, Who has better immunity system will live, who not will die ⟩
	7	pro	⟨ Economy is more worth then the young ones, The economy will improve and companies won't go bankrupt, They still go to the pub, Life has to go on regardless of the situation, Young people would be happy about going out and meeting friends ⟩

Table C.4: All argument clusters from HyEnA for the option of *Young people may come together in small groups.*

Option	ID	Stance	Argument cluster
YOUNG	8	con	⟨ Exceptions should be considered, Because this cannot be maintained, and it is already violated everywhere, We should be cautious with making big changes to the regulations because it might cause us damage, Entertainment/Events give opportunities to break rules, with this option no longer risk of breaking rules ⟩
	9	con	⟨ People should reasonably decide the distance to maintain, They wouldn't switch between 1,5m distanz with old ones and young ones, they would always be nearer. , People will be more willing to meet and they will do it in larger groups which will enable the spread of the diseas, It is impossible to tell the exact age of people or gauge their immunity, Regional measures will cause problems because people commute between cities. ⟩
	10	pro	⟨ This measure will not be respected, The average Dutchman is too stupid to control themselves when out among people, It is impossible to stop it either way, They don't do it anyway regardless of the rules, People are not responsible enough for the measure to be dropped, They didn't keep the distance before, It is too difficult to follow this rule ⟩
	11	con	⟨ Important measure to archive immunity, Nursing homes can open up only if the measures are followed, Treating all people equally and not just the young ones ⟩
	12	con	⟨ Excessive mesure, It saves a lot of tax for the police because they won't need to observe young people so closely, It is not proven yet whether this would be a good option ⟩
	13	con	⟨ To many young ones would gather ⟩
	14	con	⟨ One rule for all, The young people can contaminate others, Too early ⟩
	15	pro	⟨ Many people already dont do the 1,5m distance, Less victims if they use 1.5 meters at home with fam members ⟩
	16	con	⟨ Lack of control, Easing encourages spread, Every life is worth more than the economy, Netherlands has more than enough resources to at least keep its head above water for a considerable time ⟩
	17	pro	⟨ Only the sick people should stay at home, the same as with the regular flu ⟩
	18	pro	⟨ Young people can studie again and lern together, Children can go easier to school, The schools will be open soon anyway, Young people want to see and socialize with people again, Alternate the students that go to school and the other half attend classes at home ⟩
	19	con	⟨ People will spread the virus more quickly as they will feel more willing to meet in large groups ⟩

Table C.4 continued: All argument clusters from HyEnA for the option of *Young people may come together in small groups*.

Option	ID	Stance	Argument cluster
IMMUNE	0	pro	⟨ it is fair to give immune people freedom of movement ⟩
	1	pro	⟨ could lead to a second peak in cases, These measures are easier to follow compared to other measures, This is a relatively easy measure to take, Public transport use would be easier ⟩
	2	con	⟨ People who still need to follow restrictions will be less likely to when others are not, Immune people would have advantages over the non-immune, and this is unfair, could be seen as discrimination, Everyone should be subject to the same set of rules/restrictions. , Complacency will make it harder for individuals to follow the rules, Young people seem to be getting an advantage over older people ⟩
	3	pro	⟨ Restrictions are unnecessary for people who are immune, Immune people should not be constrained ⟩
	4	con	⟨ Hard to maintain and/or implement, Too little research has been done, It is difficult to control, People can lie if they've contracted the virus ⟩
	5	pro	⟨ People will be able to meet with friends and family members again, It will allow things to get back to normal, People will be happier if they're allowed to go outside, People will be able to see family again, making them happier. , Family can visit each other more often, There will be solidarity between groups and regions, It is fair to give people back their freedom, People will be less lonely and depressed, People want to see their families again, and this measure allows it ⟩
	6	con	⟨ it is unclear if it will be helpful or will make things worse, ICU beds will become more crowded, It's still too early to relax ⟩
	7	con	⟨ It is hard to tell if people are truly immune, Not enough is known about the coronavirus yet, There are too few opportunities to test it, You can't tell who is immune and who isn't, One can lie about having or not having the virus ⟩
	8	pro	⟨ Current restrictions do not really provide any safety, This measure can have a negative effect on society ⟩
	9	con	⟨ It is not clear how people will be able to prove that they are immune, It is hard to know at a glance if someone is immune or not and this will allow some people to fake immunity, there could be immune people with other factors that make them vulnerable, immune people are no longer infective, People who are immune are not dangerous to others, Immunity has not been proven ⟩
	10	con	⟨ will funnel people in certain areas, Risks of transmitting the virus in gatherings ⟩
	11	con	⟨ Infection numbers are still increasing, It risks causing a spike in case numbers, Could lead to the misunderstanding that the situation is safe, Lifting restrictions will cause another wave of Covid, Lifting restrictions will cause people to stop following other rules related to Covid like social distancing. , Too much risk of another spike in cases, By taking this measure, health care would become very pressured ⟩
	12	con	⟨ Infections and morality will increase ⟩
	13	pro	⟨ Advantages to the economy from having immune people working again, This will be beneficial to the economy, People in high-risk of contact jobs will be allowed to return to work, Lifting restrictions will cause economic and social damage. , Lifting restrictions will allow people to feel like things are returning to the pre-Covid normal. , People can go back to work, People who work in contact professions can go back to work, Immune people are, well immune, and can help getting the economy back up ⟩

Table C.5: Argument clusters from HyEnA for the option *All restrictions are lifted for persons who are immune*.

Option	ID	Stance	Argument cluster
REOPEN	0	pro	⟨ This will bring improvement in employment rate, This will improve the economy, This will help these industries recover, to support these sectors and to entertain and please us all, Killing the industry, This helps the economy ⟩
	1	con	⟨ will end up in another confinment, will end with a spike of infections, It is too early, There are less cases now than before ⟩
	2	con	⟨ The difference is we must first protect ourselves from this sickness to then adapt, This will help people satisfy their cravings, People will not benefit a lot from this, This can help people create social interaction and build resistance against COVID ⟩
	3	con	⟨ Leads to more COVID cases , Leads to better moral While keeping Covid cases down, If people die business will still suffer , Things aren't normal yet, Keep sick people away, This will bring more new cases and deaths ⟩
	4	pro	⟨ This can be done only on open spaces, It's already being done in other countries, There are more important industries that needs to be re-opened. , This will help people earn enough to support basic necessities, Tests can be previously made ⟩
	5	con	⟨ will gather a lot of people together, Better moral less infection , This will bring about chaos and lack of control ⟩
	6	con	⟨ These industries are very risky, Risk of spread increases significantly, Catering is a distance of 1.5 meters impossible which leads to great chance of contamination, This increases the chances for the virus to be spread ⟩
	7	pro	⟨ will decrease the number of people with breakdowns, will decrease the contact between people, Keeping group small helps ⟩
	8	pro	⟨ will increase the attendes in the shows, will be controlled environment, With the necessary restrictive measures, cultural events must be able to be visited again as they are an important part of human life, Workers are well protected ⟩
	9	pro	⟨ No evidence that the lockdown works, A distinction should be made, some contact professions are basic service and others are not, Restriction of liberty is a violation of human rights ⟩
	10	pro	⟨ Excited to do things as before for preserving mental health, This will ensure freedom for the people, In order to save people´s lives, we should be very careful and not relax too quickly, To support the churches and meet fellow believers again and pray and sing together ⟩
	11	con	⟨ It's not worth getting people sick, It's not safe yet , These are not vital industries ⟩
	12	pro	⟨ People need to let out pressure , People are tired and bored , Culture and entertainment is important in life, This will make people feel better ⟩
	13	pro	⟨ It will help everyone tremendously, This will help people go back to work, This will motivate people to be more active and healthy ⟩
	14	pro	⟨ Need freedom, It is best to know more of the virus before reopening these industries, This can be done following certain conditions, This will support small businesses recover ⟩
	15	pro	⟨ This will empower the people to be more responsible ⟩
	16	pro	⟨ Cannot be maintained, These places can't be maintained ⟩
	17	pro	⟨ It is easy to maintain social distancing in these industries. ⟩

Table C.6: All argument clusters from HyEnA for the option of *Re-open hospitality and entertainment industry*.

