



Universiteit
Leiden
The Netherlands

Opinion diversity through hybrid intelligence

Meer, M.T. van der

Citation

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

Version: Publisher's Version

License: [Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden](#)

Downloaded from: <https://hdl.handle.net/1887/4209024>

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

Opinion Diversity through Hybrid Intelligence

Michiel Theo VAN DER MEER

Opinion Diversity through Hybrid Intelligence

Proefschrift

ter verkrijging van
de graad van doctor aan de Universiteit Leiden,
op gezag van rector magnificus prof.dr.ir. H. Bijl,
volgens besluit van het college voor promoties
te verdedigen op woensdag 26 maart 2025
klokke 11:30 uur

door

Michiel Theo van der Meer

geboren te Groningen, Nederland
in 1995

Promotores:

prof.dr. C.M. Jonker
prof.dr. P.T.J.M. Vossen

Universiteit Leiden, Technische Universiteit Delft
Vrije Universiteit Amsterdam

Copromotor:

dr. P.K. Murukannaiah

Technische Universiteit Delft

Promotiecommissie:

prof.dr. A. Plaat
prof.dr. S. Verberne
prof.dr. M.M. Bonsangue
prof.dr. A. Fokkens
prof.dr. D. Grossi
prof.dr.ing. S. Kopp

Vrije Universiteit Amsterdam
Rijksuniversiteit Groningen, Universiteit van Amsterdam
Universität Bielefeld



Universiteit
Leiden



Nederlandse Organisatie voor Wetenschappelijk Onderzoek



Keywords: Natural Language Processing, Hybrid Intelligence, Perspectives, Online Deliberation

Printed by: Proefschriftspecialist: <https://www.proefschriftspecialist.nl>

Cover by: Arwen Rosenberg-Meereboer

ISBN 978-94-93431-16-4

This dissertation is available online at <https://scholarlypublications.universiteitleiden.nl/>

The work in this dissertation was funded by the Netherlands Organisation for Scientific Research (NWO) through the Hybrid Intelligence Centre via the Zwaartekracht grant (024.004.022).

SIKS Dissertation Series No. 2025-15.

The research reported in this thesis has been carried out under the auspices of SIKS, the Dutch Research School for Information and Knowledge Systems.

“Houden van het anders zijn van de ander.”
Theodorus Aris Pieter (Theo) Meereboer 1952-2020, Groningen.

Summary

Problem In a representative democracy, citizens elect representatives to act in their interest for shaping public policy. Modern democracies face a critical issue of declining citizen participation, leading to a disconnect between citizens and their elected officials. Deliberative democracy, which emphasizes open dialogue and encourages wider participation, is one way to address this issue. However, traditional in-person deliberation methods face challenges, such as including participants with a diverse set of perspectives and backgrounds, and ensuring all voices in a discussion are considered equally. Online social media platforms offer an alternative venue for large-scale deliberation, allowing for discussions with a wider audience and rapid access to information. However, concerns exist about whether these platforms can foster truly inclusive and diverse discussions. For instance, locating contributions of relevant perspectives can be difficult due to the large amounts of scattered content. Further, interactions on platforms lead to echo chambers that drive polarization, threatening the egalitarian basis of the discussions. Artificial Intelligence (AI), and Natural Language Processing (NLP) in particular, for facilitating text-based online discussions have become attractive as a solution. However, its impact on the diversity of perspectives in online deliberation is unexplored.

Methods This dissertation (1) identifies the challenges involved in facilitating large-scale online discussions with NLP, (2) suggests solutions to these challenges by incorporating hybrid human–AI technologies, and (3) investigates what these technologies can reveal about individual perspectives in online discussions. We propose a three-layered hierarchy for representing perspectives that can be obtained by a mixture of human intelligence and Large Language Models (LLMs). This combination is known as Hybrid Intelligence (HI). We illustrate how these representations can draw insights into the diversity of perspectives and allow us to investigate interactions in online discussions.

- In Part I of this dissertation, we show that existing opinion analysis methods, particularly those involving LLMs, are limited in understanding the perspectives expressed by minorities. Nonetheless, the models' capabilities of processing text-based data at scale make them attractive in analyzing online discussions. Despite the complexity of understanding free-form opinionated texts, we can effectively use models in low-resource settings. In particular, LLMs can address abstract tasks fluently and interpolate missing information. However, the sensitivity of, e.g., zero-shot prompting procedures for LLMs underscores how they still need human oversight to perform well across contexts. Further, LLMs behave differently from humans, failing to align with human disagreement and making errors different from us, thus necessitating careful supervision.
- In Part II, we harness the potential of HI systems for opinion analysis. By strategically incorporating human input with LLMs and fostering a back-and-forth process between

humans and AI, HI systems can be designed to capture diverse opinions precisely and efficiently. HI requires careful task allocation and balancing. Our methods use human annotators to provide a nuanced understanding of e.g. arguments while using NLP techniques for sampling interesting opinions from a large dataset. Through repeated interaction, both sides continuously adjust and learn from each other. Our hybrid setup paves the way for a future where humans and NLP technology can join forces to cultivate a deeper understanding of the multifaceted nature of online discussions.

- In Part III, we show that HI systems can extract an individual's *Perspective Hierarchy* based on the tasks used in Part I. To extract the perspective hierarchy, we leverage the complementary abilities of humans and NLP models. We show how arguments, next to stances and personal values, are a core component of the hierarchy by experimenting with extracting a direct relation between values and stance.

Findings We find that there are fundamental issues to fostering diversity when analyzing online discussions on social media platforms. These issues include: (1) ensuring that minority and marginalized voices are participating on the platforms, (2) an emphasis on frequently repeated opinions that fail to bridge political divides, and (3) an aggravation of this problem by the straightforward application of LLMs for opinion analysis. HI can help alleviate these problems. In our approach to HI, we encourage more explicit communication between humans through repeated interaction with LLMs. All of this feeds improvements on two ends: humans benefit from explicit communication and closely considering each other's point of view, while the rationales provided by them are useful resources for AI to learn from.

In this dissertation, we provide one of the first demonstrations of how HI can be used to integrate humans and AI, mixing human collaborative capacity with LLMs. Nonetheless, showing that HI leads to improved and diverse discussions remains difficult. Existing evaluation paradigms are insufficient for measuring how HI leads to improvements over AI-only or manual approaches, indicating the need for more dynamic and context-sensitive evaluation approaches.

Samenvatting

Probleem In een representatieve democratie kiezen burgers vertegenwoordigers die in hun belang handelen bij het vormgeven van overheidsbeleid. Moderne democratieën kampen met een kritiek probleem van afnemende burgerparticipatie, wat leidt tot een kloof tussen burgers en hun gekozen vertegenwoordigers. Een mogelijke aanpak voor dit probleem is via een deliberatieve democratie, waarin open dialogen en wijdere participatie wordt aangemoedigd. Traditionele fysieke deliberatiemethoden kennen echter aanzienlijke uitdagingen, zoals het betrekken van deelnemers met verschillende perspectieven en achtergronden, en het waarborgen van gelijkwaardige inspraak in de discussies. Als alternatief bieden online social-mediaplatformen een mogelijkheid voor grootschalige deliberatie, omdat het discussies met een groter publiek en met snelle toegang tot informatie mogelijk maakt. Er bestaan echter zorgen over de vraag of deze platformen inclusieve en diverse discussies kunnen bevorderen. Het kan bijvoorbeeld moeilijk zijn om bijdragen van relevante perspectieven te vinden vanwege de grote hoeveelheden versnipperde informatie. Bovendien kunnen interacties op platformen tot echokamers leiden die polarisatie opdrijven, wat vervolgens de egalitaire basis van de discussies bedreigt. Het gebruik van Kunstmatige Intelligentie (KI), en met name Natural Language Processing (NLP), voor het faciliteren van online tekstgebaseerde discussies is aantrekkelijk geworden als oplossing hiervoor. De impact van deze technologie op de diversiteit van perspectieven bij online deliberaties is tot dusver nog niet onderzocht.

Methoden Dit proefschrift (1) identificeert de uitdagingen die komen kijken bij het faciliteren van online discussies op grote schaal met behulp van NLP, (2) suggereert oplossingen voor deze uitdagingen door het bewerkstelligen van hybride mens-KI-technologie, en (3) onderzoekt wat deze technologieën kunnen ophelderen over individuele perspectieven in online discussies. Om perspectieven te representeren stellen wij een drielaagse hiërarchie voor die kan worden verkregen door een combinatie van menselijke intelligentie en Large Language Models (LLM's). Deze combinatie noemen wordt ook wel Hybride Intelligentie (HI) genoemd. We illustreren hoe deze representaties ons inzicht kunnen bieden in de diversiteit van perspectieven, en hoe ze ons toestaan interacties in online discussies te karakteriseren.

