HyEnA: A Hybrid Method for Extracting Arguments from Opinions

Anonymous ACL submission

Abstract

The key arguments underlying a large and noisy set of opinions help understand the opinions quickly and accurately. Fully automated methods can extract arguments but (1) require large labeled datasets and (2) work well for known viewpoints, but not for novel points of view. We propose HyEnA, a hybrid (human + AI) method for extracting arguments from opinionated texts, combining the speed of automated processing with the understanding and reasoning capabilities of humans. We evaluate HyEnA on three feedback corpora on COVID-19 relaxation measures. We find that, on the one hand, HyEnA achieves higher coverage and precision than a state-of-the-art automated method, when compared on a common set of diverse opinions, justifying the need for human insight. On the other hand, HyEnA requires less human effort and does not compromise quality compared to (fully manual) expert analysis, demonstrating the benefit of combining human and machine intelligence.

1 Introduction

To make decisions on large public issues, such as combating the COVID-19 pandemic and transitioning to green energy, policy makers often turn to the public for feedback (Kythreotis et al., 2019; Lee et al., 2020). This feedback provides insights on the public opinion and contains diverse perspectives. Further, involving the public in the decision-making process helps in gaining their support when the decisions are to be implemented.

In the face of crises, decisions must be made swiftly. Thus, the collection of feedback, its analysis, and recommendations for decision-making are done under tight time constraints. For example, when debating on relaxing COVID-19 measures in the Netherlands, researchers had one month to design the experiment, collect public feedback, and make recommendations (Mouter et al., 2021). The time constraint limits the amount of information researchers can look at, potentially painting an incomplete picture of the opinions. In the scenario above, researchers analyzed data manually, and thus could analyze less than 8% of the feedback provided by more than 25,000 participants.

Argument Mining (AM) (Lawrence and Reed, 2020) methods can assist in increasing the efficiency of feedback analysis by, e.g., separating strongly argumentative feedback from noise and classifying statements as supporting or opposing a decision. However, applying AM methods for feedback analysis poses three main challenges. First, AM methods generalize poorly across domains (Stab et al., 2018; Thorn Jakobsen et al., 2021). Thus, they require large amounts of domain-specific training data, which is often not available. While contextualized representations, using the pre- or fine-tuning paradigm, yield more promising results (Reimers et al., 2019b), they still rely on large amounts of data to be effective. Second, although AM methods can automatically detect logical connections between comments and policy decisions, they do not compress the information. That is, they do not recognize whether two identified arguments describe the same concept, leaving the policy makers with significant manual labor. Finally, analyzing a small sample of comments might cause minority opinions to be ignored (Klein, 2012), creating a bias toward popular (repeated) arguments, which can perpetuate echo chambers and filter bubbles (Price, 1989; Schulz-Hardt et al., 2000).

The *key point analysis* (KPA) task (Bar-Haim et al., 2020a) seeks to automatically compress argumentative discourse into unique *key points*, which can be matched to arguments. However, synthesizing key points is a significant challenge. Bar-Haim et al. (2020a) employ domain experts (skilled debaters) to generate key points and train a model to take over the task. However, such key points are not grounded in data (public opinion) and are subject to the perspectives and biases of the human experts. Further, making use of a few experts to generate key points defeats the purpose of engaging the public in the decision-making process.

We argue for a joint human-machine approach, exploiting both the speed of automated methods and the human understanding of subtle issues. We propose HyEnA (Hybrid Extraction of Arguments), a hybrid (human + AI) method for extracting a diverse set of key arguments from a textual opinion corpus. HyEnA breaks down the argument extraction task into argument *annotation* and *consolidation* phases. In each phase, HyEnA employs human (crowd) annotators, and supports them via intelligent algorithms based on natural language processing (NLP). See Figure 1 for an overview.

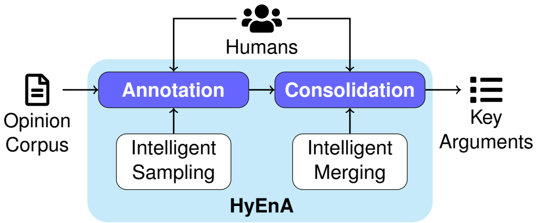

Figure 1: Overview of the HyEnA method.

We evaluate our method on three corpora, each containing more than 10K public opinions on relaxing COVID-19 restrictions (Mouter et al., 2021). We compare HyEnA with an automated approach (Bar-Haim et al., 2020b), which generates key points from the corpus using a pretrained neural argument matching model. In addition, we compare the key arguments generated by HyEnA with insights identified by experts (Mouter et al., 2021).

Contributions (1) We present a hybrid method for key argument extraction, which, given a collection of opinionated user comments, generates a diverse set of key arguments raised in the discussion. (2) We evaluate our method on real corpora of public feedback on policy options. Compared to an automated baseline, HyEnA increases the precision of the key arguments produced, and improves coverage over diverse opinions. Compared to the manual baseline, HyEnA identifies a large portion of arguments identified by experts as well as new arguments that experts did not identify.

2 Related work

We describe related works on AM and methods for extracting key arguments from opinions.

2.1 Computational Argument Analysis

Argument Mining methods (Cabrio and Villata, 2018; Lawrence and Reed, 2020) focus on computational analysis of arguments. They seek to discover arguments brought forward by speakers and identify connections between them. AM is a costly and complex process, and it often requires significant effort by human annotators for reaching moderate inter-rater agreement (Teruel et al., 2018). The ability to recognize and extract arguments from text is dependent on the argumentativeness of the underlying data. Given argumentative texts, popular NLP models are reasonably good at recognizing argumentative discourse (Niculae et al., 2017; Eger et al., 2017; Reimers et al., 2019b). Typically, the first step of AM is to identify the elemental components of arguments (e.g., *claims* and *premises*) in text (Toulmin, 2003). The combination of such components forms a structured argument. However, there is currently no consensus on the exact nature of such elemental components (Daxenberger et al., 2017). Nonetheless, a few characteristics have been recognized as important for recognizing arguments, namely that arguments (1) contain logical reasoning (Stab and Gurevych, 2014), (2) address a *why* question (Biran and Rambow, 2011), and (3) have a non-neutral stance towards the issue being discussed (Stab and Gurevych, 2014).

HyEnA is a novel AM method that combines human annotators and automated NLP models. By splitting up the argument extraction task into distinct phases, we take advantage of the diverse human perspectives, while addressing scalability problems through automation. Because annotators are only given the opinion text, we aim to achieve better grounding by preserving links between argument and the original text, all while providing condensed key arguments useful in analysis.

2.2 Summarization of Arguments

Automated methods have been proposed to create a core set of key points from a large corpus of individual comments (Bar-Haim et al., 2020b). In this paradigm, comments are filtered by a manually tuned selection heuristic, resulting in a list of key point candidates. The candidates are matched against all comments, based on a classifier trained for the argument–key point matching task (Bar-Haim et al., 2020a). We evaluate the performance of this approach on a novel domain on COVID-19 measures and compare it against HyEnA.