Option	Stance	Arguments
YOUNG	pro	in the long term, this measure is not sustainable in any case
	pro	Low risk group. Easing also gives more space for parents/families.
	pro	if it is not necessary then it is desirable. Also saves on enforcement
	pro	Easing at 1.5m may provide better motivation to comply with other measures
	pro	Youth has the future, it pays a lot for what it 'costs'
	pro	This is hard to maintain. Let's put time into more urgent matters.
	pro	young people are not going to last , a lot of fighting in home situation
	pro	Young people need to support the economy again by getting to work
	pro	Young people need freedom, encourage their own responsibility
	pro	Schools can open 100% again, so parents can also work 100% again
	pro	Can't be stopped. Maintaining this leaves society in a state of cramp.
	pro	Up to the age of 18, this must be the responsibility of parents.
	pro	Relatively little extra pressure on care. Easing this measure benefits education.
	pro	they already had a lot of trouble with it, making it better official
	pro	Untenable for that group, but appeal to solidarity with at-risk groups
	pro	young people do not have the full support to risk
	pro	Help for parents to work better at home
	con	Immunity has not yet been proven. Young people can also transmit the virus.
	con	The rules must remain uniform, otherwise there will be confusion
	con	Young people are better at fighting the Coronavirus
	con	see previous answer Health is for economic importance
	con	young people don't care much about the same problem
	con	We must all stand in solidarity. Moreover, enforcement is easier
	con	Groups with relatively small economic impact if the measures continue to exist for longer.
	con	That way you distinguish between people. This is not advisable for maintaining support.
	con	Young people can easily transfer. No physical/mental distinction between people.
	con	no exceptions for subgroups. Together we get corona under control.
	con	In fact, my motivation is: Equal monks, equal caps.
	con	I don't want to be responsible for the deaths of fellow human beings.
	con	Risk hedging in the near future. Adds nothing
	con	because I am not convinced that well-considered visionary decisions are now being taken
	con	Companies are always at the forefront. Now health comes first No generational differences
	con	Everything is making choices
	con	based on the effects in the explanatory statement, I make that choice.

Table C.7: All arguments from ArgKP for the option of *Young people may come together in small groups.*

Option	Stance	Arguments
IMMUNE	pro	Partly rekindling the economy Better availability of healthcare staff Less protective equipment needed
	pro	that can be used in crucial places
	pro	If you maintain it, I think this is a logical choice.
	pro	Positive effect on loss of income for large group of people.
	pro	Why restrict people's freedom when there's no very urgent reason for it?
	pro	No, it just has to be suffering.
	pro	people are perfectly capable of using their common sense
	pro	The psychological benefits seem much greater than the physical disadvantages.
	pro	they can be deserving of people who are sick
	pro	You can decide what you want. Some feel deprived of their freedom.
	pro	This makes travelling in public transport easier, for example
	pro	These people can therefore reduce the uneaten of the elderly
	pro	Everyone has to be free, but living in a dictatorship very sad
	pro	Survival of the fittest. Reward is in order
	pro	That should be possible n arithmetic could not predict a future
	pro	This seems like a good start to moving for the new world name corona virus
	con	Immunity has not yet been proven. Young people can also transmit the virus.
	con	Immunity has not been established Opening certain provinces gives much more travel
	con	Creates inequality that is not good for social cohesion. Possible source of polarization.
	con	this reduces the willingness of the rest of the netherlands
	con	Too much risk people don't have a size if they are allowed again
	con	Because young people don't stick to it now so it won't matter much
	con	see previous answer Health is for economic importance
	con	In my opinion, the selected items are less urgent than the other
	con	This gives a high degree of inequality within the population
	con	It's way too early for that. R values must remain well below 1
	con	Don't reward groups for already having a problem with the rules.
	con	Because we want to live a normal life again
	con	no exceptions for subgroups. Together we get corona under control.
	con	Enforceability is complicated, keeps simple rules. Moreover, these measures undermine solidarity.
	con	This is uncheckable, you have to show proof everywhere.
	con	because I am not convinced that well-considered visionary decisions are now being taken

Table C.8: All arguments from ArgKP for the option of *All restrictions are lifted for persons who are immune.*

Option	Stance	Arguments
REOPEN	pro	Catering under certain conditions. entertainment as late as possible
	pro	Empower citizens' own responsibilities
	pro	I think those at high risk can be advised to avoid hospitality.
	pro	Hospitality but not entertainment. Catering reasonably similar to shops.
	pro	Only when you're sick do you stay at home, otherwise you don't
	pro	visitors are usually under 50 years of age, can handle this
	pro	Especially lower risk groups use these facilities.
	pro	Everyone can decide for themselves whether they want to go here.
	pro	people are perfectly capable of using their common sense
	pro	People know how to do this. Sufficiently alert to allow this.
	pro	restriction of liberty is violation of human rights
	pro	Make sure the drug is widely available, then the percentages will be even lower
	pro	Who else is going to pay the extra care costs?
	pro	Have seen so many good ideas on media to open responsibly
	pro	Income is also important. Over-50s don't have to participate.
	pro	These companies are also on the rise.
	con	lifting measures northern provinces suffer from hospitality migration within the Netherlands
	con	These options can cause other problems, are uncheckable or easy to bypass.
	con	Too much risk. People will then travel to those regions.
	con	Risk of spreading is far too great. Measure 1.5 meters is impracticable
	con	No distinction between areas in NL Entertainment is less important.
	con	Too dangerous for too little added value.
	con	Somewhere we have to start slowly with normal life again, but with limitations.
	con	Equal treatment of the population
	con	I believe that public support for safety will be greatly reduced.
	con	People are well able to weigh up themselves
	con	people have common sense
	con	A personal choice is not one of the government's.
	con	This is uncheckable, you have to show proof everywhere.
	con	because I am not convinced that well-considered visionary decisions are now being taken
	con	Restaurants also cause addiction damage

Table C.9: All arguments from ArgKP for the option of *Re-open hospitality and entertainment industry*.

Option	ID	Stance	Arguments	Mapped to
YOUNG	0	pro	Young people play a minor role in the spread of the virus and their risk of getting sick is low	3
	1	pro	Social contact is relatively important for young people (to develop themselves)	0
	2	pro	For young people it is difficult not to violate the rules	10
	3	pro	Reduction of problematic psychological symptoms	0
	4	pro	Reduces the pressure on parents	–
	5	pro	Possibility to build up herd immunity	11
	6	pro	Increases support among young people for other lockdown measures	1
	7	con	Constitutes age discrimination which results in a dichotomy in society	14
	8	con	Measures are difficult to enforce. Young people will also get in contact with other people	8

Table C.10: All arguments from the expert-driven manual analysis for the option of *Young people may come together in small groups*. Arguments are **mapped to** argument clusters from HyEnA, showing the cluster ID taken from Table C.4.

Option	ID	Stance	Arguments	Mapped to
IMMUNE	0	pro	These people pose no danger to their environment	3
	1	pro	These people can keep society and the economy going again	13
	2	pro	It is pointless to demand solidarity from these people if they are already immune. Doing so will lead to fierce protests	8
	3	con	Tests for immunity are not foolproof, and this increases the risk of new infections	11
	4	con	Creates a dichotomy in society. People who are not immune can get annoyed by the behaviour of those who are allowed to resume normal life	2
	5	con	Difficult to enforce	4
	6	con	Potential confusion as immunity is not outwardly apparent	7

Table C.11: All arguments from the expert-driven manual analysis for the option of *All restrictions are lifted for persons who are immune*. Arguments are **mapped to** argument clusters from HyEnA, showing the cluster ID taken from Table C.5.

Option	ID	Stance	Arguments	Mapped to
REOPEN	0	pro	This is good for our economy and business	0
	1	pro	It is good for people's well-being	12
	2	pro	This relaxation option will increase support for the continuation of the other measures	–
	3	pro	It is enforceable	4
	4	pro	People can take responsibility for themselves by staying away if they wish	15
	5	pro	We should preserve our cultural heritage and cannot risk bankruptcies in the cultural sector	12
	6	pro	Keeping these businesses closed is too big of a sacrifice for young people	–
	7	pro	In this way, we can build up herd immunity	–
	8	pro	If the hospitality industry is not re-opened people will do other things to relax which is also risky	9
	9	con	Risk of too many people gathering together, which helps to spread the virus	3
	10	con	It is not necessary at the moment	11
	11	con	When alcohol is consumed, people are more likely to underestimate risks and are less likely to comply with distancing measures	–
	12	con	Opening up the hospitality and entertainment sectors should only be considered in the next phase if it appears that other adjustments have worked	14
	13	con	Hospitality industry has a bad impact on society. Please keep it closed	16

Table C.12: All arguments from the expert-driven manual analysis for the option of *Re-open hospitality and entertainment industry*. Arguments are **mapped to** argument clusters from HyEnA, showing the cluster ID taken from Table C.6.