- In Deel I van dit proefschrift laten we zien hoe bestaande methoden voor het analyseren van meningen, specifiek wanneer ze gebruikmaken van LLM's, beperkt zijn in het begrijpen van de perspectieven die geuit worden door minderheden. Desalniettemin maakt de kracht van deze modellen om tekstgebaseerde data op grote schaal te verwerken ze aantrekkelijk voor het analyseren van online discussies. We deze modellen effectief toepassen wanneer we weinig brondata tot onze beschikking hebben ondanks de complexiteit van het interpreteren van ongestructureerde meningen. LLM's zijn, in het bijzonder, in staat om abstracte taken effectief te voltooien en missende informatie zelfstandig aan te vullen. Hun gevoeligheid benadrukt daarentegen wel dat LLM's menselijk toezicht nodig hebben om goed te presteren in verschillende contexten. LLM's gedragen zich anders dan mensen:

ze kunnen menselijke meningsverschillen vaak niet begrijpen, en maken andere soorten fouten dan mensen. Dit maakt dat ze zorgvuldige menselijke begeleiding nodig hebben.

- In Deel II gebruiken we de potentie van HI-systemen voor het analyseren van meningen. Door middel van het strategisch combineren van menselijke input met dat van LLM's, en door het bevorderen van een heen-en-weer proces tussen mensen en KI, kunnen HI-systemen ontworpen worden die diverse meningen op een precieze en efficiënte manier kunnen vatten. HI vereist een zorgvuldige taakverdeling. Onze methoden maken gebruik van menselijke annotateerders die een genuanceerd begrip van bijvoorbeeld argumenten kunnen leveren. Tegelijkertijd kunnen we NLP-technieken gebruiken voor het selecteren van interessante meningen uit een grote dataset. Beide kanten kunnen door herhaalde interactie continu aanpassen en van elkaar leren. Onze hybride opzet baant de weg voor een toekomst waarin mensen en NLP-technologie de handen ineenslaan om een dieper begrip over de veelzijdige aard van online discussies te cultiveren.
- In Deel III laten we zien dat HI-systemen de *Perspectief Hiërarchie* van een individu kunnen extraheren, gebaseerd op de taken uit Deel I. Voor het extraheren van deze hiërarchieën gebruiken we de complementaire kracht van mensen en NLP-modellen. We tonen aan dat hoe de houding, argumenten, en persoonlijke waarden van een individu kerncomponenten zijn van de hiërarchie door een directe relatie tussen waarden en houdingen te onderzoeken.

Bevindingen We constateren dat er fundamentele problemen zijn voor het bevorderen van diversiteit in de analyse van social media platformen. Deze problemen omvatten (1) het waarborgen dat gemarginaliseerde standpunten deel nemen op de platformen, (2) een nadruk op vaak herhaalde meningen die de politieke kloof verergeren, en (3) een verergering van dit probleem door de gemakkelijke toepassing van LLM's voor het analyseren van meningen. HI kan helpen om deze problemen te verlichten. In onze aanpak voor HI moedigen wij explicietere communicatie tussen mensen aan door middel van herhaalde interactie met LLM's. Dit alles voedt verbeteringen aan twee kanten: mensen profiteren van expliciete communicatie en kunnen elkaars uitgangspunten nauwgezet overwegen terwijl hun redeneringen gebruikt kunnen worden om betere KI-modellen te trainen.

Dit proefschrift biedt een van de eerste demonstraties van de integratie van mensen en KI door het samenwerkingsvermogen van mensen te combineren met LLM's. Toch blijft het aantonen dat HI leidt tot een verbeterde en diversere discussie moeilijk. Bestaande evaluatieparadigma's voor dergelijke systemen zijn ontoereikend om te meten hoe HI leidt tot verbeteringen ten opzichte van alleen KI of handmatige aanpakken, wat de noodzaak voor meer dynamische en context-specifieke evaluaties onderstreept.

Contents

Summary	vii
Samenvatting	ix
1 Introduction	1
1.1 Research Questions	4
1.2 Research Methodology	5
1.2.1 Fundamental Issues (Q1)	5
1.2.2 Hybrid Intelligence (Q2)	6
1.2.3 Perspective Hierarchy (Q3)	9
1.3 Dissertation Scope	11
1.4 Outlook	12
I NLP for Online Discussion Analysis	15
2 An Empirical Analysis of Diversity in Argument Summarization	19
2.1 Introduction	20
2.2 Related Work	21
2.2.1 Key Point Analysis	21
2.2.2 Opinion Summarization	21
2.2.3 Diversity in Societal Decision Making.	21
2.3 Method	21
2.3.1 Task setup	22
2.3.2 Modeling Diversity in Key Point Analysis	22
2.4 Experimental Setup.	23
2.4.1 Data	23
2.4.2 Approaches.	24
2.4.3 Evaluation Metrics	25
2.5 Results and Discussion	26
2.5.1 KPG Performance	26
2.5.2 KPM Performance	26
2.5.3 Analysis	27
2.6 Conclusion	30
3 Will It Blend? Mixing Training Paradigms & Prompting for Argument Quality Prediction	33
3.1 Introduction	34
3.2 Related Work.	34
3.3 Data and Training Paradigms	35
3.3.1 Data	35

3.3.2	Training Paradigms	35
3.4	Approach.	36
3.4.1	Implementation details	37
3.5	Experiments and Results	37
3.5.1	Error Analysis	38
3.6	Conclusion	39
3.7	Access and Responsible Research	39
II	Hybrid Intelligence for NLP	41
4	A Hybrid Intelligence Method for Argument Mining	45
4.1	Introduction	46
4.2	Related work	48
4.2.1	Computational Argument Analysis	48
4.2.2	Summarization of Arguments.	49
4.3	Method	49
4.3.1	Opinion Corpora	50
4.3.2	Key Argument Annotation	51
4.3.3	Key Argument Consolidation.	52
4.3.4	Key Argument Selection	54
4.4	Experimental Setup.	54
4.4.1	Phase 1: Key Argument Annotation.	55
4.4.2	Phase 2: Key Argument Consolidation	56
4.4.3	Phase 3: Key Argument Selection	56
4.4.4	Baselines	58
4.5	Results	60
4.5.1	Annotator Agreement	60
4.5.2	Phase 1: Key Argument Annotation.	61
4.5.3	Phase 2: Key Argument Consolidation	63
4.5.4	Phase 3: Key Argument Selection	64
4.5.5	Comparison with Automated Baseline	67
4.5.6	Comparison with Manual Baseline	68
4.6	Discussion	69
4.7	Conclusion and Future Directions	71
5	Annotator-Centric Active Learning for Subjective NLP Tasks	73
5.1	Introduction	74
5.2	Related work	75
5.2.1	Learning with annotator disagreement	75
5.2.2	Active Learning	75
5.3	Method	76
5.3.1	Soft-label prediction	76
5.3.2	Active Learning	76
5.3.3	Annotator-Centric Active Learning.	77

5.4	Experimental Setup	78
5.4.1	Datasets	78
5.4.2	Evaluation metrics	78
5.4.3	Training procedure	79
5.5	Results	79
5.5.1	Highlights	79
5.5.2	Efficiency and Fairness	80
5.5.3	Convergence	83
5.5.4	Impact of subjectivity	83
5.6	Conclusion	85

III Social Science with Hybrid Intelligence 87

6	Do Differences in Values Influence Disagreements in Online Discussions?	91
6.1	Introduction	92
6.2	Related Work	93
6.2.1	(Dis)-agreement and discussion analysis	93
6.2.2	Value models	94
6.2.3	Value estimation	94
6.3	Method	94
6.3.1	Data	95
6.3.2	Value Extraction	95
6.3.3	Value Profile Estimation	96
6.4	Experiments and Results	96
6.4.1	Training Models for Value Estimation	96
6.4.2	Value Profile Estimation	97
6.4.3	Value Conflicts and Disagreement	98
6.4.4	Use Case: Predicting (Dis-)agreement	102
6.5	Conclusion	103

IV Conclusions 105

7	Contributions and Future Work	107
7.1	Research Findings	108
7.1.1	NLP for Perspective Analysis	109
7.1.2	Hybrid Intelligence for NLP	110
7.1.3	Perspective Hierarchy	112
7.2	Contributions	112
7.2.1	Scientific Relevance	112
7.2.2	Societal Relevance	114
7.3	Limitations	115
7.4	Future Work	116

V	Appendices	119
A	An Empirical Analysis of Diversity in Argument Summarization	121
A.1	Detailed Experimental Setup	121
A.1.1	Data	121
A.1.2	Per-approach Specifics	122
A.1.3	Evaluation metrics	124
A.2	Additional results.	125
A.2.1	Detailed ROUGE scores for Key Point Generation.	125
A.2.2	Additional BERTScores for Key Point Generation	125
A.2.3	Long-tail experiment for KPG	125
A.2.4	ChatGPT generated key points for PVE	126
B	Will It Blend? Mixing Training Paradigms & Prompting for Argument Quality Prediction	129
B.1	Hyperparameters.	129
B.2	Additional results.	130
B.2.1	Per-label Performance	130
B.2.2	Label confusion	130
B.2.3	Seed Variance	130
B.2.4	Topics	131
C	A Hybrid Intelligence Method for Argument Mining	133
C.1	Experiment Protocol & Description.	133
C.1.1	Preliminaries	133
C.1.2	Phase 1: Argument Annotation	133
C.1.3	Phase 2: Argument Consolidation	134
C.1.4	Comparison to Automated Baseline.	134
C.1.5	Annotation platform	134
C.2	Method Details.	135
C.2.1	Parallel Pairwise Annotation Algorithm.	135
C.2.2	Hyperparameters.	135
C.3	Detailed Results	138
C.3.1	Unclear Translation Actions	138
C.3.2	Clustering Arguments	138
C.3.3	Key Arguments.	138
D	Annotator-Centric Active Learning for Subjective NLP Tasks	149
D.1	Detailed Experimental Setup	149
D.1.1	Dataset details	149
D.1.2	Hyperparameters.	149
D.1.3	Training details.	150
D.1.4	ACAL annotator strategy details	150
D.1.5	Disagreement rates	151
D.2	Detailed results overview	152
D.2.1	Annotator-Centric evaluation for other MFTC and MHS tasks	152
D.2.2	Training process	152