Additionally, there exists a body of work on Natural Language Inference (NLI) and Semantic Textual Similarity (STS). In these works, models are trained to indicate semantic similarity or logical entailment between two sentences (Conneau et al., 2017; Reimers et al., 2019a). They have made a significant impact on general-purpose applications (Xu et al., 2018; Zhong et al., 2020). However, downstream applications often need additional fine-tuning (Howard and Ruder, 2018) in order to perform a task well. They also capture generic aspects of semantic similarity and entailment, which may not be applicable to arguments (Reimers et al., 2019a), or conversely overfit to spurious patterns in the data (McCoy et al., 2019).

3 Method

HyEnA is a hybrid method since it combines automated techniques and human judgement (Akata et al., 2020). HyEnA guides human annotators toward the creation of *key arguments* (i.e., groups of semantically distinct arguments that describe relevant aspects of the topic under discussion) from an *opinion corpus* composed of individual *opinions* (i.e., textual comments) on the topic of discussion.

HyEnA consists of two phases (Figure 1). In the first phase (*Key Argument Annotation*), an intelligent sampling algorithm guides human annotators through an opinion corpus to extract high-level information from the opinions. In the second phase (*Key Argument Consolidation*), a new group of annotators merges the results from the first phase, supported by an intelligent merging strategy, involving manual and automatic labeling. In the second phase, HyEnA aims to reduce the subjectivity in annotation. The final result of HyEnA is key arguments grounded on the opinions in the corpus.

3.1 Opinion Corpora

Our opinion corpora are composed of citizens' feedback on COVID-19 relaxation measures, a contemporary topic. The feedback was gathered in April and May 2020 using the Participatory Value Evaluation (PVE) method (Mouter et al., 2021). In the PVE, participants are offered a set of policy options and asked to select their preferred portfolio of choices. Then, the participants are asked to motivate why they picked certain options (*pro* stance) and not pick the other options (*con* stance) via textual comments. Pro- and con-opinions together form the opinion corpus. The PVE collected feedback from 26,293 Dutch citizens on eight policy options about COVID-19 relaxation measures. We analyze the feedback on three of these options, treating feedback on each option as an opinion corpus. Table 1 shows examples. In our experiments, the HyEnA method is applied to one corpus at a time. For each policy option, we use the keyword in uppercase as the policy (or corpus) identifier in the remainder of the paper. The opinions in these corpora are similar to noisy user-generated web comments, as in Habernal and Gurevych (2017). Some opinions span multiple sentences and contain more than one argument.

Table 1: Example opinions in the COVID-19 corpora.

Policy option (Corpus)	Example opinion	# Opinions
YOUNG people may come together in small groups	Then they can go back to school (Pro)	13400
All restrictions are lifted for persons who are IMMUNE	Encourages inequality (Con)	10567
REOPEN hospitality and entertainment industry	The economic damage is too high (Pro)	12814

The original opinions were provided in Dutch. To accommodate a diverse set of annotators in our experiments, we translated all comments to English using the Microsoft Azure Translation service. All experiments are performed with the translated opinions. Mixing (pretrained) embeddings and machine-translated comments has a minimal impact on downstream task performance (Sennrich et al., 2016; Eger et al., 2018; Daza and Frank, 2020). Although all experiments are conducted in English, the link to the original Dutch text is preserved for future applications.

3.2 Key Argument Annotation

In the first phase of HyEnA, human annotators extract individual key argument lists by analyzing the opinion corpus. Since a realistic corpus consists of thousands of opinions, it is unfeasible for an annotator to read all opinions. Thus, HyEnA proposes a fixed number of opinions to each annotator. HyEnA employs NLP and a sampling technique to select diverse opinions to present to an annotator.

Intelligent Opinion Sampling Each annotator is presented, one at a time, a fixed number of opinions. To sample the next opinion, we embed all opinions and arguments observed thus far using the S-BERT model (M_S, Reimers et al., 2019a). S-BERT converts sentences into fixed-length embeddings, which can be used to compute semantic similarities between pairs of sentences.

Then, we select a pool of candidate opinions using the Farthest-First Traversal (FFT) algorithm (Basu et al., 2004). FFT selects the candidate pool as the f farthest opinions in the embedding space from the previously read opinions and annotated arguments (in our experiments, we empirically select $f = 5$). Next, we use an argument quality classifier trained on Gretz et al. (2020) to select the opinion most clear and related to the policy option. In this way, we aim at increasing both diversity and quality of the opinions presented to each annotator.

Annotation Upon reading an opinion, the annotator is asked, first, to *identify* whether the opinion contains an argument or not. If so, the annotator is asked to check whether the argument is already included in their current list of key arguments. If it is not, the annotator should *extract* the argument into a standalone expression (i.e., into a key argument), and add it to the list of key arguments. When adding a new argument, the annotator is asked to indicate the *stance* of the opinion (i.e., whether it is in support or against the related policy option). To facilitate this task, HyEnA highlights the most probable stance for the user as a label suggestion (Schulz et al., 2019; Beck et al., 2021).

Topic Assignment We train a BERTopic model $\mathcal{T}$ on all the available opinions (Grootendorst, 2020). We create a short-list of topics, selected as the most frequent topics found by $\mathcal{T}$, with duplicates and unintelligible topics manually removed by two experts. We ask human annotators to associate the topics from the generated short-list to each argument. This topic assignment T is used in the second phase to compute argument similarity.

Thus, in the first phase, HyEnA yields multiple key argument lists (one per annotator), each containing key arguments and their stances, and an assignment of key arguments to pre-selected topics.

3.3 Consolidation

In the first phase, (1) the annotators are exposed to a small subset of the opinions in the corpus, and (2) the interpretation of arguments is subjective. In the second phase, HyEnA seeks to *consolidate* the key argument lists generated in the first phase. Our goal is to increase the diversity of the resulting arguments and compensate for individual biases.

First, we create the union of all lists of key arguments generated in the first phase of HyEnA. Then, we ask the annotators to evaluate the similarity of the key argument pairs in the union list. Based on the similarity labels, we employ a clustering algorithm to group similar key arguments, producing a consolidated list of key arguments.

Pairwise Annotation To simplify the consolidation task, we present to the annotators one pair of key arguments at a time and ask whether the concepts described by the key arguments in the pair are semantically similar. To reduce human effort, we select only the most informative key argument pairs for manual annotation, and automatically annotate the remaining pairs. To select the most informative pairs, we adapt the Partial-Ordering approach, POWER (Chai et al., 2016), as described below.

Let p_{ij} be a pair of key arguments $\langle a_i, a_j \rangle$. The similarity between the two key arguments in the pair is described by a set of *similarity scores*, s^h_{ij}. By using multiple scores, we seek to make the similarity computation robust. For each p_{ij}, we compute the two similarity scores described in Table 2. We use cosine similarity for s^1_{ij} since the angular distance describes the semantic textual similarity between two arguments. In contrast, we use Euclidean distance for s^2_{ij} since the absolute values of the topic assignment are relevant.

Table 2: The similarity scores between key argument pairs used to create the pairwise dependency graph.