D

Annotator-Centric Active Learning for Subjective NLP Tasks

D.1 Detailed Experimental Setup

Dataset	Task (*dimension*)	Num. Samples	Num. Annotators	Num. Annotations	Num. Annotations per item
DICES	Safety Judgment	990	172	72,103	72.83
MFTC	Morality (*care*)	8,434	23	31,310	3.71
MFTC	Morality (*loyalty*)	3,288	23	12,803	3.89
MFTC	Morality (*betrayal*)	12,546	23	47,002	3.75
MHS	Hate Speech (*dehumanize*, *genocide*, *respect*)	17,282	7,807	57,980	3.35

Table D.1: Overview of the datasets and tasks employed in our work.

D.1.1 Dataset details

We provide an overview of the datasets used in our work in Table D.1. We split the data on samples, meaning that all annotations for any given sample are completely contained in each separate split.

D.1.2 Hyperparameters

We report the hyperparameters for training passive, AL, and ACAL in Tables D.2, D.3, and D.4, respectively. For turning the learning rate for passive learning, on each dataset, we started with a learning rate of 1e-06 and increased it by a factor of 3 in steps until the model showed a tendency to overfit quickly (within a single epoch). All other parameters are kept on their default setting.

Parameter	Value
learning rate	1e-04 (constant)
max epochs	50
early stopping	3
batch size	128
weight decay	0.01

Table D.2: Hyperparameters for the passive learning.

Parameter	Dataset (task)	Value
learning rate	all	1e-05
batch size	all	128
epochs per round	all	20
num iterations	all	10
sample size	DICES	79
sample size	MFTC (care)	674
sample size	MFTC (betrayal)	1011
sample size	MFTC (loyalty)	263
sample size	MHS (dehumanize), MHS (genocide), MHS (respect)	1728

Table D.3: Hyperparameters for the oracle-based active learning approaches.

D.1.3 Training details

Experiments were largely run between January and April 2024. Obtaining the ACAL results for a single run takes up to an hour on a Nvidia RTX4070. For large-scale computation, our experiments were run on a cluster with heterogeneous computing infrastructure, including RTX2080 Ti, A100, and Tesla T4 GPUs. Obtaining the results of all experiments required a total of 231 training runs, combining: (1) two data sampling strategies, (2) four annotator sampling strategies, plus an additional Oracle-based AL approach, (3) a passive learning approach. Each of the above were run for (1) three folds, each with a different seed, and (2) the seven tasks across three datasets. For training all our models, we employ the AdamW optimizer [252]. Our code is based on the Huggingface library [435], unmodified values are taken from their defaults.

D.1.4 ACAL annotator strategy details

Some of the strategies used for selecting annotators to provide a label to a sample

$\mathcal{T}_S$ uses a sentence embedding model to represent the content that an annotator has annotated. We use `all-MiniLM-L6-v2`[1]. We select annotators that have not annotated yet (empty history) before picking from those with a history to prioritize filling the annotation history for each annotator.

[1] `https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2`

Parameter	Dataset	Value
learning rate	all	1e-05
num iterations	DICES	50
num iterations	MFTC (all), MHS (all)	20
epochs per round	DICES, MHS (all)	20
epochs per round	MFTC (all)	30
sample size	DICES	792
sample size	MFTC (care)	1250
sample size	MFTC (betrayal)	1894
sample size	MFTC (loyalty)	512
sample size	MHS (dehumanize), MHS (genocide), MHS (respect)	2899

Table D.4: Hyperparameters for the annotator-centric active learning approaches.

$\mathcal{T}_D$ creates an average embedding for the content annotated by each annotator and selects the most different annotator. We use the same sentence embedding model as $\mathcal{T}_S$. To avoid overfitting, we perform PCA and retain only the top 10 most informative principal components for representing each annotator.

D.1.5 Disagreement rates

We report the average disagreement rates per dataset and task in Figure D.1, for each of the dataset and task combinations.

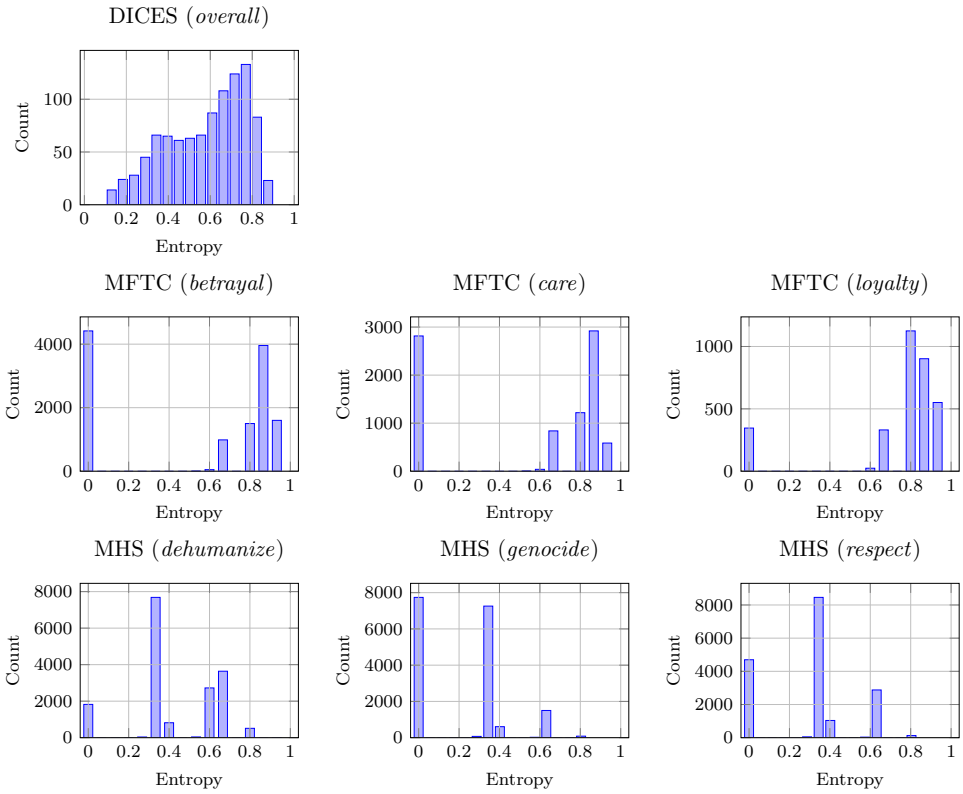


Figure D.1: Histogram of entropy score over all annotations per sample for each dataset and task combination.

D.2 Detailed results overview

D.2.1 Annotator-Centric evaluation for other MFTC and MHS tasks

We show the full annotator-centric metrics results for MFTC *betrayal* and MFTC *loyalty* in Table D.5, and MHS *genocide* and MHS *respect* in Table D.6. This follows the same format as Table 5.1. The results in this table also form the basis for Figure 5.5.

D.2.2 Training process

In Chapter 5, we report a condensed version of all metrics during the training phase of the active learning approaches. Below, we provide a complete overview of all approaches for all metrics. The results can be seen in Figures D.2 through D.8.

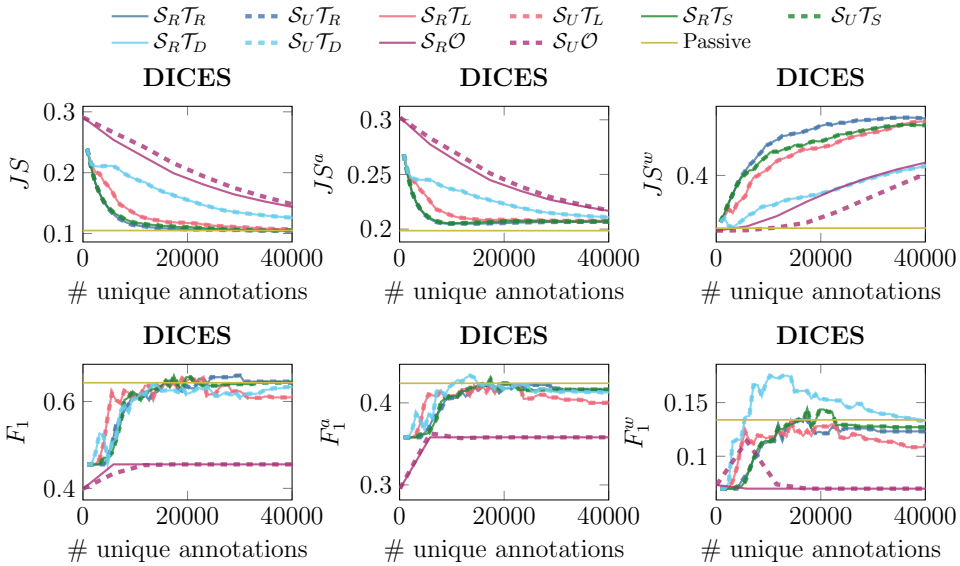


Figure D.2: Validation set performance across all metrics for DICES during training.