E	Do Differences in Values Influence Disagreements in Online Discussions?	159
E.1	Methodological details	159
E.1.1	Training Value extraction methods	159
E.1.2	Annotator experiment	160
E.1.3	Training agreement analysis models.	162
E.2	Additional Results	165
E.2.1	Value Extraction	165
E.2.2	Value Survey	165
E.2.3	Qualitative Examples of Value Conflicts and (Dis-)agreement	165
E.2.4	Decomposition of BF_{10} results	165
E.2.5	Kendall τ vs. Spearman ρ	166
E.2.6	Agreement Analysis	166
	Bibliography	171
	Acknowledgments	217
	Curriculum Vitæ	219
	List of Publications	221
	SIKS Dissertations	223

1

Introduction

1

The essence of democracy is that citizens have a say in how they are governed. From the public fora of the Ancient Greeks to the European Parliament, reasoning and arguing form the core of discussions where diverse perspectives are debated. However, modern governments struggle with declining citizen engagement [362] and diminished trust in political institutions [128]. At the same time, our society faces a multitude of complex, interwoven issues—climate change [181], misinformation [431], vaccination hesitancy [258], and many others [280]—that require democratic resolution. These societal issues share characteristics: problems are multifaceted and interdependent, they have no clear definite solution, decisions need to be made under strict time constraints, and solutions require fleshing out deeply-rooted ethical disagreements. These characteristics are typical for *wicked problems* [320]: issues that seemingly have no solutions due to the diverse needs of those involved.

Addressing wicked problems in society requires reshaping citizen participation [325]. *Deliberative democracy* underpins a wave of democratic transformation, advocating for decisions to be made through fair and reasonable discussion [289]. Central to deliberative democracy is the process of **deliberation**, where citizens, not just experts or politicians, are deeply involved in shaping solutions to societal issues [106]. Deliberation is based on egalitarian and rational debate, with expert information freely accessible [155]. Solutions stemming from deliberation benefit from the wisdom of the crowd effect: the collective judgment of a diverse crowd of humans is more accurate than any individual member in that group. Humans are good collaborative problem solvers [219], and collective decision-making builds sustainable solutions [163]. However, deliberations need careful facilitation to sustain the conditions for productive discussions and safeguard democratic ideals.

The diversity of perspectives is a driving factor in determining the quality of outcomes in a deliberation [36, 53, 87]. When citizens express their desires and provide insights from different backgrounds, diversity leads to effective decisions [227]. Diverse perspectives can spark creative solutions by challenging assumptions and encouraging innovative thinking. This is echoed in cases of democratic transformation where encouragement of diverse perspectives is hailed as a means of stabilizing democracy [114].

Facilitating diversity requires actively steering the deliberation process. First, participation from a broad group of representatives requires more organizational overhead to ensure an inclusive recruitment procedure. Second, deliberating the complex needs of individuals requires active perspective-taking from those involved in the discussion, imparting a significant cognitive and emotional load [133, 213, 391]. Third, the deliberation process requires moderators that play a crucial role in setting ground rules for respectful communication, encouraging participation from all members, managing conflicts constructively, and summarizing discussions to highlight different viewpoints [91, 136].

Existing deliberative practices have inherent limitations, such as a reliance on physical gatherings and the frequent use of small, supposedly representative, citizen groups [26]. Even small-scale deliberations see issues surrounding organization, effective participation, and collective decision-making [123]. For instance, gathering people to come together physically at a specific time is resource-intensive [115]. Further, there is a maximum number of people that can be feasibly included, limiting the diversity of that group.

Alternatively, contemporary social media platforms enable large-scale communication and may facilitate large-scale *online* deliberations [132], fostering citizen engagement [159, 348]. These platforms can serve as a channel for the rational exchange of ideas and opin-

ions, provide access to a broad range of information sources, and host facilitated discussion through moderator involvement [118]. Large technical leaps, like recommender systems [14] and automatic translation [444], can provide opportunities for all citizens to contribute to the public debate. Lowering the barrier to accessing societal discussions allows global issues like climate change to be addressed not by a limited group of representatives, but through engagement across all layers of society. However, whether such platforms serve as an inclusive public space or not remains debated [297]. Online discussion is fundamentally different from the conversations in offline deliberation [25]. Online discussions offer wider and more free participation but are less regulated and harder to moderate than offline ones. It is therefore important to highlight the prerequisites for achieving the wisdom-of-the-crowd effect in online discussions: the egalitarian participation of a diverse crowd of citizens.

Transitioning to online deliberation adds a new dimension to the challenge of facilitating diversity: that of **scale**. Considering the massive user bases online platforms can support, manual moderation becomes infeasible. Online opinions spread and evolve differently from guided offline deliberations [441, 447]. In offline deliberation, diverse participation is attained by representative sampling according to demographics. However, ubiquitous participation from online users leads to open questions on how to foster the development of diverse perspectives when such a strategy is infeasible. Since poorly designed online discussions can lead to polarized outcomes [437], this challenge needs to be considered carefully.

To effectively facilitate online discussions at scale, it is essential to have tools that can analyze these discussions. In this dissertation, we consider these interactions to be text-based exchanges of opinions. On social media platforms, humans engage with one another by communicating their viewpoints through written text. We turn to Natural Language Processing (NLP) and create new methods for harvesting insights from opinions. While investigating human behavior has long been the domain of social sciences, combining social science methodologies with NLP models has barely passed its infancy [454]. This emerging interdisciplinary approach offers new avenues for understanding large-scale human interactions. To uphold democratic ideals, it is essential to develop responsible tools [455], which requires a thorough understanding of the shortcomings of existing NLP techniques. We create an overview of these limitations and propose a strategy to overcome them in the form of Hybrid Intelligence (HI). HI refers to integrating human and machine intelligence, enhancing human capabilities instead of replacing them [5]. We dive into how we can create HI that combines citizens and NLP methods to facilitate diversity in online societal discussions.

Improving citizen engagement through deliberation requires effective collaboration between citizens and stakeholders, such as politicians or industry parties. The institutional uptake and implementation of deliberation efforts have thus far remained unfocused and scattered [140, 360]. One reason for the hesitant uptake of online deliberation is that legitimate deliberative processes need to account for non-included individuals to be considered representative [298]. Enhancing citizen participation by designing and implementing technical solutions for addressing societal issues at scale can help in achieving legitimacy [148]. This dissertation contributes to this goal by proposing to engage with a diverse public directly through NLP-supported facilitation. Focusing on finding wide-ranging perspectives in society-wide conversations leads to inclusive and informed decision-making. An integrated view of the humans involved in online discussions should limit adverse effects such as echo chambers [77], polarization [392], and other negative external and internal effects

[251, 263, 370]. In the long run, the positive effects of promoting diversity in online discussions can lead to the empowerment of citizens.

Structure The rest of this chapter is structured as follows: We provide an overview of the problem of facilitating online discussions with NLP based on the ideals of deliberation and introduce our Research Questions (RQs) in Section 1.1. We continue with a description of the relevance of each RQ in Section 1.2. We define the scope of this dissertation in Section 1.3, and finally provide an outlook on the findings of our work in Section 1.4.

1.1 Research Questions

Online discussions generate vast amounts of content, which is challenging to manage and navigate [88] because content is scattered across time and threads, and contains frequently repeating or unconnected arguments. This makes it difficult for users to know where to add new contributions, resulting in low-quality content [204]. These issues can be addressed by employing moderators, e.g., to structure the content of a discussion or to steer user interactions [390]. However, given the amount of data, manual moderation is not feasible.

Instead, we turn to NLP for interpreting text-based opinions at scale [374], powered by the recent surge of Large Language Models (LLMs) [20, 266]. LLMs have shown a remarkable ability to code novel texts with limited adaptation requirements [385]. Central to our approach to facilitation is extracting structured *perspectives* from users in a discussion. Perspectives provide high-level insights into the arguments employed by citizens [414] or the motivations underlying the opinions in a community [429]. These representations influence the facilitation strategies [121] and shape policies following the discussion [274].