Measure	**Description**
$s^1_{ij} = \frac{\mathbf{i}\cdot\mathbf{j}}{\|\mathbf{i}\|\|\mathbf{j}\|}$	Cosine similarity between embeddings $\mathbf{i} = M_S(a_i)$ and $\mathbf{j} = M_S(a_j)$
$s^2_{ij} = \frac{1}{d(T(a_i),T(a_j))}$	Inverse of the Euclidean distance d between manual topic assignments T of a_i and a_j

Given the similarity scores, we construct a dependency graph G (as in the top-left part of Figure 2), where each key argument pair is a node in G and the edges indicate a Pareto dependency ($\succ$) between two pairs as follows:

$$p_{ij} \succeq p_{i'j'} \qquad \text{if} \quad \forall h \quad s^h_{ij} \geq s^h_{i'j'} \tag{1}$$

$$\begin{aligned} p_{ij} \succ p_{i'j'} \qquad &\text{if} \quad p_{ij} \succeq p_{i'j'} \\ &\text{and} \quad \exists h \quad s^h_{ij} > s^h_{i'j'} \end{aligned} \tag{2}$$

Next, we follow POWER to extract disjoint paths from G. The highlighted path in the bottom-left part of Figure 2 is an example disjoint path. For every path, we perform a pairwise annotation as in the right part of Figure 2. We select the vertex

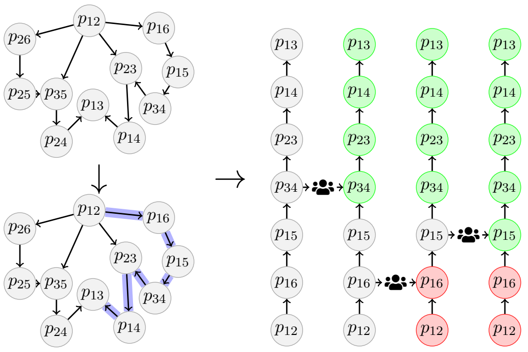

Figure 2: Pairwise annotation from dependency graph.

at the middle of the unlabeled portion of the path and ask multiple (7) humans to indicate whether the concepts described by the two arguments in the pair are similar on a binary scale, and select the label with the majority vote. Given the annotation, we can automatically label (1) all following pairs in the path as similar in case the vertex is labeled as similar or (2) all preceding pairs in the path as non-similar in case the vertex is labeled as non-similar. In essence, using the Pareto dependency, we search for threshold similarity scores for each path, above which all pairs are considered similar, and below which all pairs are non-similar. Because this is a local threshold, we prevent over-generalization. To annotate the complete graph efficiently, we employ the parallel Multi-Path annotation algorithm described in the Appendix.

Clustering Given a similarity label for each key argument pair, our goal is to identify groups of similar key arguments. However, the similarity among key arguments may not be transitive—given $\langle a_1, a_2 \rangle$ as similar and $\langle a_2, a_3 \rangle$ as similar, $\langle a_1, a_3 \rangle$ may be labeled as dissimilar. This can happen because (1) the interpretation of similarity can be subjective (for manually labeled pairs), and (2) the automatic approach is not always accurate (for automatically labeled pairs). Thus, we employ a clustering algorithm for identifying a consolidated list. First, we construct a similarity graph, where each key argument is a node and there is an edge between two arguments if they are labeled as similar. Then, we employ out-of-the-box graph clustering algorithms for constructing argument clusters. These clusters form the *key argument lists*.

4 Experimental Setup

We involve 348 Prolific (www.prolific.co) crowd workers as annotators to evaluate HyEnA. We required the workers to be fluent in English, have an approval rate above 95%, and have completed at least 100 submissions. Our experiment was approved by an Ethics Committee and we received informed consent from each subject.

Table 3 shows an overview of the tasks in the experiment. First, we ask annotators to perform the HyEnA method to generate lists of key arguments for three corpora. Then, we compare the quality of the obtained lists of key arguments with lists generated for the same corpora via two baselines. All tasks except topic generation were performed by the crowd workers. The supplemental material includes the instructions provided to the annotators.

Table 3: Overview of the tasks in the experiment. Items to be annotated can be opinions (O), arguments (A), topics (T), or combinations.

Task	Policy Option	# Items	# Annotators
Key argument annotation	YOUNG	255 (O)	5
	IMMUNE	255 (O)	5
	REOPEN	255 (O)	5
Topic generation	all	45 (T)	2
Topic assignment	YOUNG	90 (A)	10
	IMMUNE	64 (A)	5
	REOPEN	69 (A)	5
Key argument consolidation	YOUNG	1538 (A+A)	99
	IMMUNE	824 (A+A)	57
	REOPEN	940 (A+A)	87
Key argument evaluation	YOUNG	172 (O+A)	28
	IMMUNE	133 (O+A)	21
	REOPEN	157 (O+A)	21

4.1 Phase 1: Key Argument Annotation

In the first phase of HyEnA, each annotator extracts a key arguments list from an opinion corpus. In each corpus, five annotators annotated 51 opinions each, for a total of 255 opinions. Of the 51 opinions, the first is selected randomly, and the following 50 are selected by FFT. This number of opinions was empirically selected to make the annotation feasible within a maximum of one hour.

Topics We train a BERTopic model on each opinion corpus, generating 59, 56, and 72 topics for the YOUNG, IMMUNE, and REOPEN corpora, respectively. Since the number of resulting topics is too high for manual assignment of arguments to topics, we curate a short-list of topics per corpus. We select the 15 most frequent topics in a corpus and ask two experts to remove duplicates (i.e., topics covering the same semantic aspect) and rate the clarity (i.e., how well the topic describes a relevant aspect of the discussion in the corpus) of each topic. Unique topics with an average clarity score above 2.5 compose the short-list of topics. Then, we ask the annotators to assign topics to each key argument generated in the first phase of HyEnA.

4.2 Phase 2: Key Argument Consolidation

In the second phase of HyEnA, we obtain similarity labels $y(a_i, a_j)$ (1 if similar, 0 if not) for all key argument pairs $\langle a_i, a_j \rangle$—some pairs are labeled by the annotators and others are automatically labeled. Given the similarity labels, we construct an argument similarity graph, and cluster the graph to identify a consolidated list of key arguments.

Clustering We experiment with two well-known graph clustering algorithms: (1) Louvain clustering (Blondel et al., 2008) uses network modularity to identify groups of vertices based on a resolution parameter r. (2) Self-tuning spectral clustering (Zelnik-Manor and Perona, 2004) uses dimensionality reduction in combination with k-means to obtain clusters, where k is the desired number of clusters.

We select the parameters of these algorithms to minimize the error metric E shown in Equation 3. The metric penalizes clusters having dissimilar argument pairs. That is, for a cluster $k \in K$ and $\forall a_i, a_j \in k$, if $y(a_i, a_j) = 1$, the error for that cluster is 0. If a cluster contains only a single element, we manually set the error for that cluster to 1, to discourage creating single-member clusters.

$$E = \frac{1}{|K|} \sum_{k \in K} \frac{\sum\limits_{a_i, a_j \in k} \mathbb{1}_{y(a_i,a_j)=0}}{\binom{|k|}{2}} \quad (3)$$

4.3 Baselines

We employ an automated and a manual baseline.