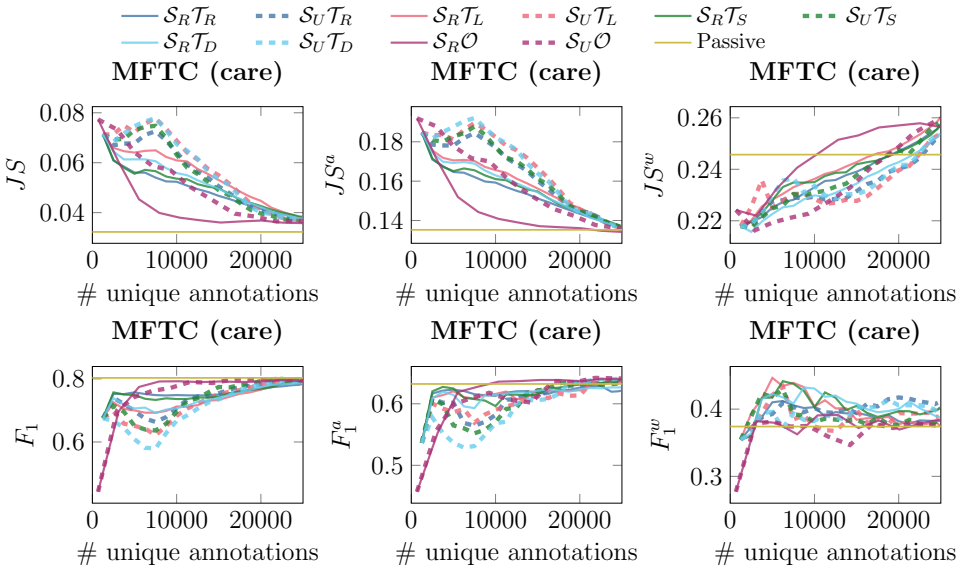


Figure D.3: Validation set performance across all metrics for MFTC (care) during training

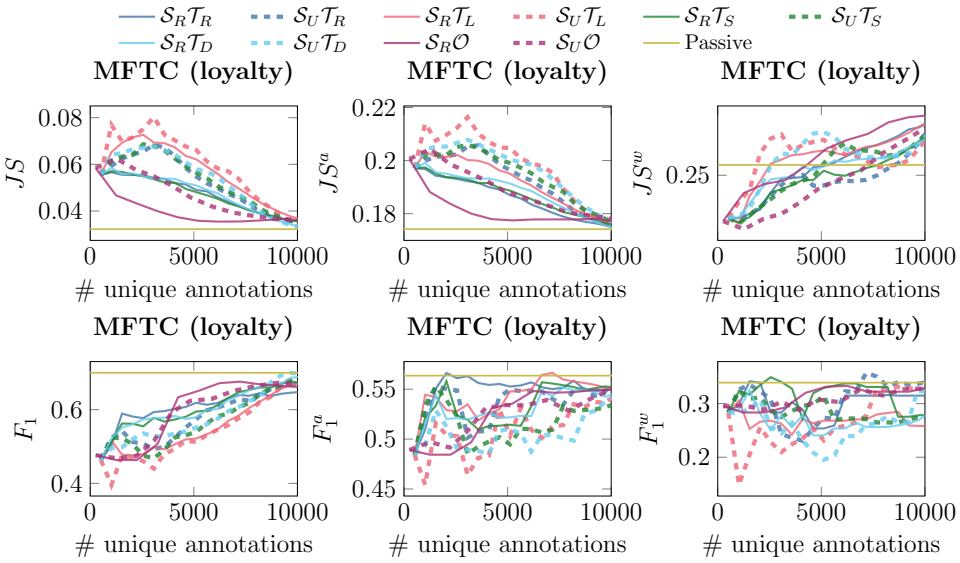


Figure D.4: Validation set performance across all metrics for MFTC (loyalty) during training

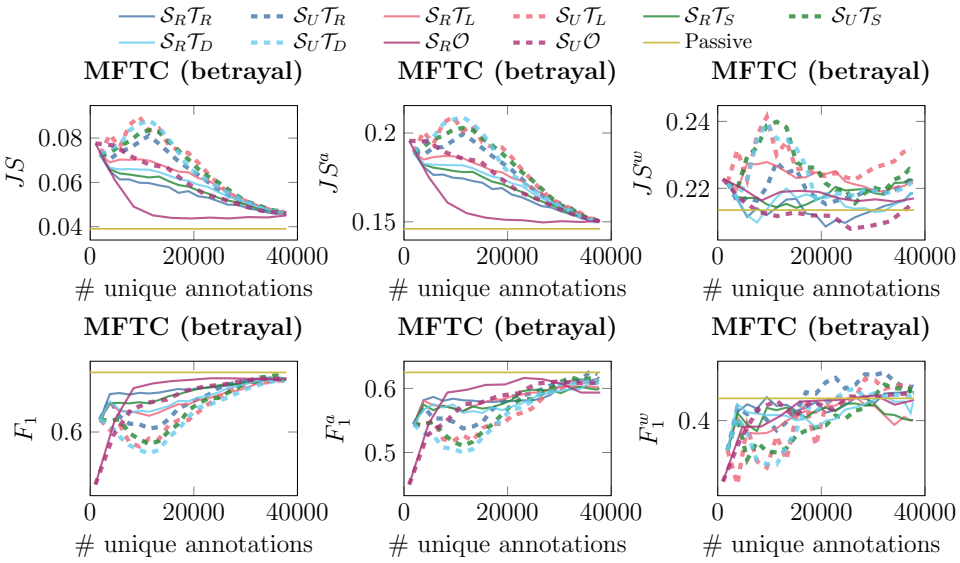


Figure D.5: Validation set performance across all metrics for MFTC (betrayal) during training

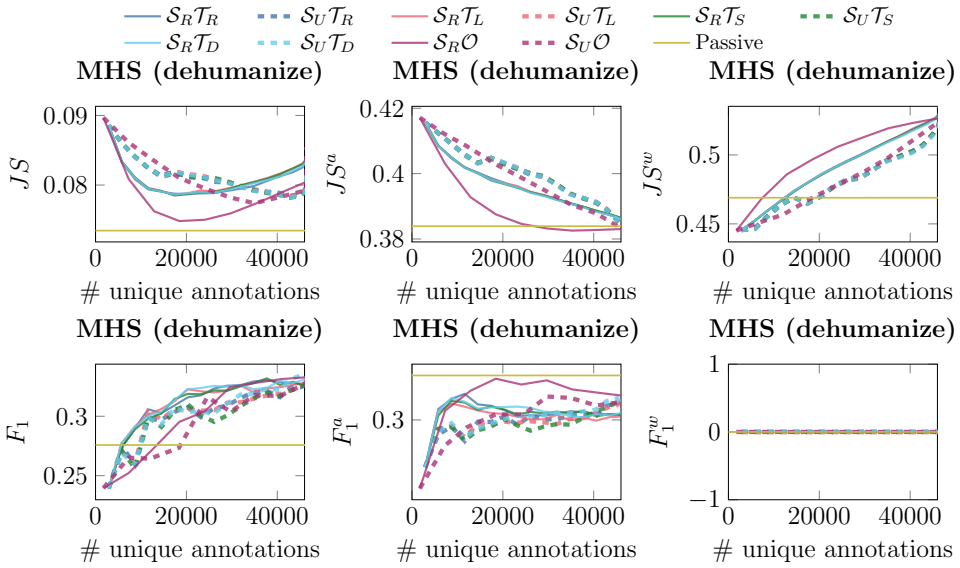


Figure D.6: Validation set performance across all metrics for MHS (dehumanize) during training