Using NLP for analyzing perspectives sourced from online discussions is challenging. For instance, social media platforms have been centered on managing large volumes of information, e.g., through personalized recommendations [3] or argument structuring [178] but have neglected inclusive design aspects [352]. This can cause majority opinions to be heard while suppressing dissent voices [282], or lead to filter bubbles [392]. Similarly, we see that LLMs capture majority opinions well, but do not distill all voices equally [e.g., 278, 405]. Further, LLMs lack deep social reasoning [232], may be biased [162, 333], and make mistakes in ways humans cannot anticipate [175]. LLMs can be readily applied in new contexts, but they remain fickle and inconsistent depending on the exact prompts used [254]. Straightforward automated discussion analysis runs the danger of ignoring diverse opinions, which undermines the wisdom-of-the-crowd effect [250]. To find out the nature of these challenges and whether they can be resolved, we ask our first research question:

Q1 *What are the fundamental issues in using NLP to analyze perspectives?*

Next, our goal is to obtain structured perspectives from online societal discussions that provide insights into the opinions involved. In particular, we aim to improve the degree to which **diverse** perspectives can be obtained. This requires us to combat the limitations of NLP by adopting a “hybrid” mindset, i.e., incorporating humans-in-the-loop to address diversity directly. We leverage LLMs and humans jointly, with their complementary capacities for interpreting opinions from text. This leads to our second research question:

Q2 *How to combine human intelligence and NLP to effectively capture diverse perspectives?*

Finally, in practice, analyzing opinions is modeled by different task formulations, all aimed at extracting various types of information based on language input. We propose a **perspective hierarchy** that incorporates *stance*, *arguments*, and *personal values* to represent perspectives at different levels of abstraction. We base our model on the complementary skills of humans and NLP methods, in which we mix higher-order abstractions with surface-level extraction tasks. Each task has been investigated separately, but little is known about their interaction in online discussions. We, therefore, ask our third research question:

Q3 *How to construct a perspective hierarchy based on diverse opinions in a discussion?*

1.2 Research Methodology

We introduce the methods for answering the research questions step by step.

1.2.1 Fundamental Issues (Q1)

There is an increasing interest [e.g., 84, 183, 440] in using NLP to facilitate online societal discussions. Existing work is focused on (1) using NLP tools, in particular few-shot prompted LLMs, to analyze the discussions [e.g., 377, 440], and (2) using discussion data to benchmark the capabilities of NLP tools [e.g., 124]. In the next two sections, we provide related work to the research methodologies adopted in this dissertation, highlighting fundamental techniques and applications.

Discussion Analysis

Using NLP to analyze large amounts of text in online interaction is studied under the broad umbrella of opinion mining [244]. Discussions happen in various contexts, such as climate change [249], pandemics [160], and others [49]. The scale of these discussions, combined with their pertinence, makes analyzing them interesting. Analyzing what humans express through text is the core task in many NLP areas, e.g., Opinion Summarization [244], Argument Mining [224], Sentiment Analysis [424], and Value Classification [237]. These tasks lie at the heart of creating insights into online (political) discourse. They can be used e.g., for estimating the quality of discussions [368], extracting the arguments involved [220], or reasoning over inconsistencies between choices and their justifications [243]. In the age of LLMs, these tasks have seen considerable performance improvements [186], although new challenges such as dealing with shortcut learning [138] or mitigating social biases [232] arise.

Extracting diverse views from online discussions is challenging for three reasons. First, data from social media platforms inherits biases present on these platforms, including fake news, trolling, and polarization [77]. This impacts how opinions are shaped [167] and the distribution of opinions [441]. Second, when analyzing the opinions about societal issues, not all citizens have equal access due to the digital divide [86] or differences in tech-literacy [206]. This makes the users in online discussions biased and less diverse. Third, since users are free to join in discussions of their choosing, there are undesired echo chambers or self-selection effects among the messages seen by users [363].

Despite these challenges, we can use NLP to investigate questions about human behavior at scale [225]. Analyses about behavior may lead to insights at both individual and group levels. This can be useful for improving democratic processes [80], but also applies in other areas, such as faithfully interpreting product feedback [34], service improvement [358], or course management for education purposes [233].

Approach

We can employ discussion analysis to benchmark how well NLP approaches understand opinionated text. In benchmarking, we test the analysis procedure, and models used, for possible mistakes and biases. Representing subjectivity is difficult since LLMs do not faithfully capture the full range of opinions [108, 166, 405]. Whether LLMs can learn to represent them in the future remains unclear [337, 427], but research suggests that they cannot [20, 124]. Therefore, we work with the assumption that this is a fundamental limitation of LLMs, and we have to find other approaches for improving diversity.¹

Creating diversity-enhancing techniques is gaining traction in NLP, but there are several aspects of diversity. For instance, creating more diverse news recommender systems is a common goal [216, 438] for shaping an individual's perspective [29]. Others strive to make LLMs better represent a diverse group of annotators based on their labeling behavior and demographics [28, 217]. In such approaches, models rely on annotated data. Labels are obtained from a few human annotators per instance and are often aggregated by majority voting, painting an incomplete picture of the true range of interpretations of opinionated text [302]. The role of subjectivity in these tasks remains unclear [21, 61]. This holds for traditional supervised learning, but also for the latest trends in instruction-tuning [393, 422].

Contributions

In Part I of this dissertation, we dive into the application of LLMs to analyze the opinions in online discussions. Our work centers on argumentation: the rationales behind human opinions. In Chapter 2, we begin by examining the diversity of the opinions in LLM-generated summaries of argumentative content. We find that automated methods for summarizing arguments struggle to represent arguments shared by few people, and such error cases usually go unnoticed using standard NLP evaluation practices. By examining how LLMs fare on complex argument quality assessment tasks under strong data constraints in Chapter 3, we aim to further investigate how we can best deal with low-resource settings. Here, we observe that zero-shot models can drive the state-of-the-art, but come with significant cost and data requirements to work well out of context. Overall, significant challenges remain when applying LLMs to tasks of analyzing opinionated data at scale.

1.2.2 Hybrid Intelligence (Q2)

In Part II, we argue that the aforementioned challenges can be overcome by using LLMs to **assist humans** in mining opinionated text, rather than replacing humans. This notion of *Hybrid Intelligence* [5, 97, 98] is central to our approach to uncovering diverse perspectives in online discussions. In Hybrid Intelligent Systems (HISs), Artificial Intelligence (AI) agents are collaborators that enhance human abilities such as reasoning, decision-making,

¹Although linguistic diversity generally refers to the diversity of language proficiencies [103, 190], we are specifically interested in diversity in arguments, communication styles, and values in online discussions.

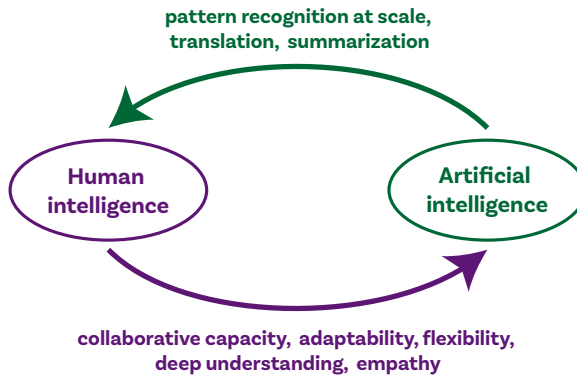


Figure 1.1: Feedback loops in Hybrid Intelligence.

and problem-solving [383]. Hybrid intelligence aims to augment intellect, creating a synergy between humans and NLP. For supporting online discussions, we combine the strengths of human intelligence with AI, highlighting bidirectional gains, as shown in Figure 1.1.

The application of AI to understanding human written language has had a profound impact on how researchers analyze human behavior at scale. To do so responsibly, we must ensure that our methods uphold democratic values, especially considering the pressing need to represent diverse perspectives. Previous work on hybrid approaches for NLP includes user adaptation [256], human-in-the-loop computing [423], human-AI interaction [164] and others [e.g., 82, 102]. Recent interest in explainable AI has focused on human understanding of NLP models [230]. Specifically for NLP, much focus is on approaches that mix crowd, expert, and automated decision-making, which have been applied to analyzing discussion content [208, 295]. However, these approaches have a one-way interaction between the NLP model and humans, as we will describe in the next section.

Approach

We observe that LLMs have many challenges to overcome in representing diverse perspectives (Section 1.2.1). Discussions are deeply human, who can adapt to incomplete and informal argumentation, behave flexibly, and provide empathic responses to foster collaboration. Thus, humans and NLP can benefit from each other. In the next paragraphs, we examine each benefit in either direction (humans aiding NLP or NLP aiding humans) separately, and lastly illustrate how both can be incorporated into an overall hybrid method.

Humans aiding NLP Humans provide the data that the NLP tools perform their analysis on, as gathered from interactions between different stakeholders, including casual and advanced users, moderators, or even site admins [336]. They provide text and behavioral data, such as likes or post-votes, which we, in turn, can use to analyze their attitude. Furthermore, NLP approaches learn from labeled data, obtained from annotators who observe a given text and draw labels from a predefined set of classes. Humans can be flexibly employed in such procedures, dealing with expanding label sets [396], free-form text response [294], asking a crowd of annotators rather than individuals [286], and more [e.g., 302, 334]. Humans contribute their opinions, either through text or by labeling, based on

lived experiences or professional expertise, and are capable of empathizing with others. While crowd annotators are usually uninformed lay users, they are assumed to adapt to tasks quickly given a set of instructions and examples. Since annotators adapt differently, addressing the problem of diverse opinion understanding requires selecting an appropriate set of annotators, to capture the human label variation [302].