4.3.1 Automated Baseline

We use the **ArgKP** argument matching model (Bar-Haim et al., 2020b) to automatically extract key points from the corpus. ArgKP selects candidate key points from opinions using a manually-tuned heuristic, which filters opinions on their length, form, and predicted argument quality (Gretz et al., 2020). We adopt the hyperparameters from Bar-Haim et al. (2020b), but relax them such that ∼10% of the opinions are selected as candidates by the heuristic instead of the recommended 20%. Candidate key points and opinions are assigned a match score using a pretrained matching network based on RoBERTa (Liu et al., 2019). Opinions only match the highest scoring candidate key points if their match score exceeds a threshold θ (corresponding to the BM+TH approach). After deduplication, this results in a single list of key arguments per option.

We use two metrics, *coverage* (C) and *precision* (P), to compare HyEnA and ArgKP.

Coverage Bar-Haim et al. (2020b) define C as the fraction of opinions mapped to an argument out of all the processed opinions. To compute C, first, we extract the set of key arguments $\mathcal{A}_H$ from HyEnA based on opinions O_H^{obs} ($\subset O$) observed by the annotators. Further, if an argument is extracted from an observed opinion $o_i \in O_H^{obs}$, we add o_i to the set of *annotated* opinions O_H^{ann}. Similarly, we extract the set of key arguments $\mathcal{A}_A$ from ArgKP based on its observed set of opinions $O_A^{obs}(\equiv O)$, producing a set of *annotated* opinions O_A^{ann}. Then, the coverage with respect to *all* observed opinions is:

$$C_H = \frac{|O_H^{ann}|}{|O_H^{obs}|} \quad (4) \qquad C_A = \frac{|O_A^{ann}|}{|O_A^{obs}|} \quad (5)$$

Comparing coverages as defined above may not be fair since the set of observed opinions (i.e., the denominators of Equations 4 and 5) are not the same for HyEnA and ArgKP. Thus, we also compute coverage with respect to a set of *common* opinions, $O_H^{obs} \cap O_A^{obs}$, observed by both methods, as:

$$C_H^{common} = \frac{|O_H^{ann} \cap O_A^{obs}|}{|O_H^{obs} \cap O_A^{obs}|} \quad (6)$$

$$C_A^{common} = \frac{|O_A^{ann} \cap O_H^{obs}|}{|O_H^{obs} \cap O_A^{obs}|} \quad (7)$$

We add the same term to both denominator and numerator in Equations 6 and 7 so that the coverage stays in the range [0, 1]. Further, note that $C_H^{common} = C_H$ since $O_H^{obs}, O_H^{ann} \subset O_A^{obs}(\equiv O)$.

Precision Bar-Haim et al. (2020b) define P as the fraction of mapped opinions for which the mapping is correct. Thus, we must map a set of opinions to arguments in order to compute precision. For this mapping, we select the common opinions, $O_H^{ann} \cap O_A^{ann}$, that are annotated in both HyEnA and ArgKP. Then for each $o_i \in O_H^{ann} \cap O_A^{ann}$, we create two pairs $\langle o_i, \mathcal{A}_H(o_i) \rangle$ and $\langle o_i, \mathcal{A}_A(o_i) \rangle$, where $\mathcal{A}_H(o_i)$ and $\mathcal{A}_A(o_i)$ are the arguments associated with o_i by HyEnA and ArgKP, respectively. Then, we ask annotators to label $z(o_i, a_i) = 1$ for all matching pairs and $z(o_i, a_i) = 0$ for all non-matching pairs, and keep the majority consensus from multiple annotators. Given the opinion-argument mapping, we compute precision as:

$$P_H^{common} = \frac{\sum_{o_i \in O_H^{ann} \cap O_A^{ann}} z(o_i, \mathcal{A}_H(o_i))}{|O_H^{ann} \cap O_A^{ann}|} \qquad (8)$$

$$P_A^{common} = \frac{\sum_{o_i \in O_H^{ann} \cap O_A^{ann}} z(o_i, \mathcal{A}_A(o_i))}{|O_H^{ann} \cap O_A^{ann}|} \qquad (9)$$

4.3.2 Manual Baseline

Mouter et al. (2021) involve six experts to manually analyze the feedback from a sample of participants (2,237 out of 26,293) over all eight policy options and identify key arguments. However, they do not report the exact number of opinions analyzed. Since there are 36,781 opinions for the three options we analyze (Table 1), we estimate the number of opinions the six experts would have analyzed to be 3,129 across the three options. In contrast, HyEnA annotators analyze 765 intelligently selected opinions across the three options.

It is evident that HyEnA reduces the number of opinions analyzed. Further, we investigate the extent to which the key argument lists generated by HyEnA and the manual baseline have comparable insights. To do so, we report the number of HyEnA key arguments that are overlapping, missing, and new compared to the expert-identified key arguments. We cannot compute precision and coverage for the manual baseline because it does not include a mapping between key arguments and opinions.

5 Results and Discussion

Before comparing with the baselines, we analyze the intelligent sampling and merging techniques HyEnA employs in Phases 1 and 2.

5.1 Phase 1: Key Argument Annotation

Table 4 shows the number of different operations annotators perform in Phase 1. On average, the annotators identified 15 unique key arguments per option. About half of the opinions were skipped, mainly because the opinion lacked a clear argument. This is a positive result since the noise (i.e., irrelevant or non-argumentative opinions) in public feedback can be much higher. Thus, the argument quality classifier we incorporate for opinion sampling is effective in filtering noise. Further, the annotators marked only about 15% of the encountered opinions as already annotated key arguments, which shows that the FFT approach is effective in sampling a diverse set of opinions for annotation.

Table 4: The average annotation operations (and their standard deviation) in Phases 1 and 2.

	Phase 1			Phase 2	
Option	**# Args**	**# Skip**	**# Already**	Δ	τ
YOUNG	18.0 (5.5)	23.4 (5.4)	11.4 (9.0)	-61.6%	0.34
IMMUNE	12.8 (2.6)	31.4 (4.5)	8.6 (4.4)	-59.1%	0.42
REOPEN	13.8 (7.6)	29.2 (11.5)	10.2 (7.6)	-59.8%	0.41

5.2 Phase 2: Key Argument Consolidation

Table 4 also shows the benefit of POWER, HyEnA's approach for consolidating key arguments. The number of pairs requiring human annotation (Δ) was on average reduced by 60%. The transitivity score τ (Newman et al., 2002) indicates the extent to which transitivity holds among the similarity labels of argument pairs. The relatively low τ scores justify the subsequent clustering we perform. Louvain clustering yields the smallest error for the YOUNG and IMMUNE corpora, and spectral clustering for REOPEN corpus (additional details in Appendix C.2).

5.3 Comparison with Automated Baseline

Figure 3 compares the coverage and precision of HyEnA and ArgKP. The low coverage (for both methods) indicates that a large number of opinions do not map to a key argument. This is not surprising since real-world opinions are noisy.

Considering *all* observed opinions (C_H and C_A), HyEnA yields slightly higher coverage than ArgKP in the YOUNG and REOPEN corpora. In contrast, ArgKP yields higher coverage than HyEnA in the IMMUNE corpus. We attribute this to the repeated arguments in the IMMUNE corpus. As 83% of opinions are con-opinions, the IMMUNE policy option (Table 1) was highly opposed and its corpus contains many repeated arguments against that option. Since the set of *all* observed opinions is the entire corpus for ArgKP, the repeated arguments inflate its coverage. However, since HyEnA observes only a small subset of diverse opinions from the corpus, the repeated arguments do not influence its coverage significantly. Thus, we compare the coverage of HyEnA and ArgKP with respect to a *common* set of diverse opinions. In this comparison (C_H^{common} and C_A^{common}), HyEnA yields consistently higher coverage (0.34 on average) than ArgKP (0.16 on average) in all three corpora.