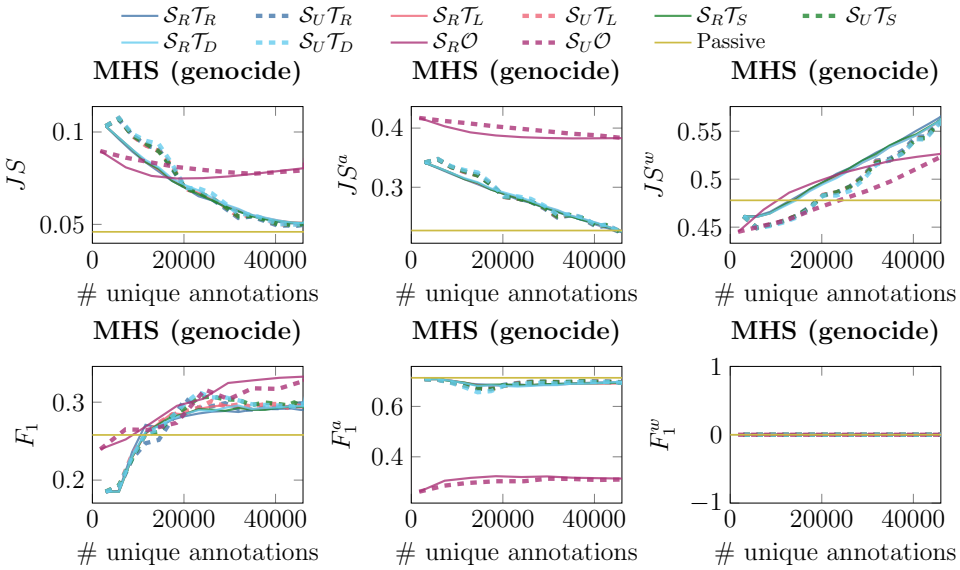


Figure D.7: Validation set performance across all metrics for MHS (genocide) during training

	App.	F_1	JS	Average F_1^a	Average JS^a	Worst-off F_1^w	Worst-off JS^w	Δ%
MFTC (*betrayal*)	$\mathcal{S}_R\mathcal{T}_R$	71.5	.047	57.8	**.147**	42.0	.199	-1.6
	$\mathcal{S}_R\mathcal{T}_L$	71.2	.046	58.1	.149	43.3	.212	-1.6
	$\mathcal{S}_R\mathcal{T}_S$	71.2	.051	59.3	.161	43.0	.239	-5.0
	$\mathcal{S}_R\mathcal{T}_D$	71.0	.046	58.3	.148	42.9	.199	-1.6
	$\mathcal{S}_U\mathcal{T}_R$	72.6	.042	**59.4**	.150	41.9	.203	-2.5
	$\mathcal{S}_U\mathcal{T}_L$	73.6	.045	58.4	.148	43.4	.200	-1.3
	$\mathcal{S}_U\mathcal{T}_S$	74.0	.045	58.8	.149	**43.5**	.204	-1.0
	$\mathcal{S}_U\mathcal{T}_D$	73.2	.044	59.1	.149	42.8	**.194**	-2.6
	$\mathcal{S}_R\mathcal{O}$	72.1	.046	58.9	**.147**	43.1	.195	**-48.6**
	$\mathcal{S}_U\mathcal{O}$	71.8	.047	58.9	.149	43.0	.200	-0.0
	PL	**75.2**	**.037**	48.1	.199	36.0	.290	0.0
MFTC (*betrayal*)	$\mathcal{S}_R\mathcal{T}_R$	66.9	.034	56.4	.177	22.2	.372	-0.4
	$\mathcal{S}_R\mathcal{T}_L$	68.9	.032	56.3	.176	22.2	.374	-0.3
	$\mathcal{S}_R\mathcal{T}_S$	67.1	.031	**57.3**	.176	22.2	.370	-0.3
	$\mathcal{S}_R\mathcal{T}_D$	68.4	.031	55.1	**.175**	22.2	.373	-0.3
	$\mathcal{S}_U\mathcal{T}_R$	61.3	.032	55.7	.177	21.7	.357	-1.1
	$\mathcal{S}_U\mathcal{T}_L$	66.5	.032	54.1	.177	22.2	.355	-0.8
	$\mathcal{S}_U\mathcal{T}_S$	62.4	.033	55.6	.177	22.2	.358	-0.9
	$\mathcal{S}_U\mathcal{T}_D$	64.4	.031	55.8	.177	22.2	.358	-1.3
	$\mathcal{S}_R\mathcal{O}$	**71.5**	.030	56.0	.176	22.2	.361	**-29.1**
	$\mathcal{S}_U\mathcal{O}$	66.5	.033	55.9	.177	22.2	.366	-0.1
	PL	62.5	**.029**	51.2	.183	**26.1**	**.309**	0.0

Table D.5: Test set results on the MFTC (*betrayal*) and MFTC (*loyalty*) datasets. Δ% denotes the reduction in the annotation budget with respect to passive learning.

	App.	F_1	JS	Average F_1^a	Average JS^a	Worst-off F_1^w	Worst-off JS^w	Δ%
MHS (*genocide*)	$\mathcal{S}_R\mathcal{T}_R$	26.5	.050	70.0	.227	0.0	.560	-6.3
	$\mathcal{S}_R\mathcal{T}_L$	28.2	.051	69.8	.225	0.0	.565	-1.7
	$\mathcal{S}_R\mathcal{T}_S$	28.1	.051	70.0	**.224**	0.0	.566	-1.7
	$\mathcal{S}_R\mathcal{T}_D$	28.3	.050	70.2	**.224**	0.0	.565	-1.7
	$\mathcal{S}_U\mathcal{T}_R$	32.8	.077	71.1	.229	0.0	.549	-12.6
	$\mathcal{S}_U\mathcal{T}_L$	27.7	.048	70.7	.231	0.0	.548	-7.9
	$\mathcal{S}_U\mathcal{T}_S$	26.7	.048	70.9	.231	0.0	.548	-7.9
	$\mathcal{S}_U\mathcal{T}_D$	27.3	.048	**71.2**	.229	0.0	.547	-12.6
	$\mathcal{S}_R\mathcal{O}$	28.0	.048	33.9	.387	0.0	**.496**	**-60.1**
	$\mathcal{S}_U\mathcal{O}$	**33.3**	.080	33.1	.390	0.0	.497	-24.7
	PL	21.6	**.044**	70.0	.245	0.0	.570	–
MHS (*respect*)	$\mathcal{S}_R\mathcal{T}_R$	41.4	.086	46.0	.331	0.0	.528	-18.8
	$\mathcal{S}_R\mathcal{T}_L$	40.8	.087	45.6	.331	0.0	.530	-18.8
	$\mathcal{S}_R\mathcal{T}_S$	41.2	.086	46.1	.331	0.0	.529	-18.8
	$\mathcal{S}_R\mathcal{T}_D$	40.6	.086	46.0	.331	0.0	.528	-18.8
	$\mathcal{S}_U\mathcal{T}_R$	32.8	**.077**	**46.6**	**.323**	0.0	.533	-4.9
	$\mathcal{S}_U\mathcal{T}_L$	41.0	.085	46.3	**.323**	0.0	.532	-4.9
	$\mathcal{S}_U\mathcal{T}_S$	**41.8**	.084	45.9	.324	0.0	.531	-4.9
	$\mathcal{S}_U\mathcal{T}_D$	40.6	.085	46.2	.324	0.0	.532	-4.9
	$\mathcal{S}_R\mathcal{O}$	41.7	.085	33.9	.387	0.0	**.496**	**-60.1**
	$\mathcal{S}_U\mathcal{O}$	33.3	.080	33.1	.390	0.0	.497	-24.7
	PL	41.0	.080	25.9	.405	0.0	.587	–

Table D.6: Test set results on the MHS (*genocide*) and MHS (*respect*) datasets. Δ% denotes the reduction in the annotation budget with respect to passive learning.

