

A Hybrid Intelligence Method for Argument Mining

Michiel van der Meer

M.T.VAN.DER.MEER@LIACS.LEIDENUNIV.NL

*Leiden Institute for Advanced Computer Science (LIACS),
Leiden University*

Enrico Liscio

E.LISCIO@TUDELFT.NL

*Interactive Intelligence (II),
Delft University of Technology*

Catholijn M. Jonker

C.M.JONKER@TUDELFT.NL

*Interactive Intelligence (II),
Delft University of Technology*

Aske Plaat

A.PLAAT@LIACS.LEIDENUNIV.NL

*Leiden Institute for Advanced Computer Science (LIACS),
Leiden University*

Piek Vossen

P.T.J.M.VOSSEN@VU.NL

*Computational Linguistics & Text Mining Lab (CLTL),
Vrije Universiteit Amsterdam*

Pradeep K. Murukannaiah

P.K.MURUKANNAIAH@TUDELFT.NL

*Interactive Intelligence (II),
Delft University of Technology*

Abstract

Large-scale survey tools enable the collection of citizen feedback in opinion corpora. Extracting the key arguments from a large and noisy set of opinions helps in understanding the opinions quickly and accurately. Fully automated methods can extract arguments but (1) require large labeled datasets that induce large annotation costs 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 citizen feedback corpora. We find that, on the one hand, HyEnA achieves higher coverage and precision than a state-of-the-art automated method when compared to 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 artificial intelligence.

1. Introduction

To make decisions on large public issues, such as combating a pandemic and transitioning to green energy, policymakers often turn to the citizens for feedback (Kythreotis et al., 2019; Lee et al., 2020). This feedback provides insights into public opinion and contains viewpoints from many individuals with different perspectives. Involving the public in the decision-making process helps in gaining their support when the decisions are to be implemented, fostering the legitimacy of the process (Ostrom, 1990).

In the face of crises, decisions must be made swiftly. Thus, collecting feedback, analyzing it, and making recommendations ought to be performed under tight time constraints. For example, when deciding on relaxing COVID-19 measures in the Netherlands, researchers had one month to design the experiment, collect public feedback, and make recommendations to the government (Mouter et al., 2021). The time constraint limits the amount of information researchers can analyze, potentially painting an incomplete picture of the opinions. In the scenario above, researchers processed data manually and they could only analyze less than 8% of the qualitative feedback provided by more than 25,000 participants.

Argument Mining (AM) (Lawrence & Reed, 2020) methods can assist in increasing the efficiency of feedback analysis by, e.g., locating and interpreting argumentative feedback and classifying statements as supporting or opposing a decision. However, applying automated 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. The use of pretrained language models, with the pre- or fine-tuning paradigm, mitigates but does not solve the reliance on large domain-specific training datasets (Reimers et al., 2019; Ein-Dor et al., 2020). Second, although AM methods can identify argumentative content, they do not compress the information. That is, they do not recognize whether two arguments describe the same point of view, leaving the policymakers with the significant manual labor of aggregating arguments. Finally, naively relying on a small sample of labeled data might cause minority opinions to be ignored as they are not well represented (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. In the ArgKP dataset, domain experts (skilled debaters) were asked to generate key points. Subsequently, a model was trained to take over the task (Bar-Haim et al., 2020b). However, the reliance on few human expert annotators introduces biases of the human experts and may not be representative of the opinions of the larger population. This defeats the purpose of engaging the larger public in a bottom-up deliberative decision-making process.

We argue for a crowd-sourced human-machine approach for argument extraction, combining the scalability of automated methods and the human understanding of others’ perspectives. 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*, *consolidation*, and *selection* phases. HyEnA employs human (crowd) annotators and supports them via intelligent algorithms based on natural language processing (NLP) techniques for analyzing opinions provided by a large audience, as shown in Figure 1.

HyEnA is evaluated 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) performing the KPA task. In addition, we compare the key arguments generated by HyEnA with manually obtained insights identified by experts (Mouter et al., 2021). We find that HyEnA outperforms the automated baseline in terms of precision and diversity, specifically when confronted with a set of varied per-

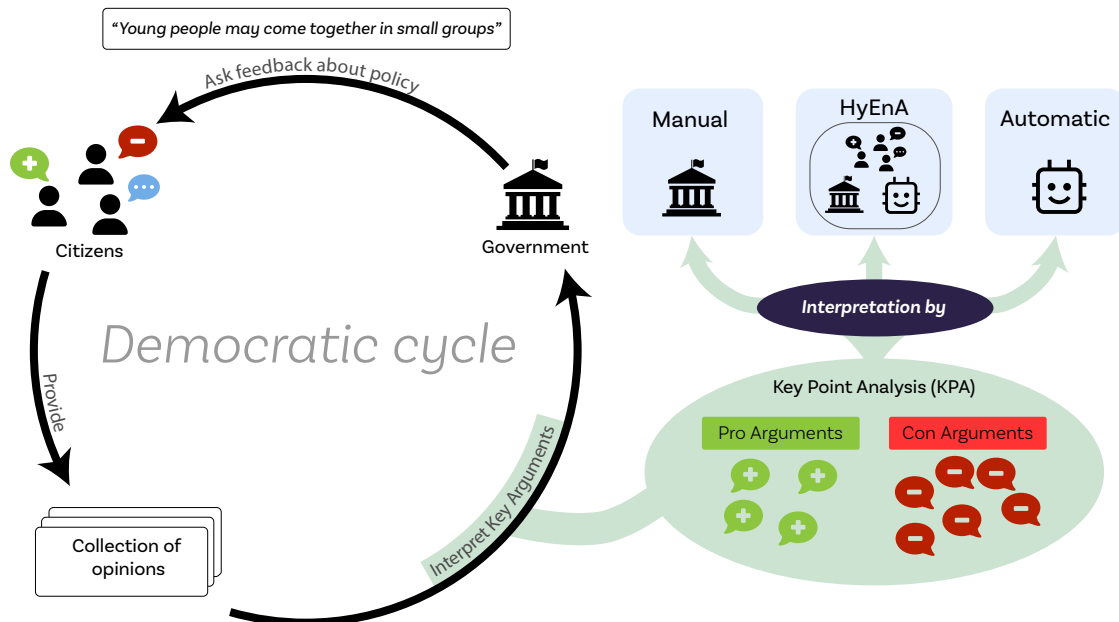


Figure 1: In a democratic cycle, citizens provide their opinions on options for governmental decision-making and their opinions need to be interpreted. Insights into the arguments embedded in their comments can be provided by Key Point Analysis (KPA). To perform KPA, most analysis is performed either manually or automatically. In our work, we propose HyEnA, a hybrid method.

spectives. HyEnA also yields better results than manual analysis, as fewer opinions needed to be analyzed in order to obtain a wider set of key arguments.

Contributions

- C1** We present a hybrid method for key argument extraction, which generates a diverse set of key arguments from a collection of opinionated user comments.
- C2** We evaluate our method on real-world 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.
- C3** We extensively discuss the implications of incorporating recent advances in NLP, such as Large Language Models (LLMs), into the workflow of our hybrid method.

Extension In this paper, we extend the HyEnA method (van der Meer et al., 2022a) to include *argument selection*. The original HyEnA method outputs argument clusters, and leverages manual annotations from the first two phases to select arguments from argument clusters. In this extension, we introduce a method for selecting the most representative

argument from each cluster. The need to summarize argument clusters is not specific to HyEnA, as previous AM applications also retrieve clusters instead of singular arguments (Boltužić & Šnajder, 2015; Wachsmuth et al., 2018; Daxenberger et al., 2020). We compare various techniques to accomplish this task, including generative large language models. Furthermore, we run additional experiments to demonstrate how the new argument selection step can be incorporated in the HyEnA pipeline. Finally, we perform additional analyses to derive further insights from annotators in HyEnA.

Organization Section 2 provides background on Argument Mining for public opinions, and Section 3 introduces the HyEnA method for extracting arguments. We outline the experimental setup in Section 4 and provide extensive results in Section 5. A discussion of our work is given in Section 6 and we conclude with Section 7.

2. Related work

We describe related work on Argument Mining, methods for summarizing arguments, and their application to opinion analysis.