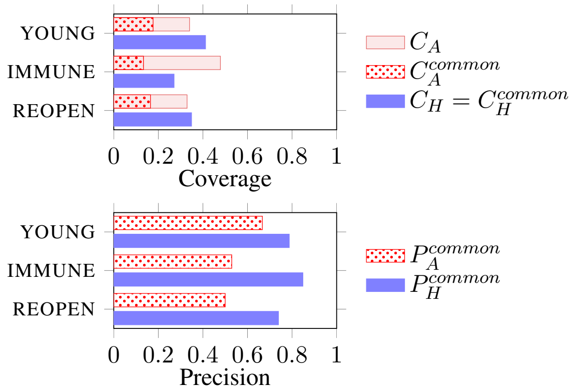

Figure 3: Comparing HyEnA and ArgKP.

ArgKP yields a larger number of key arguments (around 30 for each option) than HyEnA. However, these arguments lead to an average precision of 0.56. In contrast, HyEnA extracts fewer argument clusters (on average 17 per option), but with higher precision (0.80). Further, we notice that HyEnA annotators actively rephrase the content of the key arguments—only in 22% of the annotated key arguments, more than half of the key argument text is directly copied from the original opinion text; in contrast, the key points generated by ArgKP are composed of the original text.

5.4 Comparison with Manual Baseline

Table 5 shows a confusion matrix, comparing overlapping (yes, yes), missing (no, yes), and new (yes, no) key arguments between HyEnA and the manual baseline. HyEnA required an analysis of 765 opinions, whereas the manual baseline required 3,129 opinions to produce their respective key arguments lists. Despite the lower human effort, the HyEnA lists largely overlap with the expert lists.

HyEnA missed some key arguments that the experts identified. For example, a key argument about building herd immunity was not in the HyEnA list for the REOPEN option. We conjecture that increasing the number of opinions annotated in HyEnA would subsequently yield the missing insights.

HyEnA also led to new insights that experts missed. For instance, an argument about the physical well-being of young people was not in the expert list for the YOUNG option. Likely, the random sample of opinions experts analyzed did not include opinions supporting this argument, whereas the smaller set sampled in HyEnA did.

Table 5: Confusion matrix, comparing the key argument lists of HyEnA and manual baseline.

		Manual baseline					
		YOUNG		IMMUNE		REOPEN	
		yes	no	yes	no	yes	no
HyEnA	yes	8	7	7	2	10	1
	no	1	–	0	–	4	–

6 Conclusion and Directions

We develop and evaluate HyEnA, a hybrid method that combines human judgements with automated methods to generate a diverse set of key arguments. HyEnA extracts key arguments from noisy opinions and achieves consistent coverage, whereas the coverage of a state-of-the-art automated method for key point analysis drops by 50% when switching from all (with several repeated) opinions to diverse opinions. Moreover, the key arguments extracted by HyEnA are more precise than those extracted by the automated baseline. Additionally, HyEnA provides important insights that were not included in an expert-driven analysis of the same corpus, despite requiring fewer opinions to be analyzed.

The pairwise comparison in the consolidation phase is the most human-intensive task in HyEnA, and the effort increases with the number of analyzed opinions. Also, comparing arguments is cognitively demanding. HyEnA reduced the number of comparisons required in the consolidation phase by 60%. Additional research is necessary to reduce the consolidation effort further. For example, first clustering the key arguments and then consolidating the arguments within these clusters (reverse order as HyEnA) can influence the performance and effort, but requires further investigation.

Finding arguments in a discourse is only one aspect that constitutes the perspectives in a discussion. Future work can incorporate analysis over the same discourse for values (Liscio et al., 2021) or other perspective factors, such as sentiment, emotion, and attribution (van Son et al., 2016). By combining these rich aspects with arguments, we can merge the logical basis of the discussion with other semantic and syntactic information, allowing close scrutiny of the perspectives in opinions.

7 Ethical Considerations

Our paper develops and evaluates a hybrid (human and AI) approach to extracting key arguments from an opinion corpus. The intended use case for our method is synthesizing key arguments that are grounded in opinionated policy-related comments, by using a pool of annotators. We identify two main aspects of risk in our method.

First, we aim to mitigate the effect of individual biases by grounding the key arguments in general public user opinions. However, the key argument extraction is ultimately performed by individual annotators. We address the influence of subjectivity and noise by combining multiple annotators in the consolidation phase. Further, as our method is transparent, the complete annotation process (from opinions to consolidated key arguments) is traceable. One could implement additional checks on annotator behavior as a bias-mitigating factor, which is a significant research challenge on its own.

Second, the diversity of the opinion embeddings is contingent on the representational quality of the S-BERT model. Underlying biases in its representation may influence the opinions sampled. However, we use FFT to actively sample diverse opinions, which can reduce the impact of inaccurate embeddings.

References

Zeynep Akata, Dan Balliet, Maarten De Rijke, Frank Dignum, Virginia Dignum, Guszti Eiben, Antske Fokkens, Davide Grossi, Koen Hindriks, Holger Hoos, et al. 2020. A research agenda for hybrid intelligence: Augmenting human intellect with collaborative, adaptive, responsible, and explainable artificial intelligence. *Computer*, 53(8):18–28.

Roy Bar-Haim, Lilach Eden, Roni Friedman, Yoav Kantor, Dan Lahav, and Noam Slonim. 2020a. From arguments to key points: Towards automatic argument summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4029–4039.

Roy Bar-Haim, Yoav Kantor, Lilach Eden, Roni Friedman, Dan Lahav, and Noam Slonim. 2020b. Quantitative argument summarization and beyond: Cross-domain key point analysis. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 39–49.

Sugato Basu, Arindam Banerjee, and Raymond J. Mooney. 2004. Active semi-supervision for pairwise constrained clustering. In *Proceedings of the 2004 SIAM International Conference on Data Mining*, SDM '04, pages 333–344, Orlando, Florida, USA. Society for Industrial and Applied Mathematics.

Tilman Beck, Ji-Ung Lee, Christina Viehmann, Marcus Maurer, Oliver Quiring, and Iryna Gurevych. 2021. Investigating label suggestions for opinion mining in German covid-19 social media. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1–13, Online. Association for Computational Linguistics.

Or Biran and Owen Rambow. 2011. Identifying justifications in written dialogs. In *2011 IEEE Fifth International Conference on Semantic Computing*, pages 162–168. IEEE.

Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008.

Elena Cabrio and Serena Villata. 2018. Five years of argument mining: a data-driven analysis. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 5427–5433. International Joint Conferences on Artificial Intelligence Organization.

Chengliang Chai, Guoliang Li, Jian Li, Dong Deng, and Jianhua Feng. 2016. Cost-effective crowdsourced entity resolution: A partial-order approach. In *Proceedings of the 2016 International Conference on Management of Data*, pages 969–984.

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

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680.