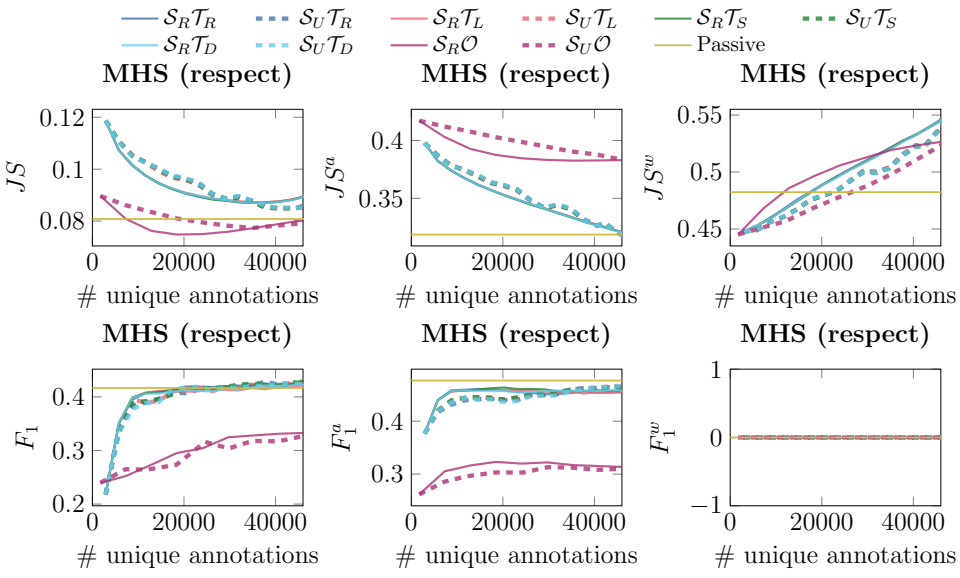


Figure D.8: Validation set performance across all metrics for MHS (respect) during training

E

Do Differences in Values Influence Disagreements in Online Discussions?

E.1 Methodological details

E.1.1 Training Value extraction methods

For training our Transformer-based NLP models, we turned to the Huggingface `transformers` Python package [436]. See Table E.1 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 E.1: 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[1], (2) gaming subreddits[2], (3) image-related subreddits[3], (4) user subreddits, all subreddits starting with "u_", (5) non-English posts (as detected

[1] `https://www.reddit.com/r/ListOfSubreddits/wiki/nsfw`
[2] `https://www.reddit.com/r/gaming/wiki/faq`
[3] `https://www.reddit.com/r/ListOfSubreddits/wiki/sfwporn`

using the FastText [191] Language Identification model[4]), (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. [304]. 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.

E.1.2 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 [270].

- **Informed consent** See Figure E.1. Shown to users before starting the experiment outlining the data protection and disclaimers of any risks.

- **Value Survey** See Figure E.2. Users fill in 21 items on a Likert scale.

- **Annotation instructions** See Figure E.3.

- **Annotation interface** See Figure E.4. 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.

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

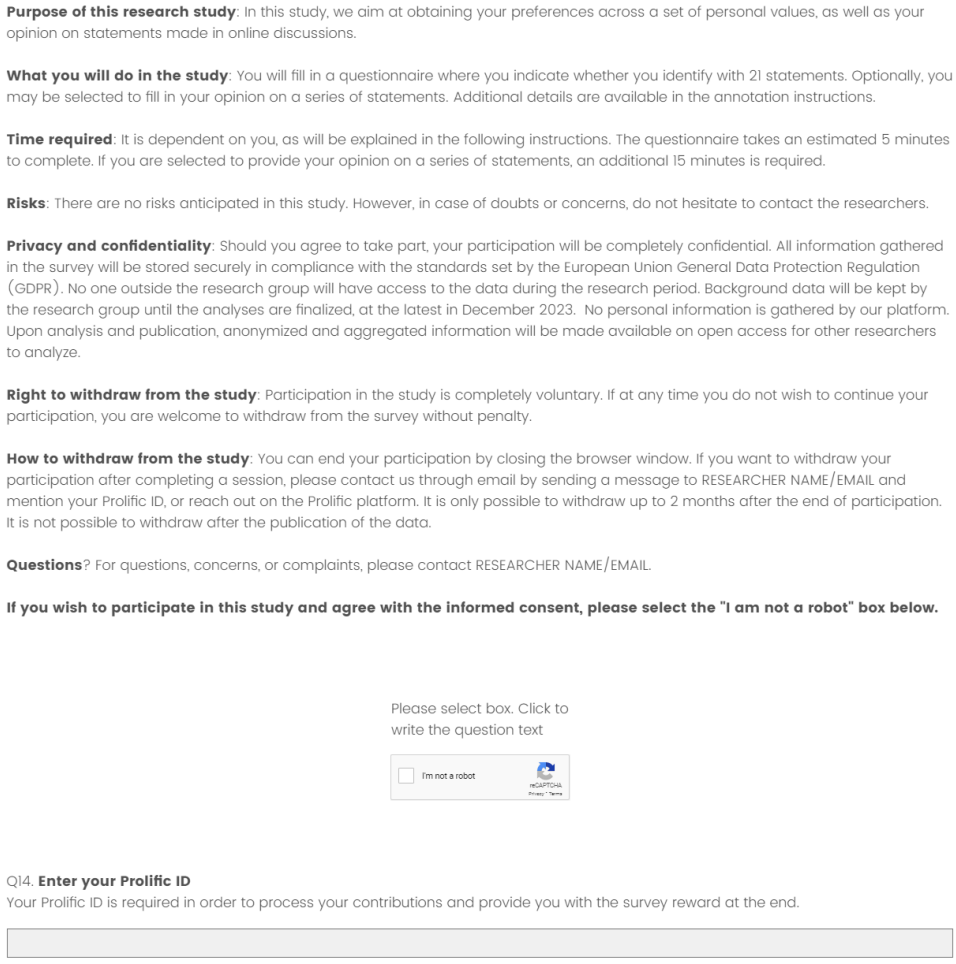


Figure E.1: Informed consent shown to users before starting the experiment.

Transforming survey responses into profiles We adopt the suggestions from Schwartz [344] for constructing a numerical value profile that reflects preferences among values. We create the following pipeline:

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.

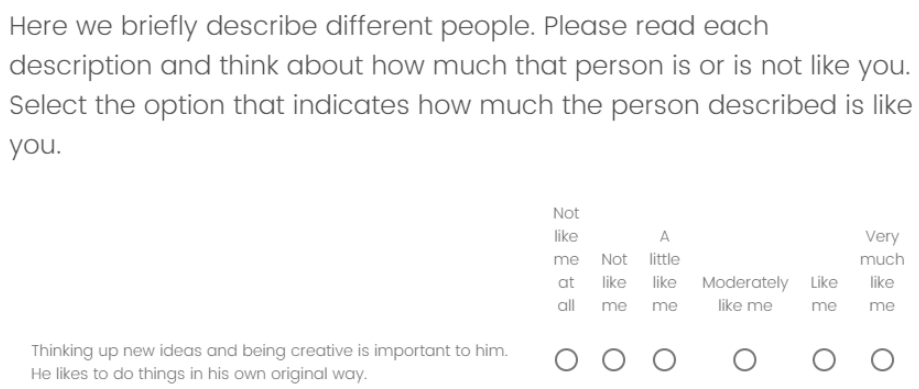


Figure E.2: Screenshot of the PVQ-21 survey.

E.1.3 Training agreement analysis models

Training models for agreement analysis takes around 4 hours for the BERT models on the subsampled Debagreement dataset. See Table E.2 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` [299] Python package. All training involving TF-IDF embeddings takes under 1 hour.