2.1 Computational Argument Analysis

Argument Mining (AM) methods (Cabrio & Villata, 2018; Lawrence & Reed, 2020) focus on the recognition, extraction, and computational analysis of arguments presented in natural language. They seek to discover arguments brought forward by speakers and identify connections between them. Typically, AM techniques concern themselves with finding the *structure* of arguments (van Emmeren et al., 1987), with the goal of finding premises for supporting or refuting conclusions.

AM is a challenging problem. The ability to recognize and extract arguments from text (for humans and machines, alike) is dependent on the argumentativeness of the underlying data. Often, significant effort is required by human annotators for reaching moderate inter-rater agreement when annotating arguments (Teruel et al., 2018). Given argumentative texts, modern NLP models are reasonably good at recognizing argumentative discourse within specific contexts (Niculae et al., 2017; Eger et al., 2017; Reimers et al., 2019).

Typically, the first step of AM is to identify the elemental components of arguments (e.g., *claims* and *premises*) in text (Palau & Moens, 2009). The combination of such components forms a structured argument. However, there is currently no consensus on the exact linguistic notion of such elemental components, with multiple levels of granularity being proposed (Daxenberger et al., 2017; Walton et al., 2008; Freeman, 2011; Bentahar et al., 2010). Nonetheless, a few characteristics have been recognized as important for recognizing arguments, namely that arguments (1) contain (informal) logical reasoning (Stab & Gurevych, 2014), (2) address a *why* question (Biran & Rambow, 2011), and (3) have a non-neutral stance towards the issue being discussed (Stab & 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 through automation.

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 an extended 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 & Gurevych, 2019). They have made a significant impact across a range of tasks (Xu et al., 2020; Zhong et al., 2020). However, downstream applications often need additional fine-tuning (Howard & 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 & Gurevych, 2019), or overfit to spurious patterns in the data (McCoy et al., 2019). Thus, such methods require significant adaptation to effectively compress information in particular domains. Recently, Large Language Models (LLMs) have been shown to perform well on inference tasks with out-of-distribution data (Wang et al., 2023). However, we argue that a plurality of (human) perspectives is necessary to perform sensitive tasks such as the summarization of arguments, which may in turn be used to inform policy-makers about the sentiment of a population (Talat et al., 2022). Yet, LLMs might be adequate for specific subtasks, as we showcase in the third phase of the HyEnA method.

3. Method

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

HyEnA consists of three phases (Figure 2). In the first phase (*Key Argument Annotation*), an intelligent sampling algorithm guides human annotators individually through an opinion corpus to extract high-level information from the opinions. In the second and third phases, HyEnA aims to reduce the subjectivity in the first phase annotations by combining and rewriting arguments that were individually annotated. In the second phase (*Key Argument Consolidation*), an intelligent merging strategy supports a new group of annotators in merging the results from the first phase into clusters of arguments, combining manual and automatic labeling. In the third phase (*Key Argument Selection*), HyEnA employs an automated method to synthesize a single argument that represents the arguments belonging to the same merged argument cluster. The final output of HyEnA is a list of 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 Par-

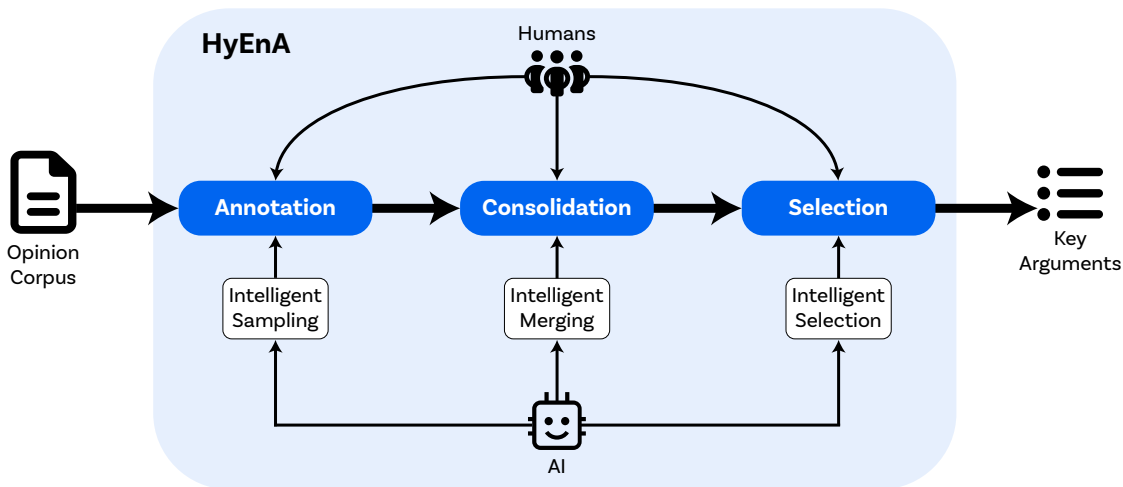


Figure 2: Overview of the HyEnA method.

participatory Value Evaluation (PVE) method (Mouter et al., 2021). In a 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. We analyze feedback from 26,293 Dutch citizens on three of these options, treating comments on each option as an opinion corpus. Table 1 shows examples of opinions provided for each different policy option. In our experiments, the HyEnA method is applied to one corpus at a time. Since we use data from a publicly run citizen feedback experiment, we observe that some options attracted more pro comments than others. We picked these three options with different pro/con ratios to investigate their impact on the key argument extraction task. The opinions in these corpora are similar to noisy user-generated web comments (Habernal & Gurevych, 2017), may span multiple sentences, and contain more than one argument at a time. For each policy option, we use the keyword in uppercase as the option identifier in the remainder of the paper.

Policy option (Corpus)	Example opinion	# Opinions	Pro/Con Ratio
YOUNG people may come together in small groups	Then they can go back to school (Pro)	13400	0.66/0.34
All restrictions are lifted for persons who are IMMUNE	Encourages inequality (Con)	10567	0.17/0.83
REOPEN hospitality and entertainment industry	The economic damage is too high (Pro)	12814	0.55/0.45

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

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 (Senrich et al., 2016; Eger et al., 2018; Daza & 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, with 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 & Gurevych, 2019). 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 the ArgQ dataset (Gretz et al., 2020) to select the opinion most clear and related to the policy option. In this way, we aim at increasing both the 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 use a BERTopic (Grootendorst, 2022) model \mathcal{T} to extract clusters of topics from the corpus. We train \mathcal{T} on all opinions in the corpus and select 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 shortlist 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 pre-selected topics to key arguments.

3.3 Key Argument 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, the annotators are presented with one pair of key arguments at a time and asked 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 adopt a 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 two *similarity scores*, s_{ij}^1 and s_{ij}^2 . 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_{ij}^1 since the angular distance describes the semantic textual similarity between two arguments. In contrast, we use Euclidean distance for s_{ij}^2 since the absolute values of the topic assignment are relevant.

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

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

Given the similarity scores, we construct a dependency graph G (as in the top-left part of Figure 3), where each key argument pair is a node in G and the edges indicate a Pareto dependency (\succ) between two pairs as follows:

$$p_{ij} \succeq p_{i'j'} \quad \text{if } \forall n \quad s_{ij}^n \geq s_{i'j'}^n \quad (1)$$

$$p_{ij} \succ p_{i'j'} \quad \text{if } p_{ij} \succeq p_{i'j'} \quad \text{and} \quad \exists n \quad s_{ij}^n > s_{i'j'}^n \quad (2)$$

Next, we follow POWER to extract disjoint paths from G . The highlighted path in the bottom-left part of Figure 3 is an example disjoint path. For every path, we perform a pairwise annotation as in the right part of Figure 3. We select the vertex 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