Johannes Daxenberger, Steffen Eger, Ivan Habernal, Christian Stab, and Iryna Gurevych. 2017. What is the essence of a claim? cross-domain claim identification. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2055–2066.

Angel Daza and Anette Frank. 2020. X-srl: A parallel cross-lingual semantic role labeling dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3904–3914.

Steffen Eger, Johannes Daxenberger, and Iryna Gurevych. 2017. Neural end-to-end learning for computational argumentation mining. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11–22.

Steffen Eger, Johannes Daxenberger, Christian Stab, and Iryna Gurevych. 2018. Cross-lingual argumentation mining: Machine translation (and a bit of projection) is all you need! In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 831–844, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Shai Gretz, Roni Friedman, Edo Cohen-Karlik, Assaf Toledo, Dan Lahav, Ranit Aharonov, and Noam Slonim. 2020. A large-scale dataset for argument quality ranking: Construction and analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7805–7813.

Maarten Grootendorst. 2020. Bertopic: Leveraging bert and c-tf-idf to create easily interpretable topics. `https://maartengr.github.io/BERTopic/`.

Ivan Habernal and Iryna Gurevych. 2017. Argumentation mining in user-generated web discourse. *Computational Linguistics*, 43(1):125–179.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339.

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

Andrew P Kythreotis, Chrystal Mantyka-Pringle, Theresa G Mercer, Lorraine E Whitmarsh, Adam Corner, Jouni Paavola, Chris Chambers, Byron A Miller, and Noel Castree. 2019. Citizen social science for more integrative and effective climate action: A science-policy perspective. *Frontiers in Environmental Science*, 7:10.

John Lawrence and Chris Reed. 2020. Argument mining: A survey. *Computational Linguistics*, 45(4):765–818.

Sabinne Lee, Changho Hwang, and M Jae Moon. 2020. Policy learning and crisis policy-making: quadruple-loop learning and covid-19 responses in south korea. *Policy and Society*, 39(3):363–381.

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

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448.

Niek Mouter, Jose Ignacio Hernandez, and Anatol Valerian Itten. 2021. Public participation in crisis policymaking. how 30,000 dutch citizens advised their government on relaxing covid-19 lockdown measures. *Plos one*, 16(5):e0250614.

Mark EJ Newman, Duncan J Watts, and Steven H Strogatz. 2002. Random graph models of social networks. *Proceedings of the national academy of sciences*, 99(suppl 1):2566–2572.

Vlad Niculae, Joonsuk Park, and Claire Cardie. 2017. Argument mining with structured svms and rnns. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 985–995.

Vincent Price. 1989. Social identification and public opinion: Effects of communicating group conflict. *Public Opinion Quarterly*, 53(2):197–224.

Nils Reimers, Iryna Gurevych, Nils Reimers, Iryna Gurevych, Nandan Thakur, Nils Reimers, Johannes Daxenberger, and Iryna Gurevych. 2019a. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Nils Reimers, Benjamin Schiller, Tilman Beck, Johannes Daxenberger, Christian Stab, and Iryna Gurevych. 2019b. Classification and clustering of arguments with contextualized word embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 567–578.

Claudia Schulz, Christian M Meyer, Jan Kiesewetter, Michael Sailer, Elisabeth Bauer, Martin R Fischer, Frank Fischer, and Iryna Gurevych. 2019. Analysis of automatic annotation suggestions for hard discourse-level tasks in expert domains. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2761–2772.

Stefan Schulz-Hardt, Dieter Frey, Carsten Lüthgens, and Serge Moscovici. 2000. Biased information search in group decision making. *Journal of personality and social psychology*, 78(4):655.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the*

54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Patrick E Shrout and Joseph L Fleiss. 1979. Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin*, 86(2):420.

Julius Sim and Chris C Wright. 2005. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. *Physical therapy*, 85(3):257–268.

Christian Stab and Iryna Gurevych. 2014. Annotating argument components and relations in persuasive essays. In *Proceedings of COLING 2014, the 25th international conference on computational linguistics: Technical papers*, pages 1501–1510.

Christian Stab, Tristan Miller, Benjamin Schiller, Pranav Rai, and Iryna Gurevych. 2018. Cross-topic argument mining from heterogeneous sources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3664–3674.

Milagro Teruel, Cristian Cardellino, Fernando Cardellino, Laura Alonso Alemany, and Serena Villata. 2018. Increasing argument annotation reproducibility by using inter-annotator agreement to improve guidelines. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).*

Terne Sasha Thorn Jakobsen, Maria Barrett, and Anders Søgaard. 2021. Spurious correlations in cross-topic argument mining. In *Proceedings of *SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics*, pages 263–277, Online. Association for Computational Linguistics.

Stephen E Toulmin. 2003. *The uses of argument*. Cambridge university press.

Chantal van Son, Tommaso Caselli, Antske Fokkens, Isa Maks, Roser Morante, Lora Aroyo, and Piek Vossen. 2016. Grasp: A multilayered annotation scheme for perspectives. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1177–1184.

Jun Xu, Xiangnan He, and Hang Li. 2018. Deep learning for matching in search and recommendation. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 1365–1368.

L Zelnik-Manor and P Perona. 2004. Self-tuning spectral clustering. *Advances in Neural Information Processing Systems*, 17:1601–1608.

Ming Zhong, Pengfei Liu, Yiran Chen, Danqing Wang, Xipeng Qiu, and Xuan-Jing Huang. 2020. Extractive summarization as text matching. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6197–6208.

A 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, inside the `instructions/` directory. In addition, we provide details on the average run times per experiment, as well as any other auxiliary details.

A.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 the instructions for their task. The instructions consist of 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.

Although the opinion corpora contain comments on Dutch policy, the annotators were not restricted to certain (geographical) demographics. 88% of the annotators resided in continental Europe at the time of annotation, with the next 9% residing in Middle and North America. The average age of annotators was 28 ($SD = 7.7$). For 71% of annotators, data on student status was available, and around half (36 p.p.) indicated currently being a student.

A.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 the original paper.

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.

A.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.

A.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.”

A.5 Annotation platform

To run the HyEnA experiments and employ the workers from Prolific (www.prolific.co), we created our own web platform that supports all phases in HyEnA. The platform allows annotators to work in parallel, and is equipped with control mechanisms for conducting the experiments. 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.

B Method Details

B.1 Opinion Corpus

For an overview of the options, see Table 6. Opinions were entered by Dutch citizens in April 2020 following a Participatory Value Estimation (PVE) study. We manually split the data into separate corpora of opinions related to each of the options. Since some opinions entered in the original questionnaire were applicable to multiple options, we copy the opinion for all relevant options. We provide the full dataset of opinions, as well as the annotations performed by the annotators in HyEnA.

Table 6: Statistics for the three policy proposals (options) in the COVID-19 corpus.