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

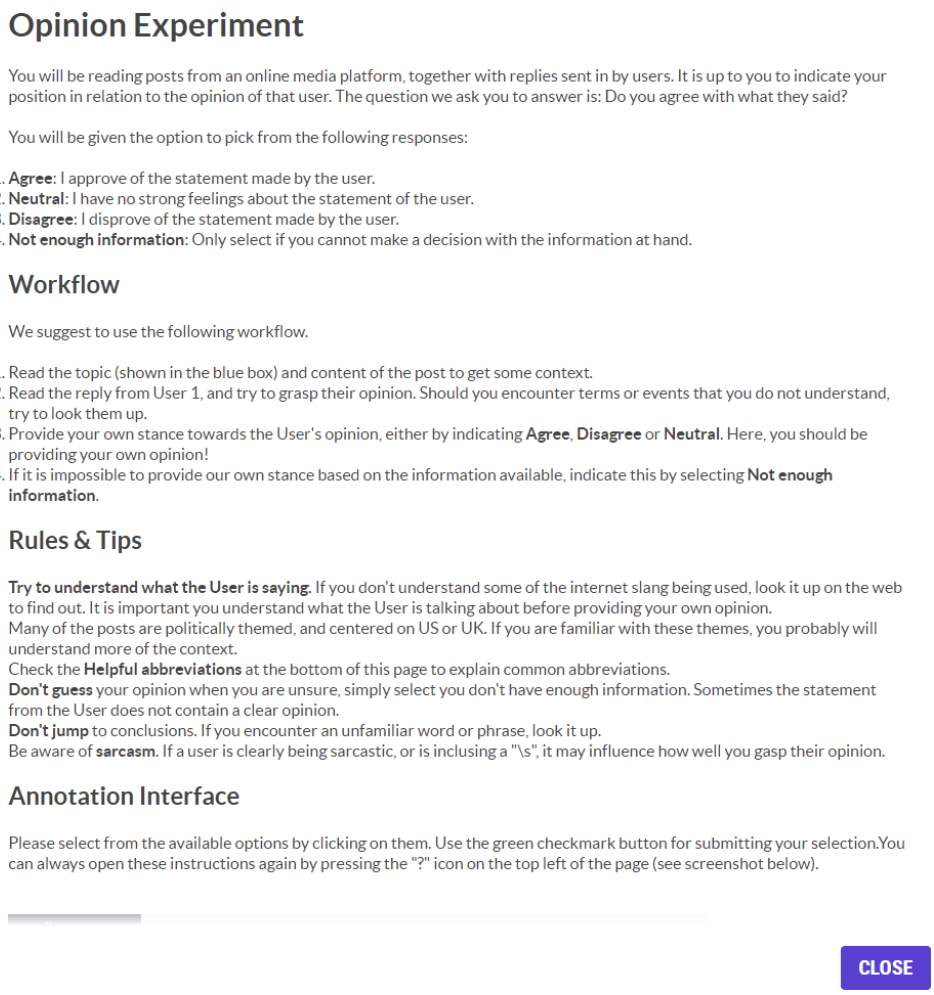
Table E.2: Hyperparameters used for training models for agreement analysis

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 $[0,1)$.

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 E.3 for the features used.



Opinion Experiment

You will be reading posts from an online media platform, together with replies sent in by users. It is up to you to indicate your position in relation to the opinion of that user. The question we ask you to answer is: Do you agree with what they said?

You will be given the option to pick from the following responses:

1. **Agree**: I approve of the statement made by the user.
2. **Neutral**: I have no strong feelings about the statement of the user.
3. **Disagree**: I disprove of the statement made by the user.
4. **Not enough information**: Only select if you cannot make a decision with the information at hand.

Workflow

We suggest to use the following workflow.

1. Read the topic (shown in the blue box) and content of the post to get some context.
2. Read the reply from User 1, and try to grasp their opinion. Should you encounter terms or events that you do not understand, try to look them up.
3. Provide your own stance towards the User's opinion, either by indicating **Agree**, **Disagree** or **Neutral**. Here, you should be providing your own opinion!
4. If it is impossible to provide our own stance based on the information available, indicate this by selecting **Not enough information**.

Rules & Tips

- **Try to understand what the User is saying.** If you don't understand some of the internet slang being used, look it up on the web to find out. It is important you understand what the User is talking about before providing your own opinion.
- Many of the posts are politically themed, and centered on US or UK. If you are familiar with these themes, you probably will understand more of the context.
- Check the **Helpful abbreviations** at the bottom of this page to explain common abbreviations.
- **Don't guess** your opinion when you are unsure, simply select you don't have enough information. Sometimes the statement from the User does not contain a clear opinion.
- **Don't jump** to conclusions. If you encounter an unfamiliar word or phrase, look it up.
- Be aware of **sarcasm**. If a user is clearly being sarcastic, or is inclusing a "\s", it may influence how well you gasp their opinion.

Annotation Interface

Please select from the available options by clicking on them. Use the green checkmark button for submitting your selection.You can always open these instructions again by pressing the "?" icon on the top left of the page (see screenshot below).

CLOSE

Figure E.3: Instructions shown to users for the annotation experiment.

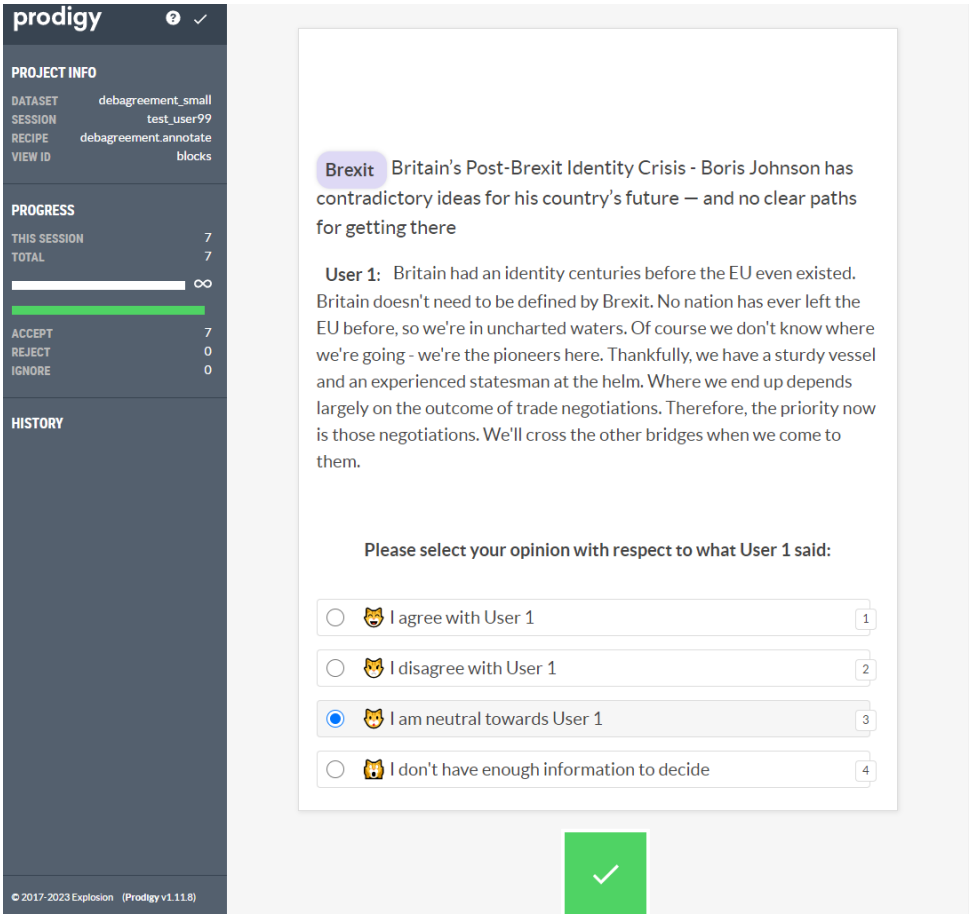


Figure E.4: Annotation interface.

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 E.3: Features used to represent a user from Reddit

E.2 Additional Results

E.2.1 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 E.4.

Method	Training data	P(VN)	R(VN)	F1(VN)	P(VA)	R(VA)	F1(VA)
All-ones	–	0.34	0.50	0.40	0.11	0.50	0.18
VD	–	0.56	0.55	0.45	0.64	0.58	0.59
Kiesel et al. [201]*	*VA*	0.20	0.21	0.15	0.47	0.34	0.37
Qiu et al. [311]*	*VN*	0.64	0.65	0.59	0.53	0.52	0.52
BERT	*VN*	$0.66_{\pm 0.00}$	$0.68_{\pm 0.00}$	$0.66_{\pm 0.00}$	$0.57_{\pm 0.02}$	$0.60_{\pm 0.02}$	$0.57_{\pm 0.03}$
	VA	$0.57_{\pm 0.00}$	$0.56_{\pm 0.00}$	$0.46_{\pm 0.00}$	$0.79_{\pm 0.02}$	$0.74_{\pm 0.01}$	$0.76_{\pm 0.01}$
	Both	$0.63_{\pm 0.00}$	$0.64_{\pm 0.00}$	$0.63_{\pm 0.00}$	$0.84_{\pm 0.02}$	$0.79_{\pm 0.00}$	$0.81_{\pm 0.01}$
RoBERTa	*VN*	$0.61_{\pm 0.15}$	$0.66_{\pm 0.05}$	$0.62_{\pm 0.12}$	$0.58_{\pm 0.02}$	$0.61_{\pm 0.02}$	$0.59_{\pm 0.02}$
	VA	$0.57_{\pm 0.00}$	$0.56_{\pm 0.00}$	$0.46_{\pm 0.00}$	$0.79_{\pm 0.02}$	$0.74_{\pm 0.01}$	$0.76_{\pm 0.01}$
	Both	$0.63_{\pm 0.00}$	$0.64_{\pm 0.00}$	$0.63_{\pm 0.00}$	$0.83_{\pm 0.02}$	$0.78_{\pm 0.01}$	$0.80_{\pm 0.01}$

Table E.4: 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.