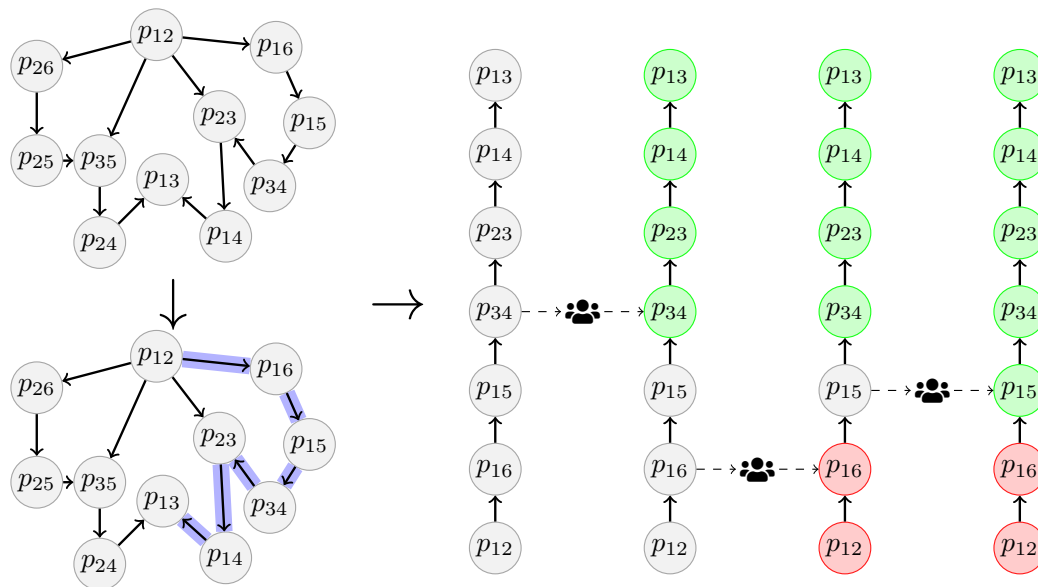


Figure 3: Pairwise annotation of the dependency graph, combining human and automatic judgments.

(1) all following pairs in the path as similar (green) in case the vertex is labeled as similar or (2) all preceding pairs in the path as non-similar (red) 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 (Chai et al., 2016).

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*.

3.4 Key Argument Selection

In the third step of HyEnA, we extract a single argument from each cluster, obtaining the final list of key arguments for the opinion corpus. Formally, for every cluster $k \in K$, we create an argument a_k that is *representative* of that cluster. Argument selection methods

can be extractive (select an argument from the cluster) or abstractive (generate a new argument that summarizes the cluster). Since there are many methods available for selecting arguments, we can experiment with multiple, and pick the best-performing method. In that case, we again pick an intermediate evaluation metric, which we use to select the best selection method. While there is no human annotation involved in this step, we still consider this higher-level algorithmic design a hybrid process, and thus a collaboration between humans and AI. For the task of selecting relevant arguments, we compare the following four types of approaches.

Centroids For every cluster k , we compute a sentence embedding of every argument a_k using M_S . Then, we compute pairwise distances between all arguments inside the same cluster. We select the argument with the lowest average distance, measured using cosine similarity, to all other arguments.

Argument Quality We use a model that measures argument quality to select the argument from each cluster with the highest quality. We use the same argument classifier as in the Key Argument Annotation phase, trained on the ArgQ dataset (Gretz et al., 2020).

Prompting We prompt an LLM to synthesize a single argument out of the arguments provided in the argument cluster (Brown et al., 2020). We experiment with an open-source and a closed-source model.

Random As a baseline, we randomly select an argument from the cluster to represent the entire argument cluster.

4. Experimental Setup

We involve 378 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. We provide supplemental material, containing instructions provided to the annotators, experiment protocol, experiment data, analysis code, and additional details on the experiment (van der Meer et al., 2022b).

Table 3 shows an overview of the tasks in the experiment. First, we ask annotators to perform the HyEnA method to generate key argument lists for three corpora. Then, we compare the quality of the obtained lists with lists generated for the same corpora via two baselines. All tasks except topic generation were performed by the crowd workers, with most of the task instances annotated by multiple annotators to investigate the agreement between annotators.

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 per corpus. 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

Task	Option	# Items	# Annotators	Overlap
Key argument annotation	YOUNG	255 (O)	5	1
	IMMUNE	255 (O)	5	
	REOPEN	255 (O)	5	
Topic generation	all	45 (T)	2	2
Topic assignment	YOUNG	91 (A)	10	5
	IMMUNE	66 (A)	5	
	REOPEN	69 (A)	5	
Key argument consolidation	YOUNG	1538 (A+A)	99	3
	IMMUNE	824 (A+A)	57	
	REOPEN	940 (A+A)	87	
Key argument evaluation	YOUNG	248 (O+A)	42	7
	IMMUNE	193 (O+A)	29	
	REOPEN	221 (O+A)	29	

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

annotation feasible within a maximum of one hour. We instantiate the S-BERT model M_S using the Huggingface Model Hub¹.

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 the 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, the first two authors, 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 shortlist of topics. Then, we ask crowd 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

1. <https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2>

on a resolution parameter r . (2) Self-tuning spectral clustering (Zelnik-Manor & 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 Eq. 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.

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

4.3 Phase 3: Key Argument Selection

In the third phase, we use a mechanism for selecting single arguments per argument cluster. We experiment with multiple methods and different models for selecting arguments. An overview of the methods used is given in Table 4. Below, we explain the setup for each method, and how we select the best-performing method to be used in the final output for HyEnA.

Method	Model	Type	Open	Size
Random	–	extractive	–	–
Centroid	S-BERT	extractive	yes	22M
Prompting	ChatGPT	abstractive	no	175B
	Llama	abstractive	yes	7B
Quality	ArgQ	extractive	no	125M

Table 4: Argument selection algorithms.

Prompts We construct different prompts for the two models to extract the desired argument selection output. *ChatGPT* is an instruction-tuned model and can be prompted to answer questions or follow instructions (Ouyang et al., 2022). *Llama* lacks instruction-tuning, and thus requires prompts designed for next-token generation (Touvron et al., 2023).

Prompt 1: ChatGPT

Consider the context of the COVID-19 pandemic and the following arguments:

- Argument 1
- ⋮
- Argument k

Write a key argument that summarizes the above arguments, and make it short and concise.

Prompt 2: Llama

Consider the context of the COVID-19 pandemic and the following arguments:

- Argument 1

⋮

- Argument k

A short and concise key argument that summarizes the above arguments is:

Testing Cluster Coherence First, we investigate the coherence of the clusters generated in Phase 2 according to each argument selection method, with the intent of measuring how each (automated) method aligns with the results of the first two phases of the (hybrid) HyEnA process. In cases of low coherence, semantically different arguments may end up together. Vice versa, in highly coherent clusters, only arguments that are the same are actually put together. While the error metric E (Equation 3) gives an error rate, it is mostly a *comparative* method, designed to select the best clustering method. Whether or not the clusters make sense to a human interpreter remains unclear. As such, we devise a so-called *odd-one-out* task, in which we use the Argument Selection methods for selecting arguments from a triple of arguments. In this triple, two arguments stem from the same cluster, and the third from a different cluster. The task for each argument selection method is to select which is the deviating argument. Here, we expect an adequate method to succeed well beyond random performance. Because Argument Quality is not intended for pairwise comparisons of arguments, we omit it in the odd-one-out task. We evaluate the remaining methods on a sample of 1K triples uniformly chosen from all possible triple combinations.

Evaluating Argument Selection We use different methods and different models for experimenting with the argument selection phase. As before, we employ an error metric to select the best-performing method, which we then inspect through a human evaluation. We use BERT score (Zhang et al., 2020), a metric designed for model selection that uses a trained BERT model to compare the semantic similarity between the selected argument and the original opinions. Specifically, BERT score recall correlates well with human *consistency* judgments, the factual alignment between selected argument and references (original opinions) (Fabbri et al., 2021). We pick the best-performing method for argument selection based on this metric. This way, we penalize any possible effect of hallucinations of LLMs on the HyEnA method. We take the argument selected by each approach in the Key Argument Selection phase of the HyEnA procedure. As references, we take all comments that were involved in the creation of the cluster. We compute BERTscore and compare it across our approaches.

4.4 Baselines

We compare the output of HyEnA to the results of an automated and a manual approach to key argument extraction.

4.4.1 COMPARISON TO 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). The original approach suggests relaxing heuristic parameters such that 20% of the opinions are selected as candidates. However, this caused overly specific arguments as candidates. Instead, we departed from the parameters used for the ArgKP dataset (Bar-Haim et al., 2020b), and only relax them slightly such that $\sim 10\%$ of opinions are selected as key point arguments.

Candidate key points and opinions are assigned a match score using a model trained for matching arguments 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 best match and threshold (BM+TH) approach. After deduplication, this results in a single list of key arguments per option. We use three metrics, *coverage* (C), *precision* (P), and *diversity* (D) to compare HyEnA and ArgKP.