Policy option	**Size**	**Pro/Con ratio**
YOUNG people do not need to maintain 1.5 meter distance among each others	13400	0.66/0.34
All restrictions are lifted for persons who are IMMUNE	10567	0.17/0.83
REOPEN hospitality and entertainment industry	12814	0.55/0.45

B.2 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 1 for computational description of the parallel pairwise annotation algorithm (Chai et al., 2016). 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 1: 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
```

B.3 Hyperparameters

B.3.1 HyEnA

An overview of hyperparameters for HyEnA is given in Table 8.

B.3.2 ArgKP

Table 9 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 in Bar-Haim et al. (2020b), but also extract ∼10% of comments 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 4, 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 7 shows the number of comparisons made by the model in use in our experiments.

Option	Stance	# Op.	# Cand.	# Comp.
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 7: Numer of opinions seen (# Op.), candidates extracted (# Cand.) and comparisons made (# Comp.) for running ArgKP.

C Detailed Results

C.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.2 Clustering Arguments

C.2.1 Optimizing for E

Figure 5 show the optimal parameter setting for the clustering methods over each corpus. We also present an alternative visualization, now separated in Figure 6. The lowest observed score is indicated with the red line, obtained by the method in bold.

C.2.2 $E = 1$ 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). The results are displayed in Figure 7, overlaid over the previous results where $E = 1$ for single-member clusters (Figure 5). As

Table 8: Hyperparameters used by HyEnA.

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 9: Hyperparameters for the ArgKP baseline used in the comparison against HyEnA. We also show the values proposed by Bar-Haim et al. (2020b).

Parameter	Option	Value	Baseline Values	Description
min_words	all	1	1	Minimum number of words in an opinion to be considered a key point candidate.
max_words	all	15	10,12	Maximum number of words in an opinion to be considered a key point candidate.
Q	all	0.5	0.4,0.5,0.7	Minimum argument quality according to a model trained on Gretz et al. (2020).
θ	all	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.

expected, error is low when a large number of clusters are obtained by each method (low r, high k). The optimal parameter settings chosen in our approach corresponds to the tipping point where E switches between low E to high E.

C.3 Unclear Argument votes

We show the histogram of unclear argument votes in Figure 8. The majority of arguments receive between 0–10 votes of being unclear, which amount to, on average, less than 5% of the times the argument was observed in a pair. We place a cutoff on 10%, which is around 20 votes or higher depending on the option. After the cutoff, arguments are removed from being included in the clustering. The number of removed arguments can be deduced in Table 10.

Option	$\lvert\mathcal{A}_H\rvert$	After removal	K
YOUNG	90	76	20
IMMUNE	64	60	14
REOPEN	69	64	18

Table 10: Descriptive statistics for the second consolidation phase of HyEnA. $\mathcal{A}_H$ are the arguments extracted from Phase 1, reduced **after removal** of unclear arguments. Finally they are clustered into K groups.

C.4 Clustered Argument Stances

All arguments that were clustered in the second phase of HyEnA were extracted with a particular stance. The clustering method, either Louvain or

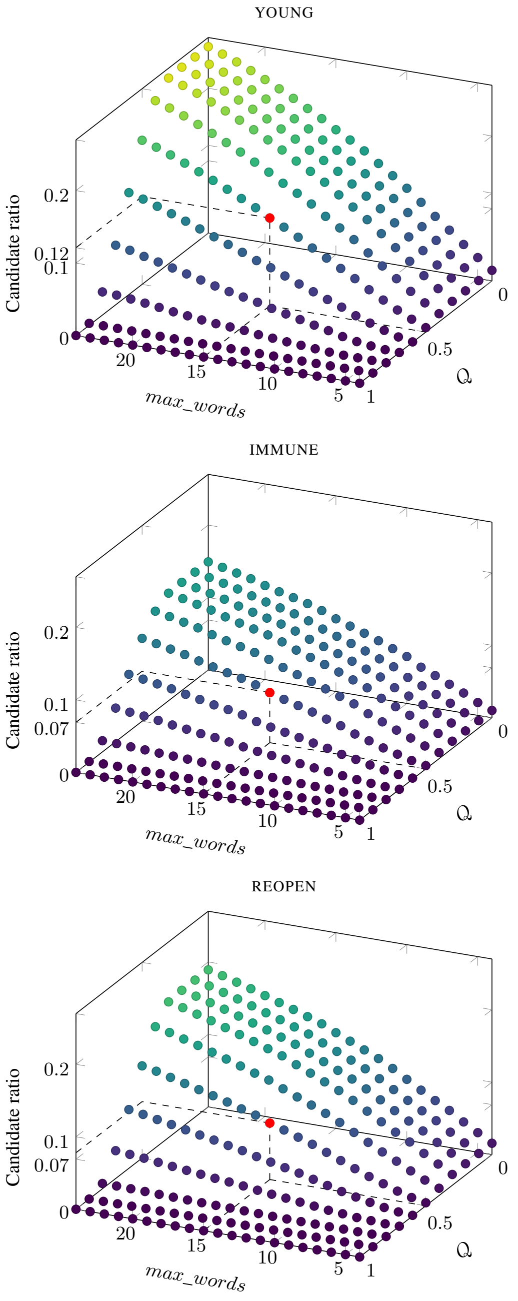

Figure 4: Hyperparameter sweep for ArgKP (*max_words* and Q) and its impact on the ratio of candidates picked. The indicated red dot shows the chosen parameter settings.

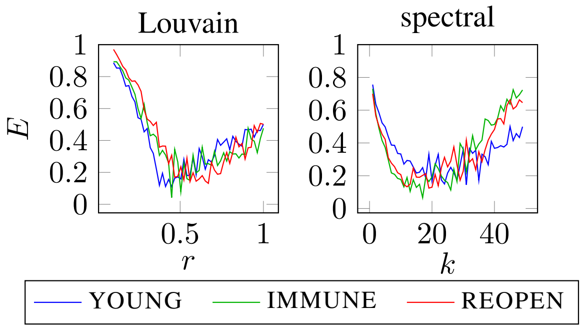

Figure 5: Parameter tuning in key argument clustering.

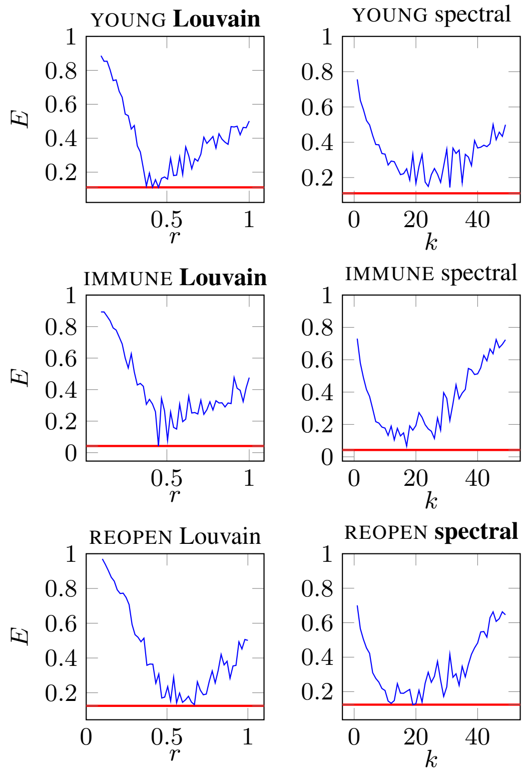

Figure 6: Parameter tuning for argument clustering with $E = 1$ for argument clusters of size 1. Repeats results from Figure 3 from the main paper, now showing the best score (red line) obtained by Louvain for YOUNG and IMMUNE, and spectral for REOPEN.

spectral, clusters based on the obtained similarity labels. However, we can check the correspondence of all stances of the arguments within one cluster, as they should all match. Figure 9 reports the average stance errors per cluster for the three policy options. Stance error is defined as the proportion of stances that do not match the majority stance. In general, the error among stance labels is low; only in some cases mixed stances occur in the clustered arguments. Moreover, only in 5 out of 24 cases the non-majority stance occurs more than once, showing a high agreement between stances inside clusters.

C.5 (Dis-)agreement Analysis

C.5.1 Annotator Reliability

Table 11 shows the inter-rater reliability (IRR) for four steps with overlapping human annotations. In the topic generation phase (Section 4.1), we use the intraclass correlation coefficient ICC$(3, k)$ (Shrout and Fleiss, 1979) since it involves ordinal ratings. In the other three tasks, multiple binary labels are