NLP aiding humans NLP aids humans in online discussions in multiple ways. While we have mostly discussed the analysis of large-scale discussion data, there is a broader potential impact of NLP technologies in online deliberations [384]. First, NLP may enable, rather than restrict, access to certain services, for example by summarizing large amounts of content through summarization or using automatic translation to account for different language proficiencies. Second, since humans suffer from cognitive biases, NLP models may offer an alternative interpretation of the content. Machines do not get bored and treat each sample with equal consideration. Third, NLP models mirror biases captured in the data, which allows for obtaining synthetic opinion data or exposing biases in discussions. Lastly, since their scale, speed, and accessibility to researchers are advancing quickly, we can experiment with them rapidly.

Combination Existing work mostly offers one-directional benefits, either machine- or human-oriented. By constructing hybrid approaches, we aim to improve both humans and AI through an iterative process. We see that NLP methods are biased, leading to questions about the soundness of the analysis. Humans can repair biases and provide deeper interpretations, contexts, and explanations. Furthermore, we see that there are many opportunities for NLP to aid humans. Completing the loop allows bootstrapping: traversing the two feedback loops shown in Figure 1.1, iteratively refining the analysis procedure while performing discussion analyses. In this procedure, a human interprets opinions shown from the output from a model and possibly corrects it in a human-in-the-loop fashion [273]. However, to guide the human through a large amount of data, the NLP models will steer it through what content to observe. Through this collaborative approach, we hope to synthesize **bidirectional gains**. Bidirectional gains in hybrid intelligence refer to the mutual increase in capabilities achieved when human and artificial intelligence work together. We emphasize the synergistic nature of human-AI collaboration, where each side strengthens the other, leading to more powerful, efficient, and reliable intelligence than either could achieve alone. Hybrid approaches combine the strengths of humans and machines, offering immediate and long-term benefits. By keeping humans in the loop, their task proficiency improves, and additional data is generated to develop the hybrid collaboration. Further, advancements in NLP models can be integrated into the framework. However, doing so effectively requires broad contextual understanding.

Developing hybrid approaches necessitates a new evaluation paradigm. We must assess the effectiveness of our method by comparing it against both human-only and machine-only baselines. In the field of NLP, test sets are typically compiled manually and with hidden data issues [99, 154, 329], which might introduce an unfair advantage to the upper bound of performance [56, 200]. Initial work shows that there are considerable performance gaps between hybrid and manual approaches [127, 443].

Contributions

We present our approach to incorporating humans and NLP methods for analyzing opinionated text data. First, we introduce a method for mining diverse arguments from citizen feedback in Chapter 4. Our method, HyEnA, finds more diverse arguments and improves the precision of the argument analysis by efficiently querying human annotators across three distinct phases. In Chapter 5, we further investigate how differences between annotators in subjective tasks, such as interpreting texts for extraction of arguments or personal values, can be modeled more efficiently. Our approach steers models to learn diverse label distributions by picking from a large pool of annotators. Central to our work, we create discussion analysis approaches that (1) select samples for human inspection that are interesting to annotate, (2) account for diversity (e.g., leveraging contextualized embeddings [314]), and (3) seek labels from multiple annotators. The hybrid nature of our methodology leads to bidirectional gains, serving the NLP system as well as the humans involved. For instance, we create approaches to capture more diverse interpretations of the arguments in discussions using a crowd of annotators. After the annotation, our method outputs a summary of the high-level argument involved, while annotators were able to develop their understanding of controversial discussions. We achieve a cost-effective crowd annotation, while actively engaging with the annotators, and developing their perspective. Moreover, we can also actively diversify which annotator we query an annotation from. We observe that an active selection of diverse annotators can inform a model more quickly of the label distribution underlying subjective tasks in cases where the annotator pool is large.

1.2.3 Perspective Hierarchy (Q3)

Given that NLP can process large amounts of discussion data, but is limited in its capabilities (Section 1.2.1) and that we may construct hybrid procedures to account for these limits (Section 1.2.2), we address the challenge on how to capture perspectives. Uncovering perspectives from online societal discussions requires a representation for identifying how people feel about potential decisions, how the considerations are communicated in the discussions, and the motivations underlying preferences held by individuals. There is a large amount of literature concerned with addressing these questions through separate NLP tasks. We attempt to integrate these tasks and find out how they model various aspects of perspectives. We propose a hierarchy to structure our approach to facilitating online discussions at scale.

Few attempts to comprehensively represent perspectives exist [71, 412]. These works focus on annotating utterances for low-level claim information [272], or investigating the reasoning behind the views held in discussions [104]. Stances and arguments are inherently linked in argumentation models [386, 408], and form the basis of frameworks for representing perspectives [72, 432]. Existing work on mapping deliberative discussions has focused extensively on capturing this reasoning and using it for facilitation [158, 205].

However, stances and arguments do not represent opinions at a deeper personal level. A fundamental concept for explaining the motivations underlying opinions is personal values [344]. There are various theories of personal values [e.g., 143, 321, 344]. Preferences among values describe the attitude of individuals and groups [304], and can be extracted from behavioral cues to investigate political affiliation [326], perform moral reasoning [271] or positively influence lifestyle [95]. Values are abstract and need to be interpreted inside their context, making it difficult for both humans and NLP methods to measure them reli-

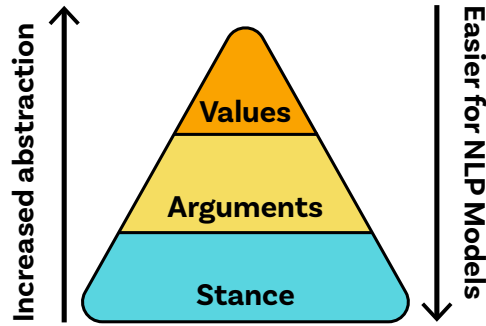


Figure 1.2: The perspective hierarchy. The higher the level of abstraction, the more human intelligence is required for interpreting the component.

ably [241]. One way to contextualize them is to connect values to argumentation, focusing on how choices are justified [201]. Using this insight, we incorporate personal values into our perspective representation and aim to obtain them using a hybrid approach.

Approach

We propose a perspective hierarchy to represent a person's perspective at different levels of abstraction, shown in Figure 1.2. Our perspective hierarchy is composed of stances, arguments, and values. We adopt the following definitions:

Stance Whether, or how much, support or opposition is expressed to a claim. Stance detection has been studied extensively and remains a popular NLP task [214].

Arguments The reasons given for adopting a stance towards a claim. In real-world contexts, argumentation manifests in many forms and is predominantly informal [146]. Mining arguments from text works well within known contexts [112], but suffers from implicit reasoning [157]. Hence, we require more human guidance to correct for possible mistakes in automated methods.

Values The motivations underlying opinions and actions [344]. Values are communicated implicitly through actions or written motivations. Estimating values automatically remains difficult even within their context [202]. Only through iterative hybrid procedures can we accurately reason about preferences among human values.

We combine the three components into a layered hierarchy, to indicate a tradeoff with respect to (1) the capabilities of NLP models to capture information from text, and (2) the level of abstraction that the component captures. Higher-order abstraction requires “filling in” more implicit knowledge. For instance, for stance detection, one or a couple of sentences can be enough to determine the stance of an individual concerning a topic [31]. However, for estimating value preferences, we need continued interaction over time to infer how values are prioritized within their context [240].

We illustrate how we used data from large online social media platforms to investigate perspective hierarchies for individuals [400]. Our main objective is to investigate whether

we can connect stances and values directly, omitting arguments to challenge their inclusion in the hierarchy.

Given a societal discussion on an online platform [305], we first identify relevant controversial topics and apply our automated methods for obtaining stances and value preferences. Because of the aforementioned limitations, we utilize the human-in-the-loop approach to uncover possible mistakes from the extraction pipeline. In particular, we compare human-provided self-reported value preferences to those estimated from behavioral data. Using this data, we can (1) compare how well the automated approaches work versus manual ones, (2) mix information from self-reported and behavior-based value preferences, and (3) investigate the relationship between components of the perspective hierarchy.

Contributions

In Part III, we make use of our hybrid setup to investigate the perspectives of participants in online discussions at scale. In Chapter 6, we investigate connections between value conflicts and disagreements in online discussions on societally relevant topics. Our experiments show that only when values are diverse, automatically-identified conflicts in values can correlate to stance disagreement. No strong evidence points towards a consistent and context-independent link between disagreement and value conflicts. However, when we incorporate human-provided self-reports, the evidence becomes stronger, showing that the hybrid approach is crucial to performing a meaningful analysis. When strong value diversity is absent, we cannot correlate disagreement and value conflict directly at all. A lack of a direct link means we require a more complete picture, and thus we incorporate arguments to complete the perspective hierarchy.