Coverage (C) is defined as the fraction of opinions mapped to an argument out of all the processed opinions (Bar-Haim et al., 2020b). 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:

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

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

Comparing the coverage scores as defined above naively 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 $[0, 1]$. Note that $C_H^{common} = C_H$ since $O_H^{obs}, O_H^{ann} \subset O_A^{obs} (\equiv O)$.

Precision (P) is the fraction of mapped opinions for which the mapping is correct (Bar-Haim et al., 2020b). 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:

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

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

Diversity (D) is defined as the ratio of key arguments and the number of comments seen by the method. We use diversity to signify how well our method is able to preserve the perspectives present in the opinions seen by the method. In order to compare across methods, we take (1) only correct mappings ($z(o_i, a_i) = 1$) using the labels from P and (2) take the opinions seen by both A and H . We define diversity as follows:

$$D_H = \frac{\mathcal{A}_H}{|O_H^{obs} \cap O_A^{obs}|} \quad (10)$$

$$D_A = \frac{\mathcal{A}_A}{|O_H^{obs} \cap O_A^{obs}|} \quad (11)$$

4.4.2 COMPARISON TO MANUAL BASELINE

A manual analysis involving six experts analyzed the feedback from a sample of participants (2,237 out of 26,293) over all policy options and identify key arguments (Mouter et al., 2021). 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

First, we analyze the inter-rater reliability of annotations. Then, we analyze the intermediate results of the three phases of HyEnA. Finally, we compare our hybrid approach with the automated and manual baselines.

5.1 Annotator Agreement

Table 5 shows the inter-rater reliability (IRR) for four steps with overlapping human annotations. We don’t obtain IRR ratings for the argument extraction task in Phase 1 since the annotation is designed to be disjoint, and raters had little to no overlap in their extractions. In the topic generation phase (Section 4.1), we use the intraclass correlation coefficient $ICC(3, k)$ (Shrout & Fleiss, 1979) since it involves ordinal ratings. In the other three tasks, multiple binary labels are obtained for the same subjects. In these tasks, we use prevalence- and bias-adjusted κ (PABAK) (Sim & Wright, 2005), which adjusts Fleiss’ κ for prevalence and bias resulting from small or skewed distribution of ratings.

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	–	0.36 (0.04)

Table 5: IRR scores per task in HyEnA. We show the average (and standard deviation) over the three option corpora.

The lowest reliability score we obtain for the last two annotation tasks, Key argument consolidation and Key argument evaluation task. The obtained scores may be due to the difficulty of the task—for instance, lay annotators are asked to characterize the similarity between two arguments, and they may not stick to the provided definition of argument similarity. However, task difficulty may not be the only factor at play here. Argument comparisons are made with limited context, and the personal perspective or background of the annotator may influence their judgment. Thus, the low IRR scores may indicate a combination of task difficulty and the relatively subjective nature of the task (Aroyo & Welty, 2015). Similar reasoning holds for the task of evaluating arguments with respect to the original opinions.

Prolific annotators were generally young (M=29.2, SD=7.8) and typically active users with a median of over 300 tasks completed (M=404, SD=418). A little over half of our annotators were male (58.8%), another 38.6% reported as female, and the rest had no data available. 76.7% reported a language other than English as their native language (we did require all annotators to be fluent in English). Annotators mostly resided in European countries, with the UK, Mexico, and the US being the only non-EU countries with more than 10 annotators. 23.8% reported as being a full-time student, with the rest either reporting as not being a student or having no data available. Further work is required in order to investigate the impact of demographic factors on the subjective interpretation of the opinions and arguments involved (Shortall et al., 2022).

5.2 Phase 1: Key Argument Annotation

In Phase 1, individual annotators were guided through 51 opinions each and asked to annotate the observed arguments. Table 6 shows the number of different operations annotators

perform over the 51 opinions. 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. Since the opinions had been automatically translated, we also provided annotators with the option to skip an opinion due to an unclear translation. Out of 51 actions, annotators reported mistranslations in 6, 7, and 2 opinions on average for YOUNG, IMMUNE, and REOPEN, respectively.

Option	Phase 1			Phase 2	
	# 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

Table 6: The average annotation operations (and their standard deviation) in Phase 1, and obtained statistics for Phase 2.

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.

To further examine the behavior of the annotators, we measured the amount of text that was literally copied from the opinions. In Figure 4, we show the distribution of overlap ratios across all extracted arguments. While some arguments do get copied verbatim (overlap ratio of 1), across all three corpora annotators generally rephrase the arguments themselves. This shows that, in HyEnA, human intervention acts in shaping the arguments extracted from the opinions, rather than simply copying part of an opinion (as automated methods would do). Table 7 shows some examples of arguments extracted with different overlap ratios.

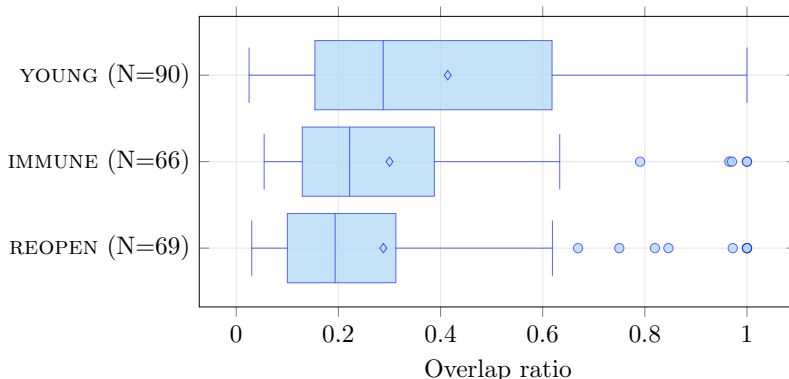


Figure 4: Distribution of argument overlap ratio for arguments generated by Key Argument Annotation in Phase 1.

Option	Opinion Text	Extracted Argument	Overlap Ratio
YOUNG	Our daughter misses her friends so much and I notice that she really needs it	Positive for the psychological health of children	0.060
IMMUNE	Keep one system, keep it simple. Not too many deviations.	Everyone should be subject to the same set of rules/restrictions.	0.091
IMMUNE	Too little research has been done to limit the measures for people who are immune and too few opportunities to test it. In addition, it is difficult to control.	It is difficult to control.	1.000
REOPEN	These measures are quite easy to take compared to the unselected measures.	Measures are easy to take compared to the unselected measures	0.820

Table 7: Examples of copied and paraphrased arguments in Phase 1 of HyEnA.

The topic models for each option generated a large variety of topics. After the generation of the topic models \mathcal{T} , we retain only the top-15 most frequent topics to make the annotation feasible. Our experts eliminated one, two, and zero topics as duplicates in the three options (Table 8). On average, the coherence scores—ranging from 1 (low) to 5 (high)—are high. This suggests that these topics were suitable for assignment to the arguments stemming from the crowd-extracted arguments. Table 9 shows the final list of topics, with low-scoring topics removed.

Option	$ \mathcal{T} $	Number of duplicates	Kept	Average rating
YOUNG	59	1	12	4.4
IMMUNE	56	2	12	4.4
REOPEN	72	0	14	4.0

Table 8: Expert topic generation statistics in Phase 1.

5.3 Phase 2: Key Argument Consolidation

In Phase 2, HyEnA uses the POWER algorithm to guide human annotations on arguments similarity, with the intent of creating clusters of similar arguments across all arguments individually annotated in Phase 1. Table 6 (right side) shows the benefit of the POWER algorithm—the number of pairs requiring human annotation (Δ) was on average reduced by 60%. The transitivity scores τ (Newman et al., 2002) indicate the extent to which transitivity holds among the similarity labels of argument pairs. The relatively low τ scores justify the subsequent clustering we perform.

Option	Topics		
YOUNG	immune entertainment hospitality restrictions	infection immunity risk in- fected	virus susceptible spread transmit
	schools reopen education students	risk limited low dangerous	group risk target least
	enforcement rules enforce measures	pressure care effect in- crease	distance keep difficult keeping
	exercise sports play exer- cising	easing relaxation options effects	mental development life psychological
IMMUNE	homes nursing care vulner- able	netherlands country provinces dutch	risk contamination danger dangerous
	work companies home economy	entertainment hospitality catering industry	infect sick cant anymore
	virus spread transmit young	testing test immunity anti- bodies	pressure care healthcare in- crease
	young distance age 15	exceptions rules enforce- ment same	easing options relaxations step
REOPEN	homes nursing care vulner- able	netherlands friesland groningen dutch	risk hospitality entertain- ment dangerous
	mental health income de- crease	measures rules these should	entrepreneurs work work- ing industry
	care healthcare psychologi- cal decrease	life economy social enter- tainment	relaxation easing options relax
	corona immune had immu- nity	bankrupt companies bankruptcies hospitality	infections increase risk out- break
	economy needs soon now	meters distance possible industry	

Table 9: Top 4 words per topic generated in Phase 1.