E.2.2 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 E.5. 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.

E.2.3 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 E.6 shows examples of the types of pairs we analyze.

E.2.4 Decomposition of BF_{10} results

We create overviews of the different tests performed in Sections 6.4.3 and 6.4.3. 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

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 E.5: Internal consistency scores (Cronbach's α) for the values in the PVQ-21 questionnaire.

show the decomposition for the BF_{10} scores obtained for comparisons between two VPE-estimated profiles in Figures E.5 and for the comparison between VPE and self-reports in Figure E.6. 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 E.7 and E.8 between two VPE profiles, and in Tables E.9 and E.10 for the experiments comparing VPE and self-reported profiles.

E.2.5 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 E.7. 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.

E.2.6 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 E.11.

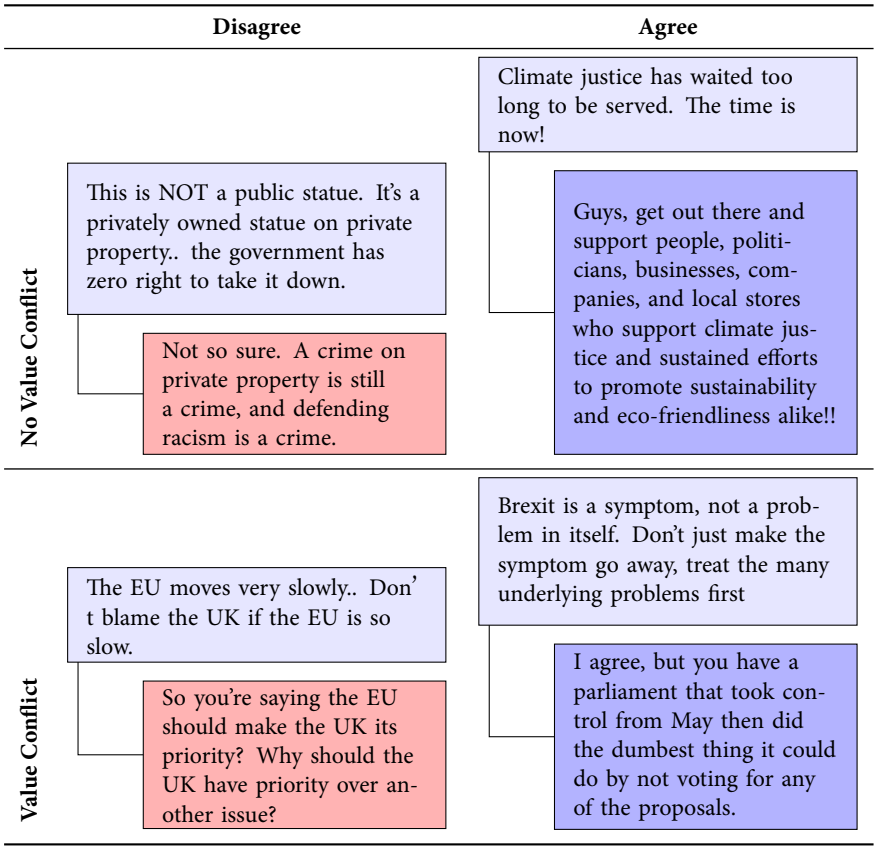


Table E.6: Confusion matrix of qualitative examples of the match between value conflict and (dis-)agreement.

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 E.7: The five tests between two VPE-constructed profiles with the highest BF_{10} scores.

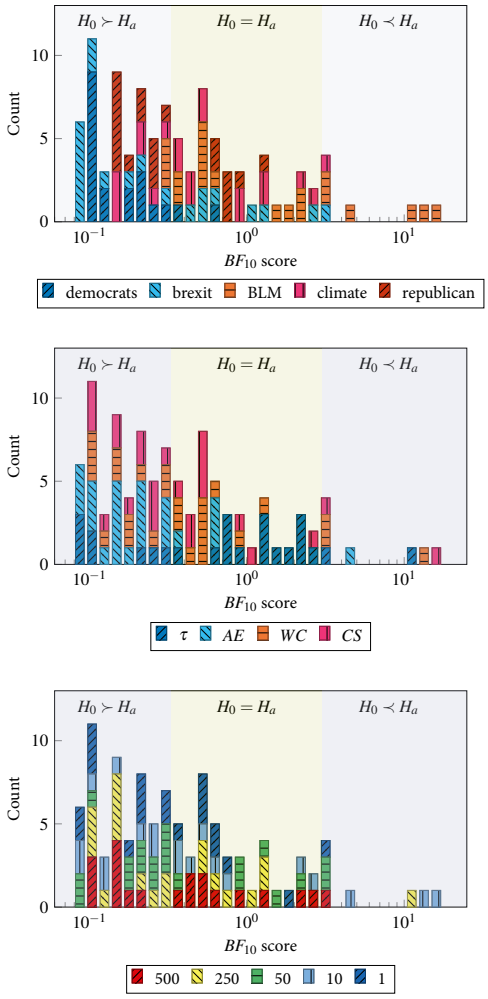


Figure E.5: BF_{10} scores when testing between two VPE-constructed profiles, obtained for all combinations of subreddits (top figure), similarity scores (middle figure) and profile thresholds (bottom figure).

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 E.8: The five tests between two VPE-constructed profiles with the lowest BF_{10} scores.

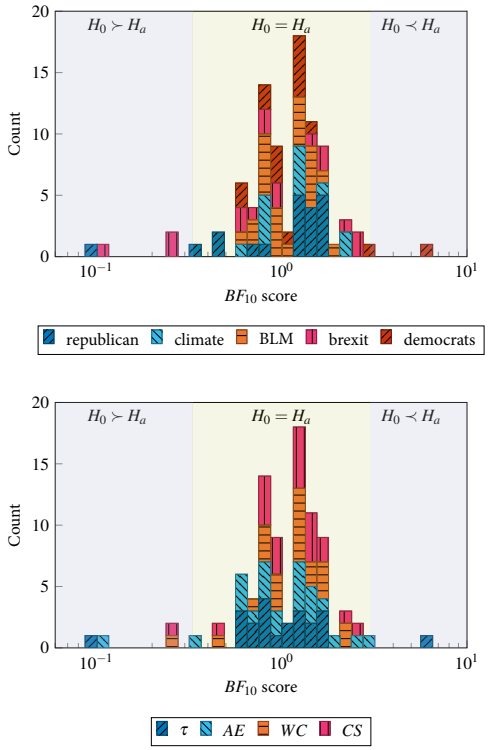


Figure E.6: BF_{10} scores when testing between a VPE-constructed profile and a self-reported profile, split into different subreddits (top figure) and different similarity scores (bottom figure).

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

Table E.9: 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 E.10: The five tests between a VPE-constructed profile and a self-reported profile with the highest BF_{10} scores.

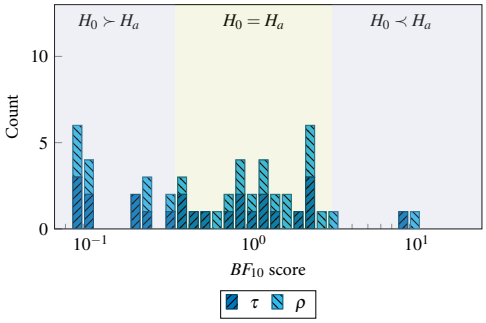


Figure E.7: BF_{10} scores when testing between two VPE-constructed profiles, obtained for the similarity scores Kendall τ and Spearman ρ.

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 E.11: Performance of the agreement classification on a subset of Debagreement (sentence pairs for which both users were available on Reddit).