1.3 Dissertation Scope

The topic of this dissertation lies in the intersection of computer science, natural language processing, social science, and political theory. It is, therefore, inherently interdisciplinary and therefore can be approached from multiple angles. We provide a scope of the research involved before we dive into the description of how we address each research question.

In our work, we consider *online discussions* as text-based user interactions that happen on contemporary online platforms such as Twitter/X² or Reddit. Furthermore, we include data from specific questionnaires that gather citizen responses on proposed policy. We focus on topics that are *societally relevant*, such as climate change, due to the difficulty of addressing them. Lastly, we concern ourselves with deliberation among a group of people, as opposed to individual deliberation for self-reflection purposes.

Core to our work is diversity of perspectives. Depending on the context, the definition of diversity encompasses differences in various attributes, including social categories (e.g., gender, age, race) and informational or functional attributes (e.g., functional background, educational background) [30, 168, 409]. Research on group deliberation and diversity has primarily focused on a limited set of dimensions within these categories, or on the interplay between these two dimensions. In this work, we adopt *diversity of perspectives* as the full range of beliefs, opinions, stances, and values held by a given group of people. For any two people, these components might be in conflict at arbitrary levels, requiring extensive deliberation to uncover common ground. Our definition is similar to those adopted in other work

²Starting from July 2023, Twitter was renamed to “X.”

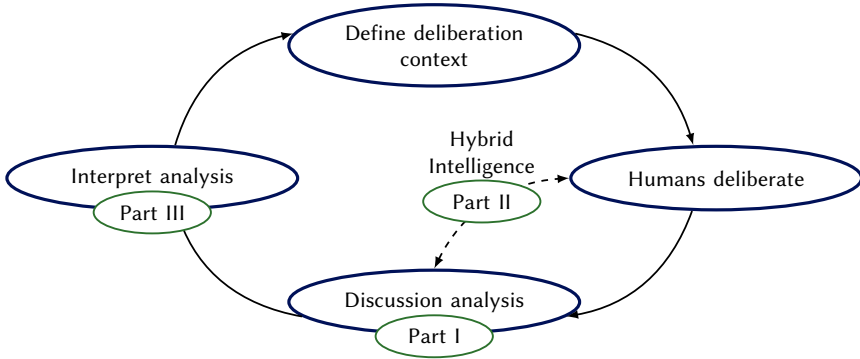


Figure 1.3: The deliberation cycle, annotated with the three parts addressed in this dissertation.

in group deliberation, often referred to as cognitive diversity [194], or viewpoint diversity [105]. It is distinct from demographic diversity [221] since we target the opinions and not the opinion holders, or linguistic diversity [190], which focuses on language proficiency.

This dissertation is focused on developing hybrid approaches to analyzing discussions from a technological perspective, with *hybrid* indicating human–AI cooperation [5, 97, 312]. We make modifications to computational artifacts (such as NLP models and datasets) and design processes for discussion analysis. Other strategies for improving discussion analysis, such as teaching humans analytical skills or implementing interventions for behavioral change, are left as future work but are compatible with our setup.

Lastly, our work is concerned with creating AI methods that focus on understanding human opinions based on digital text. Neural approaches from Natural Language Processing, in particular Transformer-based models, are the workhorse in the experiments performed in this dissertation. Other behavioral information, such as direct polling, referenda, post-voting, and others may provide different and possibly conflicting information for interpreting an individual’s perspective. Consolidating such information with text-based opinions is nontrivial and requires careful prioritization of signals [354].

1.4 Outlook

Our goal is to augment the diversity of the opinions present in online societal discussions. These discussions are rooted in deliberative ideals, aiming to foster inclusive, informed, and respectful exchanges that lead to collective decision-making and problem-solving. We enhance the discussion analysis process by considering discussion analysis a hybrid undertaking, bringing HI to aid the deliberative cycle as shown in Figure 1.3. We separate our work into three parts as follows. First, we identify the strengths and weaknesses of using NLP to analyze discussions with diverse perspectives in Part I. Second, we see how HI can improve the capture of diverse perspectives in societal discussions in Part II. Our work proposes hybrid methods to sustain a high degree of diversity in discussions with a large crowd. Third, in Part III, we propose a perspective hierarchy to guide the investigation of human opinions in online societal discussions at scale.

The outlook of using HI, where we augment human intellect with AI, particularly sup-

ports deliberative discussions and decision-making processes. Our approach can democratize access to information and enhance the quality of public discourse by providing structured data analysis, fact-checking, and summarization pipelines, enabling more informed and evidence-based conversations. HI also facilitates inclusivity by assisting individuals with different abilities and backgrounds, ensuring a broader range of voices are heard. It aids in navigating complex societal challenges, such as climate change or public health crises, by integrating diverse data sources and perspectives. However, it is crucial to ensure that these technologies are developed and deployed ethically, mitigating biases and maintaining transparency to foster trust and acceptance in society. Ultimately, HI has the potential to empower communities, strengthen democratic processes, and drive more effective problem-solving for societal issues.

I

NLP for Online Discussion Analysis

Introducing Part I: NLP for Online Discussion Analysis

In Part I of this dissertation, we dive into the application of LLMs to analyze the perspectives in online societal discussions. Our work centers on argumentation: the rationales behind human opinions. In Chapter 2, we begin by examining the diversity of the opinions in LLM-generated summaries of argumentative content. We find that automated methods for summarizing arguments struggle to represent arguments shared by few people, and such error cases usually go unnoticed using standard NLP evaluation practices. By examining how LLMs fare on complex argument quality assessment tasks under strong data constraints in Chapter 3, we aim to further investigate how we can best deal with low-resource settings. Zero-shot prompting of LLMs can drive the state-of-the-art under realistic data constraints but still incur significant costs and highlight how diverse data improves their effectiveness in generalization to novel contexts. Overall, numerous challenges emerge when applying LLMs to tasks of analyzing opinionated data at scale. Later, in Part II, we will argue that the aforementioned challenges can be overcome by using LLMs to **assist humans** in mining opinionated text data, rather than replacing them.

Part I focuses on the following research question:

Q1 What are the fundamental issues in using NLP to analyze perspectives?

2

2

An Empirical Analysis of Diversity in Argument Summarization

Presenting high-level arguments is a crucial task for fostering participation in online societal discussions. Current argument summarization approaches miss an important facet of this task—capturing diversity—which is important for accommodating multiple perspectives. We introduce three aspects of diversity: those of opinions, annotators, and sources. We evaluate approaches to a popular argument summarization task called Key Point Analysis, which shows how these approaches are ill-equipped for (1) dealing with data from various sources, (2) representing arguments shared by few people, and (3) aligning with subjectivity in human-provided annotations. We find that both general-purpose LLMs and dedicated Key Point Analysis models vary along these three criteria, but have complementary strengths. Further, we observe that diversification of training data may ameliorate generalization. Addressing diversity in argument summarization requires a mix of strategies to deal with subjectivity.

2.1 Introduction

Getting an overview of the arguments concerning controversial issues is often difficult for those participating in ongoing discussions. In these discussions, many points are being communicated, there is no way to track which arguments were already encountered, and participants engage in haphazard miscommunication or conflicts. Automatic summarization is a way to provide a comprehensible overview of the opinions [15, 281]. However, generating summaries representative of the arguments involved in a discussion is difficult [32]. Argument summarization extends beyond text summarization because it separates argumentative and non-argumentative content, preserves the argumentative structure, and provides explicit stances on a central claim or hypothesis.

Summarizing arguments is challenging in many contexts, but the potential impact is high. For instance, after summarizing the arguments from societal discussions, the extracted arguments may shape new policies and may be used to justify decision-making [17, 153]. Similarly, businesses depend on review data to find customer feedback, which can steer product design [18].

Although arguments are often summarized by hand in practice [e.g., 264, 274, 279], recent developments in Argument Mining (AM) allow automatic analysis of argumentative text [224]. Obtaining summaries that faithfully represent open-ended opinions requires careful evaluation, especially in sensitive contexts, e.g., summarizing citizen feedback [109, 267].

One approach for generating comprehensive summaries of arguments is Key Point Analysis [KPA, 32]. In KPA, a corpus of opinions is analyzed for the *key points*, those arguments that are salient and repeated multiple times. However, some aspects of the KPA experimental design misalign with respect to real-world applications. We illustrate these blind spots, in particular, when applied to summarizing online societal discussions. We highlight three dimensions of **diversity** that are central to empowering citizens' opinions at scale [352]: (1) incorporating the long tail of opinions, (2) including diverse perspectives from annotators, and (3) being robust in handling data from multiple sources.

How current KPA approaches deal with the above dimensions of diversity is unexplored. We incorporate the standardized benchmark and two other datasets to experiment with different approaches. We develop specific analyses to uncover how KPA approaches fare on each dimension of diversity. In addition to the existing approaches, we use LLMs by prompting them to perform KPA, as they may be an attractive alternative to current models.