Figure 5 compares Louvain and spectral clustering for extracting argument clusters. Generally, both methods show a clear minimum for obtaining the final argument clusters. Louvain clustering yields the smallest error for the YOUNG and IMMUNE corpora, and spectral clustering for REOPEN corpus. These methods create 20, 14, and 18 clusters respectively. We pick these clusters as input to the argument selection phase.

Not all arguments inside the same cluster are constrained to have the same stance (pro or con) towards the policy option. We count what proportion of arguments in the cluster do not adhere to the majority stance. The distribution of stances scores is visualized in Figure 6. While we see that the upper limit is that half the arguments in each cluster are not agreeing with the majority label, the average ratio denotes that only a small fraction of argument stances do not agree with the majority stance label. This shows that the clusters generally represent a coherent distribution of arguments with similar stances to each policy

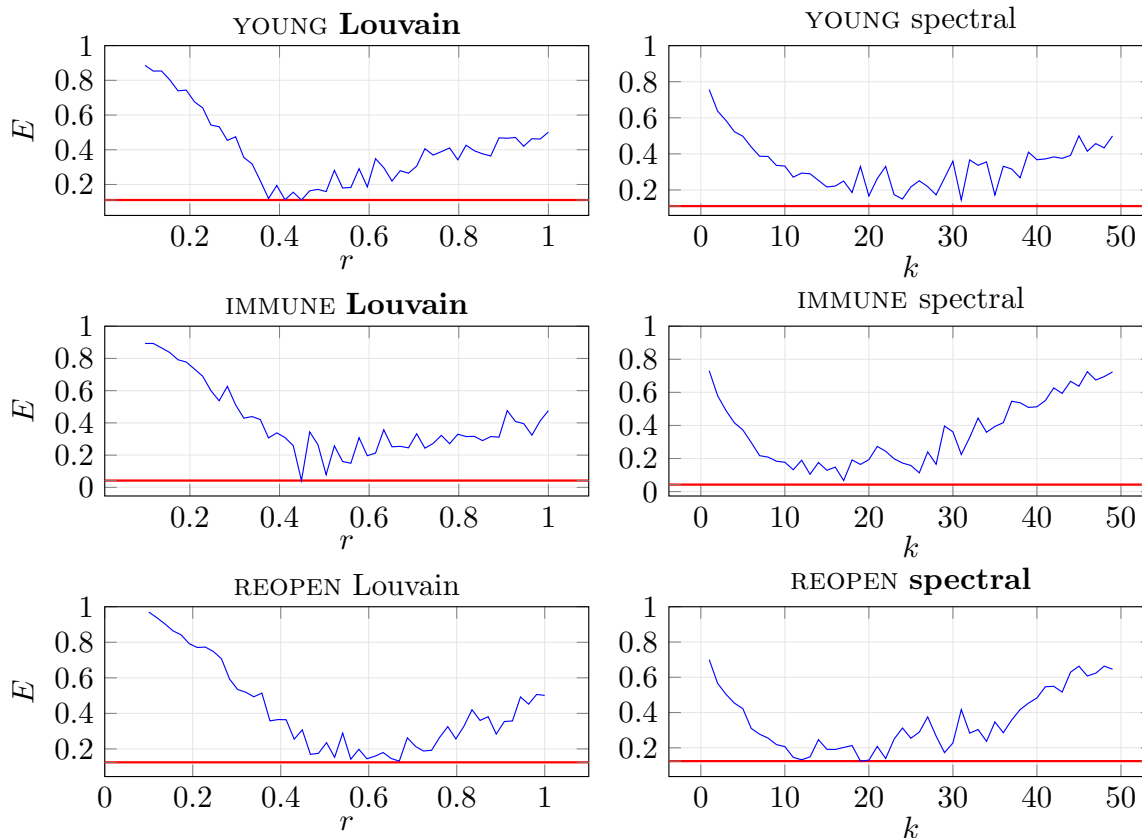


Figure 5: Error rate E for different parameters per clustering method for each subcorpus in Phase 2.

option. The ratio on average is lowest for IMMUNE, which is the option with the highest ratio of con opinions.

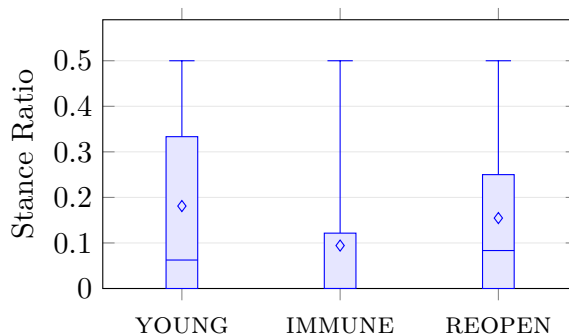


Figure 6: Stance distribution for clusters extracted for each corpus in Phase 2. A ratio of 0.5 denotes an equal number of pro and con arguments inside a cluster.

5.4 Phase 3: Key Argument Selection

In Phase 3, we compare five Argument Selection methods for extracting a representative argument for each of the clusters obtained in Phase 2. We first perform an odd-one-out task to evaluate the coherence of the clusters according to each tested Argument Selection method (see Section 4.3 for additional details). Then, we evaluate the quality of the arguments that are selected to represent clusters.

Odd-one-out task Figure 7 shows the results of the odd-one-out evaluation. Here, we see that two methods outperform the random baseline. This indicates that these methods identify cluster membership relatively consistently with the results of HyEnA, although with considerable error rates. For Llama, we encountered a strong position bias with respect to the ordering of the triple: independently of which was the odd-one-out argument, the model primarily picks arguments at a specific index. This causes its performance to be similar to random picking. We attribute this to the lack of instruction tuning for the Llama model.

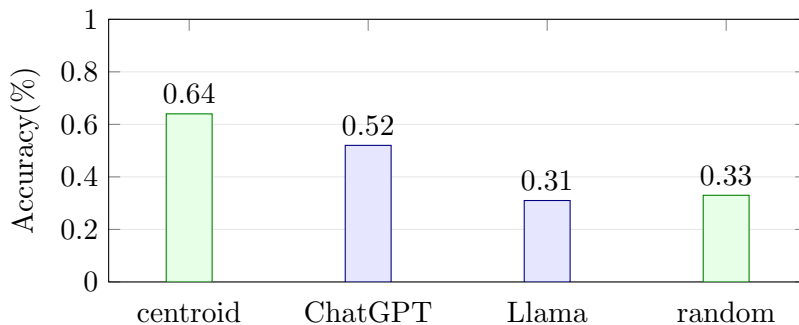


Figure 7: Accuracy on the odd-one-out task per method.

Evaluating Argument Selection To select the best-performing Argument Selection method, we compare BERTscores in Figure 8. We observe that all extractive methods have a higher standard deviation when compared to the generative methods. The high maxima for extractive methods are likely caused by the fact that some selected representative arguments were verbatim copied from an opinion in the corpus. The low minima are instead due to the fact that the extractive methods are required to use an argument from the cluster, which may not be representative of the cluster (since there may be noise in the clusters, see Figure 7). Conversely, the lower bound for the abstractive methods is higher, showing how rephrasing the selected argument allows it to be more related to all arguments inside a cluster. Between the abstractive methods, ChatGPT achieves a better score than Llama on average but has a higher standard deviation as well. Since we did not perform extensive prompt engineering, there is room for improvement in both methods with better-crafted prompts.