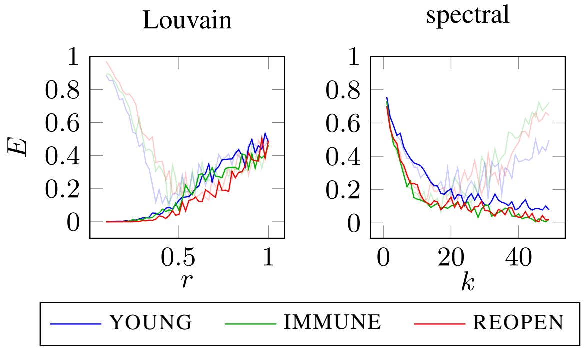

Figure 7: Parameter tuning for argument clustering with $E = 0$ for argument clusters of size 1. Results are overlaid on Figure 3 from the main paper.

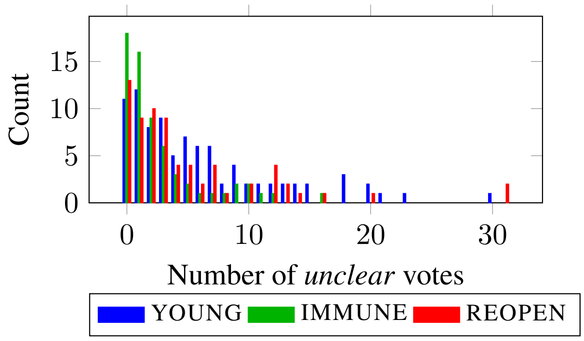

Figure 8: Histogram showing the distribution of unclear votes. Votes are highly left-skewed, and only those with high votes are removed from the final key argument pool using in clustering.

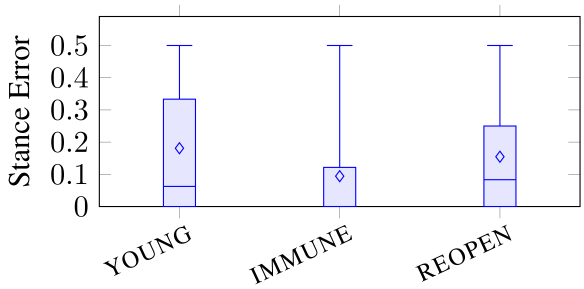

Figure 9: Stance error per final cluster of HyEnA. Overall, low error scores are achieved, indicating high stance correspondence inside clusters.

obtained for the same subjects. In these tasks, we use prevalence- and bias-adjusted κ (PABAK) (Sim and Wright, 2005), which adjusts Fleiss' κ for prevalence and bias resulting from small or skewed distribution of ratings.

The IRR for topic generation and assignment tasks are substantial. The IRR for key argument consolidation and argument evaluation are *fair* and *moderate*, respectively. We pose that the relatively low IRR scores of the latter two tasks are not shortcomings of the HyEnA method in itself. Instead, they demonstrate the complexity of language understanding, and the subtleties involved in interpreting and reasoning about arguments and opinions. Hence, hybrid approaches which use human insight are a key component for public feedback analysis. Uncovering these subtleties and making them explicit is a crucial task for enabling effective perspective taking (Chen et al., 2019). This also justifies the need for a robust argument consolidation phase that integrates judgements from a range of interpretations.

Table 11: Average (and standard deviation) IRR scores.

Task	**ICC3k**	**PABAK**
Topic generation	0.66 (0.14)	–
Topic assignment	–	0.81 (0.10)
Key argument consolidation	–	0.34 (0.03)
Key argument evaluation	–	0.40 (0.06)

C.5.2 Sources of Disagreement

Following the observed IRR scores, we set out to identify the sources of the disagreement between raters for two out of the four tasks mentioned in Table 3 of the main paper.

Topic Generation The main source of the disagreement stems from from a single option: REOPEN. Here, the annotators rated two topics almost inverted (rating 4 versus rating 2) out of a 1–5 Likert scale, resulting in a ICC score of 0.46. The two topics contained the words *"mental health income decrease,"* and *"measures rules these should"*. For the other two options, YOUNG and IMMUNE, a higher score of 0.71 and 0.80 were obtained respectively.

Key Argument Evaluation In the evaluation, annotators observed opinion and argument combinations $\langle o_i, a_i \rangle$, and provided a binary label indicating whether they matched. For analyzing the differences between raters, we compare those combinations where large disagreement was observed (DISAGREE), and compare that to combinations with low disagreement (AGREE). DISAGREE denotes $\langle o_i, a_i \rangle$ combinations that received four annotations for one label, and three for the other, for a total of seven annotations. All other $\langle o_i, a_i \rangle$ are said to AGREE. Specifically, we compared the lengths of the arguments a_i and opinions o_i involved.

The lengths of the arguments in the opinion/argument combinations with large disagreement did not differ from those with low disagreement. However, we found considerably longer opinions on average when annotators disagreed. Possibly, such long opinions contain multiple arguments, which in turn may cause the annotator to mismatch the provided argument.

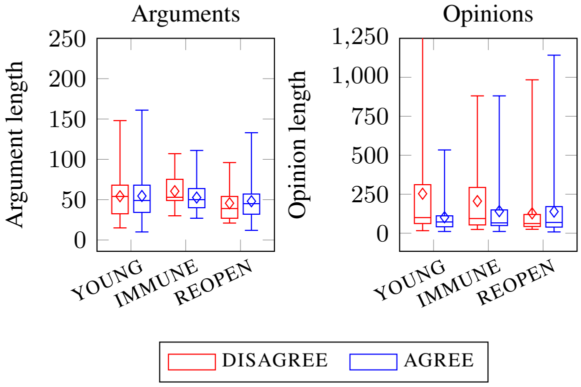

Figure 10: Disagreement analysis for ArgKP automated baseline comparison. On the left figure, argument lengths are the same when annotators agree or disagree. However, as can be seen on the right, when annotators disagree opinions are usually longer.

C.6 Key Arguments

The key arguments extracted by HyEnA are shown in Tables 12, 13 and 14. The results for the ArgKP automated baseline are shown in Tables 15, 16 and 17. Tables 18, 19 and 20 show the results from the manual expert-driven baseline.

Table 12: All argument clusters from HyEnA for the option of *Young people may come together in small groups*.

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 ⟩
	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 13: All argument clusters from HyEnA for the option of *All restrictions are lifted for persons who are immune*.

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

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

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

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

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 18: 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 12.

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 19: 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 13.

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 20: 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 14.

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