Applying KPA approaches across several datasets that vary in how they address diversity leads to mixed results. KPA approaches generalize poorly across data sources when used in transfer learning settings, though approaches reveal complementary merits across tasks. Further, their performance degrades when dealing with low-frequency opinions, i.e., opinions repeated by relatively few individuals. Finally, we observe that KPA approaches disregard subjective interpretations among individual annotators.

Contributions (1) We critically examine three dimensions of diversity—of opinions, annotators, and sources—in the KPA setup. (2) We analyze the behavior of existing metrics on one existing and two novel datasets. (3) We analyze multiple methods, including prompt-based LLMs, broadening the scope of methods that can perform KPA.

2.2 Related Work

2.2.1 Key Point Analysis

KPA serves to separate argumentative from non-argumentative content, and condense argumentative content by matching arguments to key points [32]. Key points can be seen as high-level arguments that capture the gist of a set of arguments. While most work on KPA selects high-quality arguments as representatives, generating novel key points has been proposed as an alternative [376]. KPA has been applied across topics using data from discussion portals or online reviews [33, 34]. KPA is usually divided into Key Point Generation and Key Point Matching steps (see Section 2.3.1).

Multiple approaches exist for KPA [131]. Modeling choices consist of popular Transformer models such as BERT [301], enhanced representational quality using contrastive learning [10], and the incorporation of clustering techniques [231]. Our work aims to investigate some of the modeling choices employed in these works. For instance, in Li et al. [231], the authors discarded unmapped arguments, which may hurt the ability to represent minority opinions.

2.2.2 Opinion Summarization

Opinion summarization aims to generate summaries of an individual's subjective opinions [48, 180], often applied to product reviews [75]. Leveraging Transformer models is popular for opinion summarization [13, 16], though generic extractive summarization techniques are strong baselines [373]. Measuring bias in generated summaries has seen recent interest, specifically acknowledging that diverse opinions should be taken into account [176, 355] or postulating that diversity is a desirable trait when generating opinions [12, 420]. Our work applies these techniques to argumentation to obtain a high-level summary of opinions, and analyses differences in behavior for (in-)frequent viewpoints.

2.2.3 Diversity in Societal Decision Making

Sensitive decision-making contexts call for responses rooted in reason that serve social good rather than specific interests. One way of obtaining such responses is through evidence-based policymaking, which involves stakeholders and the broader public to strike decisions [64]. Citizen participation improves the support of the decisions when some requirements are met [260]. A key factor among those requirements is the involvement of a diverse group of citizens, independently voicing opinions [375]. Approaches to summarizing arguments in such citizen feedback face similar requirements.

In Argument Mining, we find recent work that aligns with these views, e.g., by a strong focus on the diverging perspectives among annotators in AM tasks [322]. Further, some preliminary work adjusts visualization for minority opinions [38]. However, in terms of data sources, most work is still centered on English-speaking content, with few multi-lingual or multi-cultural resources available [414].

2.3 Method

We formulate the KPA subtasks—*Key Point Generation* (KPG) and *Key Point Matching* (KPM). We then introduce the three dimensions of diversity and consider them when applying KPA.

Dataset	Data Source	Filter low freq.	Key Point source	Non-aggregated annotation	IRR
ARGKP	Human annotation	✓	Expert	✗	0.50-0.82 (κ)
PVE	Citizen consultation	✗	Crowd	✓	0.35 ($\kappa^\dagger$)
PERSPECTRUM	Debate platforms	✗	Crowd	✗	0.61 (κ)

Table 2.1: Datasets and their diversity characteristics when considering the KPA task. The inter-rater reliability (IRR) is measured via Cohen’s κ scores or prevalence and bias-adjusted Cohen’s $\kappa^\dagger$ [PABAK, 357].

2.3.1 Task setup

We outline the two subtasks that constitute KPA, as originally introduced by [131].

Key Point Generation (KPG) focuses on generating *key points* $\mathcal{K}$ given a corpus of arguments $\mathcal{D}$ on a particular claim. Key points are high-level arguments that capture the gist of a collection of arguments. Key points oppose or support the claim.

Key Point Matching (KPM) *matches* arguments to key points. An argument matches a key point if the key point directly summarizes the argument, or if the key point represents the essence of the argument. We ensure that the stance of the key point (pro or con) matches the stance of the argument. Formally, given a set of key points $\mathcal{K}$ and a corpus $\mathcal{D}$, we score the match between an argument $d \in \mathcal{D}$ and a key point $k \in \mathcal{K}$ with a matching model $M(d, k)$. Assigning arguments to key points using match scores is flexible, and multiple strategies can be taken to reach a final decision (e.g. imposing a match score threshold) [33]. Since the assignment strategy is largely context-dependent, we evaluate the scoring mechanism itself, instead.

2.3.2 Modeling Diversity in Key Point Analysis

We focus on three main aspects of diversity.

Long tail opinions Several NLP models imitate biases that exist in datasets [51]. For argument summarization, focusing on majority arguments is one such form of bias, as it leads to possible misrepresentations. Failing to capture low-frequency arguments runs the danger of further estranging underrepresented viewpoints [204]. These methods need active correction from humans to account for this “long tail of opinions” [397]. For the KPA task, approaches have largely unknown behavior on capturing the long tail of opinions [278]. Additionally, LLMs struggle with learning long-tail knowledge [193], aggravating this issue. We experiment with subsampling the datasets to investigate the imbalanced data settings, which are representative of real-world use cases.

Annotators Datasets are labeled using a mix of crowd and expert annotators. Querying experts for key points may leave the impacted users (e.g., lay citizens) out of consideration [60]. Similarly, labels stemming from crowd annotation that are filtered for high agreement may disregard controversial or diverse opinions. Disagreement is a complex signal that includes subjective views, task understanding, and annotator behavior [21]. Having access to non-aggregated annotations would, e.g., allow for further modeling of patterns [89] or the

reasons [241] underlying opinions. We investigate whether models trained on such annotations can identify disagreement.

Data sources Existing works investigate cross-domain generalization of KPA methods using data stemming from a single dataset, focusing on a cross-topic setting [33, 231, 330]. This dataset is gathered at a specific time. As discussions evolve, more nuanced positions may become relevant, and new real-world events impact the opinions. Further, these discussions usually take place on a single platform (e.g., Reddit threads, Twitter discussions), inheriting biases from the source [170]. Measuring the performance of KPA approaches should rely on diverse datasets, based on data gathered from different sources at different points in time. There have been some efforts in applying KPA across different contexts [34, 66, 145], but they apply approaches to a single dataset at a time, making direct comparison difficult. Our work examines the cross-dataset performance of these approaches to assess their relative strengths and weaknesses.

Table 2.1 shows the current datasets, and how they relate to the dimensions discussed above. In all three datasets, the arguments stem from user-submitted content. In one dataset (ARGKP), low-frequency arguments (i.e., opinions repeated by few individuals) are disregarded. Further, the ARGKP benchmark relies on expert-generated key points and does not include annotator-specific match labels. PERSPECTRUM contains aggregated counts of match labels, but due to aggregation, we cannot identify annotator-specific patterns. Lastly, the inter-rater reliability differs for each dataset, with wide ranges, showing that the tasks are fundamentally subjective. We employ these three datasets for evaluating various KPA approaches and dive deeper into the three aspects of diversity.

2.4 Experimental Setup

We describe the data, KPA methods, and metrics involved in our experiments. The source code will be publicly available upon publication.

2.4.1 Data

Most work on KPA has used ARGKP, the dataset introduced by Friedman-Melamed et al. [131] in a shared task. We add two new datasets that match the KPA subtasks but have different characteristics.

ArgKP We adopt the shared task dataset, keeping the same split across claims as the original data. The ARGKP dataset contains claims taken from an online debate platform, together with crowd-generated arguments and expert-generated key points [32]. The arguments were produced by asking humans to argue for and against a claim, followed by filtering on high-quality and clear-polarity arguments. Key points were generated by an expert debater, who generated the key points without having access to the arguments. The final test set was collected after the initial dataset and has been curated to match some of the distributional properties of the training and validation sets.

PVE We use the crowd-annotated data stemming from a human-AI hybrid key argument analysis [397] based on a Participatory Value Evaluation (PVE), a type of citizen consultation. In this consultation process, citizens were asked to motivate their choices for new

Dataset	Train	Val	Test
ARGKP	24 (21K)	4 (3K)	3 (3K)
PVE	–	–	3 (200)
PERSPECTRUM	525 (6K)	136 (2K)	218 (2K)

Table 2.2: Number of claims (and arguments) when splitting the dataset into training, validation, and test sets.

COVID-19 policy through text, which formed a set of comments for each proposed policy option. The performed key argument analysis resulted in crowd-generated key points, matching individual comments to key points per option. Since this is a small dataset, we only use it for evaluation.

Perspectrum Similar to ARGKP, PERSPECTRUM contains content from online debate platforms. It extracts claims, key points, and arguments from the platform directly [71]. Part of the dataset is further enhanced by crowdsourcing paraphrased arguments and key points. The PERSPECTRUM dataset is ordered into claims, which are argued for or against by perspectives, with evidence statements backing up each perspective. We use perspectives as key points, and evidence as arguments. We retain the same split over claims as the original data. The authors provide aggregated annotations on the match between arguments and key points. While this allows us to compute the agreement scores per sample, we cannot distill individual annotator patterns.