All methods, except for Llama, achieve similar average BERTscores. Surprisingly, none of the approaches on average performs considerably better than random. This suggests that selecting a representative argument from the cluster is relatively simple—in practice, since the argument clusters are coherent, a random selection is generally sufficient. However, in the final evaluation, humans will be judging the match between selected arguments and

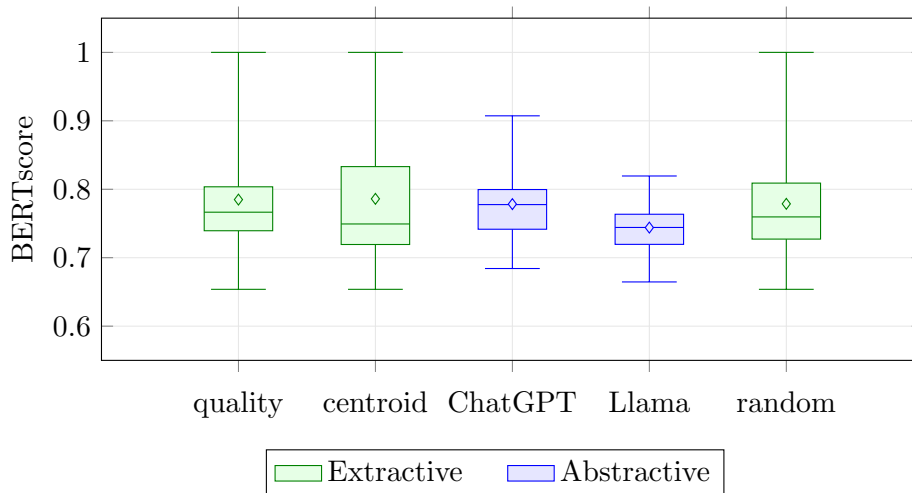


Figure 8: Aggregated BERTscore for the different Key Argument Selection methods across all corpora and argument clusters (Phase 3).

individual opinions. In this case, we strive for a better worst-case performance—we care less about having perfect matches, but rather wish to have fewer misrepresentation errors. Thus, given the comparable averages, we opt for the method with the highest lower boundary and higher media score (ChatGPT), which we use for the remainder of the experiments.

Finally, we compare the output of Phase 3 of HyEnA against a version where the selection was manual. In particular, we take the extractions from Phase 1 and re-evaluate them using a new set of annotators. In Table 10, we show the difference in Precision (Equation 8).

We find that the addition of Argument Selection has some impact on the ability of annotators to match opinions and arguments, although, generally, the decrease in performance is minor. Most interestingly, when comparing argument matches for the same set of opinions before and after the addition of Argument Selection, we find that there is only fair agreement between the re-matched labels (Cohen $\kappa = 0.255$). This indicates that the argument selection phase causes some opinions easier to be matched to selected key arguments while making others more difficult. ChatGPT generated as a selected argument a key argument that is representative of the cluster, which sometimes is more general than the arguments extracted from individual opinions. As such, for these opinions, external annotators do not recognize the specific argument being present, but some others are now captured in the selected argument.

5.5 Comparison with Automated Baseline

Figure 9 compares the coverage, precision, and diversity scores 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

Method	YOUNG	IMMUNE	REOPEN	Overall
HyEnA	0.816	0.833	0.641	0.765
HyEnA w/o Phase 3	0.787	0.848	0.739	0.789

Table 10: Comparing Precision (P) scores with and without Phase 3 (Key Argument Selection phase).

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. Since the set of *all* observed opinions is the entire corpus for ArgKP, the repeated arguments inflate its coverage. However, since HyEnA is designed to observe only a small subset of diverse opinions from the corpus, the repeated arguments do not influence its coverage significantly. This is corroborated in the diversity scores, where we observe HyEnA to consistently output a set of arguments that is more diverse than the ones produced by ArgKP.

In addition to comparing coverage over *observed* opinions, 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. ArgKP often fails to recognize the key arguments in the diverse set of opinions included by HyEnA.

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).

5.6 Comparison with Manual Baseline

Table 11 shows counts of 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, compared to the estimated 3,000 opinions seen in the manual baseline. Despite the lower human effort, the HyEnA lists largely overlap with the expert lists.

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

Table 11: A confusion matrix comparing the key argument lists generated by HyEnA and manual baseline.

HyEnA missed some key arguments that the experts identified, e.g., a key argument about building herd immunity was not in the HyEnA list for the REOPEN option. We

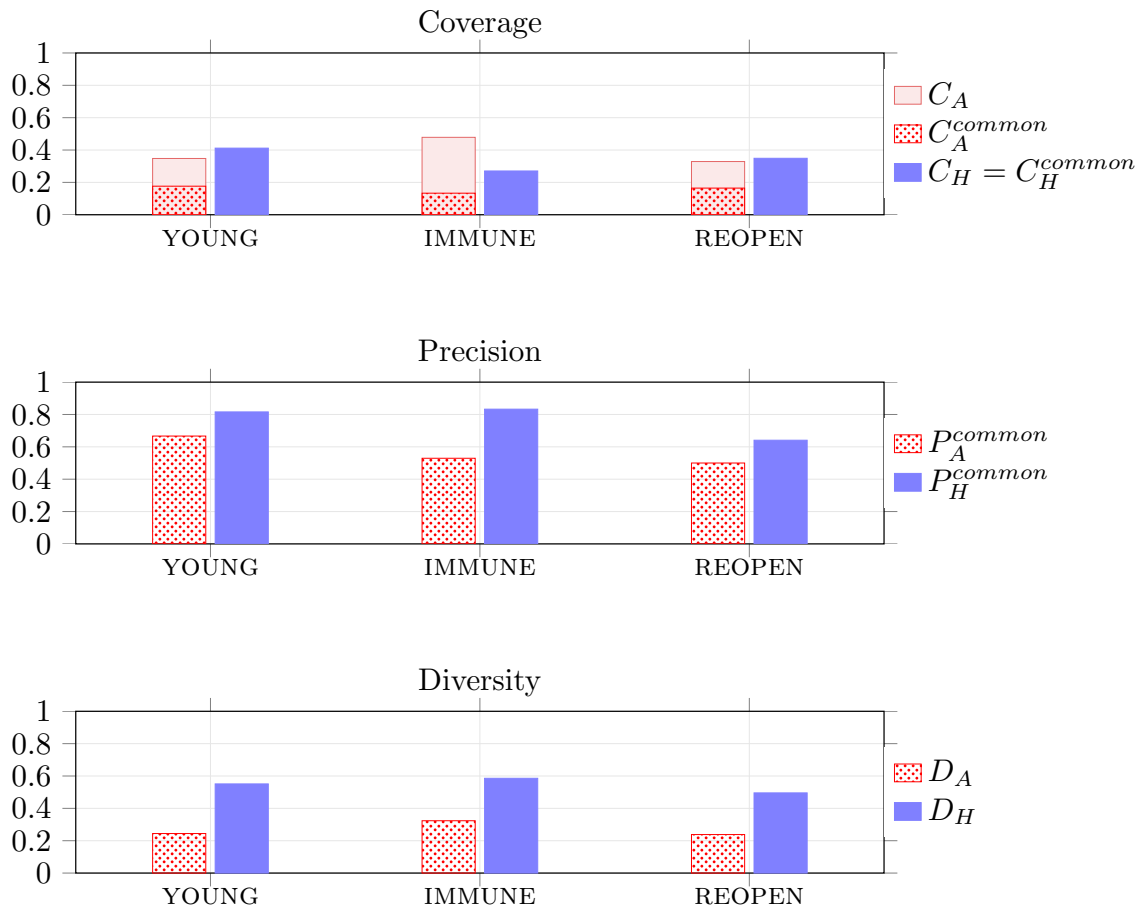


Figure 9: Comparing HyEnA and ArgKP.

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, e.g., an argument about the physical well-being of young people was not on the expert list for the YOUNG option. Likely, the larger (random) sample of opinions experts analyzed did not include opinions supporting this argument, whereas the smaller (intelligently selected) set sampled in HyEnA did.

6. Discussion

We find that HyEnA exploits the strengths of automated methods and the insights from human annotation. HyEnA outperformed an automated KPA model in terms of precision and diversity, and on a diverse set of opinions, is able to capture more nuanced arguments. Further, HyEnA expanded beyond an expert analysis, showing how a fully manual procedure may also be limited. In the remainder of this section, we expand on three specific aspects.

Limitations Our experimental setup and comparisons are limited in their scope in multiple ways, thus making our conclusions hard to generalize. Our choice of baseline is the

ArgKP model, which was optimized for the task of extracting Key Arguments from a corpus of opinions. However, other automated baselines are conceivable, especially with the introduction of the current generation of flexible LLMs (e.g., ChatGPT, Bard). Those models may be employed for KPA by using prompting techniques (Liu et al., 2023). The capabilities of these models seem to imply that they have access to higher order argumentation knowledge (Lauscher et al., 2022), and thus would fare better than the basic ArgKP model. However, having such LLMs reliably process large amounts of citizen feedback without hallucinations is a nontrivial task. In this process, due diligence to preserve a variety of perspectives is required (e.g., by optimizing for a range of opinions instead of single-annotator labels (Bakker et al., 2022)) in order to prevent rampant misrepresentation of marginalized demographics.

Instead of relying solely on the judgment of an LLM for the task of KPA, we opted to include one in the final step of HyEnA. While some of the criticism for using an LLM for end-to-end KPA still holds for the Argument Selection step as well, our method investigated a more controlled setup, supported by an objective task definition. Through our comparisons with random and human-generated labels, we aim to show where, how, and to what extent LLMs may aid in the KPA process. As ever, the choice of metrics remains important for measuring the effect size.

Balancing Task Allocation 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, partly evidenced by the low IRR. While HyEnA reduced the number of comparisons required in the consolidation phase by 60%, we may experiment with different setups or other techniques for comparing arguments to remove this overhead. For example, first clustering the key arguments and then consolidating the arguments within these clusters (reverse order as HyEnA) may drastically reduce the number of judgments required in the second phase.

We place human efforts in places where there are multiple bidirectional benefits possible stemming from performing the task. For instance, the Argument Annotation task both serves the purpose of analyzing the opinions to progress our method, as well as actively making annotators perform *perspective-taking*. On multiple occasions, annotators noted on their increase in sympathy and recognition of the issues raised in the comments, showcasing how the task could further help bring understanding to a group of citizens.

Ablations studies Any part of the HyEnA pipeline is open to adjustment and can be performed by humans, machines, or a combination of both. HyEnA presents a general framework that allows individual phases to be supported by different types of technologies. Within this hybrid framework, we considered the following criteria when deciding to allocate tasks to humans or to AI methods: (1) let humans read other’s opinions to promote perspective-taking, (2) use humans to solve tasks where AI methods may incur considerable error, (3) leverage AI methods for routine tasks, and (4) use task-specific intrinsic evaluation metrics for selecting the right method.

In each phase, we perform both intrinsic evaluations (e.g., observe error rates for particular tasks or annotator behavior) and extrinsic evaluation against two baselines. This fits a standardized machine learning pipeline, except that we are now able to (1) evaluate annotator behavior and model performance jointly, and (2) make decisions on which techniques

to use based on some intermediate statistic. We believe this setup to be generalizable for Hybrid Intelligence systems, as it makes the role of the designer and their decisions explicit (Akata et al., 2020). Furthermore, the results remain interpretable, as any decision made by either annotators or models can be traced from opinion to selected key argument.

Different configurations of the HyEnA framework are possible, and the one we have presented is an instance that tackles the problem of policy feedback analysis. HyEnA is a complex combination of AI methods and human annotation. Our main objective was to present the HyEnA framework, as well as a real-world use case to show the benefit of using a Hybrid Intelligent methodology. However, other choices for individual components of HyEnA can be used, or parts of the method can be performed solely by humans or AI methods. We leave this open for future work, as running ablations requires considerable amounts of work, and we envision research to come up with similar use cases where HI can make a significant impact.

7. Conclusion and Future Directions

We develop and evaluate HyEnA, a hybrid method that combines human judgments 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 drops by 50% when switching from all (containing 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.

Finding arguments in a discourse is only one aspect that constitutes the perspectives in a discussion. Future work can incorporate analysis of other perspective factors, such as values (Liscio et al., 2021, 2022), 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.

Acknowledgements This research was (partially) funded by the Hybrid Intelligence Center, a 10-year program funded by the Dutch Ministry of Education, Culture, and Science through the Netherlands Organisation for Scientific Research.

References

- Akata, Z., Balliet, D., de Rijke, M., Dignum, F., Dignum, V., Eiben, G., Fokkens, A., Grossi, D., Hindriks, K., Hoos, H., Hung, H., Jonker, C., Monz, C., Neerincx, M., Oliehoek, F., Prakken, H., Schlobach, S., van der Gaag, L., van Harmelen, F., van Hoof, H., van Riemsdijk, B., van Wynsberghe, A., Verbrugge, R., Verheij, B., Vossen, P., & Welling, M. (2020). A research agenda for hybrid intelligence: Augmenting human intellect with collaborative, adaptive, responsible, and explainable artificial intelligence. *Computer*, 53(8), 18–28.

- Aroyo, L., & Welty, C. (2015). Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1), 15–24.
- Bakker, M., Chadwick, M., Sheahan, H., Tessler, M., Campbell-Gillingham, L., Balaguer, J., McAleese, N., Glaese, A., Aslanides, J., Botvinick, M., et al. (2022). Fine-tuning language models to find agreement among humans with diverse preferences. *Advances in Neural Information Processing Systems*, 35, 38176–38189.
- Bar-Haim, R., Eden, L., Friedman, R., Kantor, Y., Lahav, D., & Slonim, N. (2020a). From arguments to key points: Towards automatic argument summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4029–4039, Online. Association for Computational Linguistics.
- Bar-Haim, R., Kantor, Y., Eden, L., Friedman, R., Lahav, D., & Slonim, N. (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)*, pp. 39–49, Online. Association for Computational Linguistics.
- Basu, S., Banerjee, A., & Mooney, R. J. (2004). *Active Semi-Supervision for Pairwise Constrained Clustering*, chap. 32, pp. 333–344. Society for Industrial and Applied Mathematics.
- Beck, T., Lee, J.-U., Viehmann, C., Maurer, M., Quiring, O., & Gurevych, I. (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)*, pp. 1–13, Online. Association for Computational Linguistics.
- Bentahar, J., Moulin, B., & Bélanger, M. (2010). A taxonomy of argumentation models used for knowledge representation. *The Artificial Intelligence Review*, 33(3), 211.
- Biran, O., & Rambow, O. (2011). Identifying justifications in written dialogs. In *2011 IEEE Fifth International Conference on Semantic Computing*, pp. 162–168.
- Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10), P10008.
- Boltužić, F., & Šnajder, J. (2015). Identifying prominent arguments in online debates using semantic textual similarity. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pp. 110–115.
- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33, 1877–1901.
- Cabrio, E., & Villata, S. (2018). Five years of argument mining: A data-driven analysis. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence, IJCAI’18*, p. 5427–5433. AAAI Press.
- Chai, C., Li, G., Li, J., Deng, D., & Feng, J. (2016). Cost-effective crowdsourced entity resolution: A partial-order approach. In *Proceedings of the 2016 International Conference*

on Management of Data, SIGMOD '16, p. 969–984, New York, NY, USA. Association for Computing Machinery.

- Conneau, A., Kiela, D., Schwenk, H., Barrault, L., & Bordes, A. (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*, pp. 670–680, Copenhagen, Denmark. Association for Computational Linguistics.
- Daxenberger, J., Eger, S., Habernal, I., Stab, C., & Gurevych, I. (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*, pp. 2055–2066, Copenhagen, Denmark. Association for Computational Linguistics.
- Daxenberger, J., Schiller, B., Stahlhut, C., Kaiser, E., & Gurevych, I. (2020). Argumentext: argument classification and clustering in a generalized search scenario. *Datenbank-Spektrum*, 20, 115–121.
- Daza, A., & Frank, A. (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)*, pp. 3904–3914, Online. Association for Computational Linguistics.
- Eger, S., Daxenberger, J., & Gurevych, I. (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)*, pp. 11–22, Vancouver, Canada. Association for Computational Linguistics.
- Eger, S., Daxenberger, J., Stab, C., & Gurevych, I. (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*, pp. 831–844, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- Ein-Dor, L., Shnarch, E., Dankin, L., Halfon, A., Sznajder, B., Gera, A., Alzate, C., Gleize, M., Choshen, L., Hou, Y., Bilu, Y., Aharonov, R., & Slonim, N. (2020). Corpus wide argument mining—a working solution. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05), 7683–7691.
- Fabbri, A. R., Kryściński, W., McCann, B., Xiong, C., Socher, R., & Radev, D. (2021). Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9, 391–409.
- Freeman, J. B. (2011). *Argument structure*: (2011 edition). Argumentation Library. Springer, Dordrecht, Netherlands.
- Gretz, S., Friedman, R., Cohen-Karlik, E., Toledo, A., Lahav, D., Aharonov, R., & Slonim, N. (2020). A large-scale dataset for argument quality ranking: Construction and analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05), 7805–7813.
- Grootendorst, M. (2022). Bertopic: Neural topic modeling with a class-based tf-idf procedure..
- Habernal, I., & Gurevych, I. (2017). Argumentation mining in user-generated web discourse. *Computational Linguistics*, 43(1), 125–179.

- Howard, J., & Ruder, S. (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)*, pp. 328–339, Melbourne, Australia. Association for Computational Linguistics.
- Klein, M. (2012). Enabling large-scale deliberation using attention-mediation metrics. *Computer Supported Cooperative Work (CSCW)*, 21(4-5), 449–473.
- Kythreotis, A. P., Mantyka-Pringle, C., Mercer, T. G., Whitmarsh, L. E., Corner, A., Paavola, J., Chambers, C., Miller, B. A., & Castree, N. (2019). Citizen social science for more integrative and effective climate action: A science-policy perspective. *Frontiers in Environmental Science*, 7.
- Lauscher, A., Wachsmuth, H., Gurevych, I., & Glavaš, G. (2022). Scientia potentia est—on the role of knowledge in computational argumentation. *Transactions of the Association for Computational Linguistics*, 10, 1392–1422.
- Lawrence, J., & Reed, C. (2020). Argument Mining: A Survey. *Computational Linguistics*, 45(4), 765–818.
- Lee, S., Hwang, C., & Moon, M. J. (2020). Policy learning and crisis policy-making: quadruple-loop learning and covid-19 responses in south korea. *Policy and Society*, 39(3), 363–381. PMID: 35039726.
- Liscio, E., van der Meer, M., Siebert, L. C., Jonker, C. M., Mouter, N., & Murukannaiah, P. K. (2021). Axes: Identifying and evaluating context-specific values. In *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS '21*, p. 799–808, Richland, SC. International Foundation for Autonomous Agents and Multiagent Systems.
- Liscio, E., van der Meer, M., Siebert, L. C., Jonker, C. M., & Murukannaiah, P. K. (2022). What values should an agent align with? an empirical comparison of general and context-specific values. *Autonomous Agents and Multi-Agent Systems*, 36(1), 23.
- Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9), 1–35.
- Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 1(1).
- McCoy, T., Pavlick, E., & Linzen, T. (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*, pp. 3428–3448, Florence, Italy. Association for Computational Linguistics.
- Mouter, N., Hernandez, J. I., & Itten, A. V. (2021). Public participation in crisis policymaking. how 30,000 dutch citizens advised their government on relaxing covid-19 lockdown measures. *PLOS ONE*, 16(5), 1–42.
- Newman, M. E. J., Watts, D. J., & Strogatz, S. H. (2002). Random graph models of social networks. *Proceedings of the National Academy of Sciences*, 99(suppl.1), 2566–2572.

- Niculescu, V., Park, J., & Cardie, C. (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)*, pp. 985–995, Vancouver, Canada. Association for Computational Linguistics.
- Ostrom, E. (1990). *Governing the commons: The evolution of institutions for collective action*. Cambridge University Press, Cambridge, UK.
- Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35, 27730–27744.
- Palau, R. M., & Moens, M.-F. (2009). Argumentation mining: the detection, classification and structure of arguments in text. In *Proceedings of the 12th international conference on artificial intelligence and law*, pp. 98–107.
- Price, V. (1989). Social Identification and Public Opinion: Effects of Communicating Group Conflict. *Public Opinion Quarterly*, 53(2), 197–224.
- Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992, Hong Kong, China. Association for Computational Linguistics.
- Reimers, N., Schiller, B., Beck, T., Daxenberger, J., Stab, C., & Gurevych, I. (2019). Classification and clustering of arguments with contextualized word embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 567–578, Florence, Italy. Association for Computational Linguistics.
- Schulz, C., Meyer, C. M., Kiesewetter, J., Sailer, M., Bauer, E., Fischer, M. R., Fischer, F., & Gurevych, I. (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*, pp. 2761–2772, Florence, Italy. Association for Computational Linguistics.
- Schulz-Hardt, S., Frey, D., Lüthgens, C., & Moscovici, S. (2000). Biased information search in group decision making. *Journal of Personality and Social Psychology*, 78(4), 655–669.
- Sennrich, R., Haddow, B., & Birch, A. (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)*, pp. 86–96, Berlin, Germany. Association for Computational Linguistics.
- Shortall, R., Itten, A., van der Meer, M., Murukannaiah, P., & Jonker, C. (2022). Reason against the machine? future directions for mass online deliberation. *Frontiers in Political Science*, 4, 01–17.
- Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin*, 86(2), 420.

- Sim, J., & Wright, C. C. (2005). The kappa statistic in reliability studies: use, interpretation, and sample size requirements. *Physical therapy, 85*(3), 257–268.
- Stab, C., & Gurevych, I. (2014). Annotating argument components and relations in persuasive essays. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pp. 1501–1510, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.
- Stab, C., Miller, T., Schiller, B., Rai, P., & Gurevych, I. (2018). Cross-topic argument mining from heterogeneous sources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 3664–3674, Brussels, Belgium. Association for Computational Linguistics.
- Talat, Z., Blix, H., Valvoda, J., Ganesh, M. I., Cotterell, R., & Williams, A. (2022). On the Machine Learning of Ethical Judgments from Natural Language. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL '22*, pp. 769–779, Seattle, USA.
- Teruel, M., Cardellino, C., Cardellino, F., Alonso Alemany, L., & Villata, S. (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)*, pp. 4061–4064, Miyazaki, Japan. European Language Resources Association (ELRA).
- Thorn Jakobsen, T. S., Barrett, M., & Søgaaard, A. (2021). Spurious correlations in cross-topic argument mining. In *Proceedings of *SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics*, pp. 263–277, Online. Association for Computational Linguistics.
- Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. (2023). Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 0(0).
- van der Meer, M., Liscio, E., Jonker, C. M., Plaat, A., Vossen, P., & Murukannaiah, P. K. (2022a). Hyena: A hybrid method for extracting arguments from opinions. In *Proceedings of the first International Conference on Hybrid Human-Artificial Intelligence (HHAI 2022)*, pp. 17–31, Amsterdam, the Netherlands. IOS Press.
- van der Meer, M., Liscio, E., Jonker, C. M., Plaat, A., Vossen, P., & Murukannaiah, P. K. (2022b). Hyena: A hybrid method for extracting arguments from opinions: Supplementary material. <https://liacs.leidenuniv.nl/~meermtvander/publications/hyena/>.
- van Eemeren, F. H., Grootendorst, R., & Kruiger, T. (1987). *Handbook of Argumentation Theory*. De Gruyter Mouton, Berlin, Boston.
- van Son, C., Caselli, T., Fokkens, A., Maks, I., Morante, R., Aroyo, L., & Vossen, P. (2016). GRaSP: A multilayered annotation scheme for perspectives. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pp. 1177–1184, Portorož, Slovenia. European Language Resources Association (ELRA).

- Wachsmuth, H., Syed, S., & Stein, B. (2018). Retrieval of the best counterargument without prior topic knowledge. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 241–251.
- Walton, D., Reed, C., & Macagno, F. (2008). *Argumentation schemes*. Cambridge University Press.
- Wang, J., Hu, X., Hou, W., Chen, H., Zheng, R., Wang, Y., Yang, L., Huang, H., Ye, W., Geng, X., et al. (2023). On the robustness of chatgpt: An adversarial and out-of-distribution perspective. *arXiv preprint arXiv:2302.12095, 2023*, 1–19.
- Xu, J., He, X., & Li, H. (2020). Deep learning for matching in search and recommendation. *Foundations and Trends® in Information Retrieval*, 14(2–3), 102–288.
- Zelnik-Manor, L., & Perona, P. (2004). Self-tuning spectral clustering. In *Proceedings of the 17th International Conference on Neural Information Processing Systems, NIPS’04*, p. 1601–1608, Cambridge, MA, USA. MIT Press.
- Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2020). Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.
- Zhong, M., Liu, P., Chen, Y., Wang, D., Qiu, X., & Huang, X. (2020). Extractive summarization as text matching. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 6197–6208, Online. Association for Computational Linguistics.