2.4.2 Approaches

We investigate different approaches with respect to their performance on the aspects of diversity. Appendix A.1 includes a detailed overview of the setup, parameters, and prompts. Similar to summarization techniques, most KPG methods are either *extractive*, taking samples as representative key points, or *abstractive*, formulating new key points as free-form text generation [113].

ChatGPT We use the OpenAI Python API [290] to run the KPA task by prompting ChatGPT. We differentiate between open-book and closed-book prompts. For the open-book prompts, we input the claim and a random sequence of arguments up to the maximum window (given a response size of 512 tokens) in the KPG task. For the closed-book model, we only input the claim, and the model synthesizes key points. In both approaches, KPG is abstractive. In KPM, ChatGPT predicts matches for a batch of arguments at a time, all related to the same claim.

Debater We use the Project Debater API [179], which supports multiple argument-related tasks, including KPA [35]. This approach uses a model trained on ARGKP and performs extractive KPG. We query the API for KPG and KPM separately.

SMatchToPR We adopt the approach from the winner of the shared task, which uses a state-of-the-art Transformer model and contrastive learning [10]. During training, the model learns to embed matching arguments closer than non-matching arguments. These representations are used to construct a graph with embeddings of individual argument sentences as nodes, and the matching scores between them as edge weights. Nodes with the

maximum PageRank score are selected as key points. In our experiments, the model is trained using the training set of ARGKP and PERSPECTRUM. This method performs extractive KPG. We experiment with RoBERTa-base and RoBERTa-large to estimate the effect of model size [248].

2.4.3 Evaluation Metrics

We evaluate models for KPG and KPM separately. For KPG, we adopt the set-level evaluation approach from Li et al. [231]. For KPM, we reuse the match labels provided by each dataset.

Key Point Generation (KPG)

KPG can be considered as a language generation problem [135] for evaluation. We rely on a mixture of reference-based and learned metrics, measuring both lexical overlap and semantic similarity. We use the following metrics:

ROUGE-(1/2/L) to measure overlap of unigrams, bigrams, and longest common subsequence, respectively. We average scores for all stance and claim combinations. Additional details on the ROUGE configuration are in Appendix A.1.3.

BLEURT [347] to measure the semantic similarity between a candidate and reference key point, which correlates with human preference scores. BLEURT introduces a regression layer over contextualized representations, trained on a set of human-generated labels.

BARTScore [445] to evaluate the summarization capabilities directly by examining key point generation. In contrast to BLEURT, BARTScore evaluates the likelihood of the generated sequence when conditioning on a source.

For each metric $\mathcal{S}$ that scores the overlap between two key points, we aggregate scores into Precision P and Recall R scores using Equations 2.1 and 2.2. For P , we take the maximum score between a generated key point a and the reference key points $\mathcal{B}$, averaging over all $n = |\mathcal{A}|$ generated key points. We perform the analogous for R . We report F_1 scores to balance precision and recall.

$$P = \frac{1}{n} \sum_{a \in \mathcal{A}} \max_{b \in \mathcal{B}} \mathcal{S}(a, b) \quad (2.1)$$

$$R = \frac{1}{m} \sum_{b \in \mathcal{B}} \max_{a \in \mathcal{A}} \mathcal{S}(a, b) \quad (2.2)$$

Key Point Matching (KPM)

We perform the KPM evaluation by obtaining match scores for key point-argument pairs. That is, for a key point k and an argument d , we check if a new model used in the KPA method would assign d to k . We reuse existing labels and do not use the results from KPG. Since we do not consider unlabeled examples between arguments and key points, we do not need to distinguish for undecided labels (as in Friedman-Melamed et al. [131]).

We evaluate each approach using mean average precision (mAP), taking the mean over average precision scores computed for claims C . Given a claim, we compute precision P_τ and recall R_τ for all match score thresholds τ , as in Equation 2.3. In case an approach outputs a

binary match label instead of scores, we remap the scores to 0 and 1 for non-matching and matching pairs, respectively.

$$\text{mAP} = \sum_C \frac{\sum_{\tau} (R_{\tau} - R_{\tau-1}) P_{\tau}}{|C|} \quad (2.3)$$

2

2.5 Results and Discussion

First, we report on the KPG and KPM evaluation. Then, we analyze how the aspects of diversity impact performance beyond a cross-dataset evaluation. We show results when conditioning on the long tail of opinions, look into the connection between annotator agreement and match score, and how performance changes for diverse data sources.

2.5.1 KPG Performance

Table 2.3 shows the results of KPG evaluation. Overall, no single approach performs best across all datasets. All models perform best on ARGKP except for closed-book ChatGPT, which performs the best on the PVE dataset. Thus, by adopting diverse datasets, we demonstrate that experimenting with a single dataset may inflate KPG performance.

ChatGPT consistently scores well on ROUGE and semantic similarity. This indicates that the abstractive generation of key points is beneficial. For PVE, we observe a strong tendency for open-book ChatGPT to adjust the generated key points to the linguistic style of the arguments. This clashes with the reference key points, which are paraphrased to make sense without the context of the original arguments. Hence, the closed-book model, which does not observe the source arguments, performs better, adopting more neutral language.

SMATCHToPR performs best for PERSPECTRUM. Although general-purpose LLMs are strong in zero-shot settings, a dedicated model for representing arguments achieves state-of-the-art results. The Debater approach is ranked lowest across all datasets, showing that training on a single dataset generalizes poorly to other datasets.

ROUGE and semantic similarity scores mostly agree, except for BLEURT on ARGKP. Here, we see that SMATCHToPR slightly outperforms ChatGPT. We attribute this to the optimized representational qualities of SMATCHToPR: it selects key points with high semantic similarity to many arguments, which is similar to how BLEURT provides scores based on contextualized representations.

Increasing model size (of SMATCHToPR) improves performance for PERSPECTRUM, but not for ARGKP and PVE. Because PVE is small, the pool of sentences to pick key point candidates from is limited, and possible improvements of the model are negligible when extracting the key points. For ARGKP, the ROUGE scores deteriorate, while the semantic similarity scores improve slightly. Intuitively, this matches expectations: the model can navigate the embedding space better, selecting key points that may be phrased differently but contain semantically similar content.

2.5.2 KPM Performance

Table 2.4 shows the results of KPM evaluation. ChatGPT, despite its strong performance on KPG, does not accurately match arguments to key points. Interestingly, the Debater outperforms the SMATCHToPR model on the ARGKP dataset, but SMATCHToPR is stronger on the PVE and PERSPECTRUM datasets. SMATCHToPR's strong performance on PERSPECTRUM

Dataset	Approach	R-1	R-2	R-L	BLEURT	BART
ARGKP	ChatGPT	34.3	12.5	30.3	0.556	0.540
	ChatGPT (closed book)	29.5	7.1	25.6	0.314	0.256
	Debater	25.6	5.5	22.5	0.334	0.307
	SMatchToPR (base)	31.7	11.1	29.7	0.553	0.494
	SMatchToPR (large)	30.5	8.3	26.8	0.563	0.497
PVE	ChatGPT	18.5	3.9	15.3	0.329	0.369
	ChatGPT (closed book)	27.1	8.6	21.4	0.376	0.378
	Debater	13.3	0.0	13.3	0.294	0.188
	SMatchToPR (base)	21.3	3.7	16.6	0.351	0.344
	SMatchToPR (large)	21.3	3.7	16.6	0.351	0.344
PERSPECTRUM	ChatGPT	21.3	5.7	18.2	0.355	0.322
	ChatGPT (closed book)	17.1	3.8	15.0	0.291	0.258
	Debater	9.4	0.4	8.5	0.197	0.210
	SMatchToPR (base)	22.5	6.5	19.3	0.257	0.232
	SMatchToPR (large)	22.7	6.7	19.4	0.403	0.363

Table 2.3: ROUGE scores and semantic similarity scores for the Key Point Generation task.

Name	mAP		
	ARGKP	PVE	PERSPECTRUM
ChatGPT	0.17	0.27	0.46*
Debater	0.82	0.51	0.51
SMatchToPR (base)	0.76	0.53	0.80
SMatchToPR (large)	0.80	0.61	0.82

Table 2.4: Results for the Key Point Matching task. Closed-book ChatGPT scores are not available, since its KPA is made without observing arguments. The scores for ChatGPT on PERSPECTRUM (*) were estimated on a subset of the test set to cut down costs.

and ARGKP is expected—they were included in its training. However, its good performance on PVE is interesting and it suggests that generalization is aided by more diverse data in training.

2.5.3 Analysis

Long tail diversity Most key points and claims are heavily skewed in the number of data points, except for PVE. Even for ARGKP, where key points with few matching arguments were removed, there is a strong imbalance across claims and key points in terms of associated arguments (see Figure 2.1).

Following this imbalance, we sort key points by the number of associated arguments such that the least frequent key points are considered first. Then, we introduce a cutoff parameter f to include arguments from a fraction of key points, starting with the least frequent. Using this parameter we perform matching only on low-frequency key point–arguments matches.

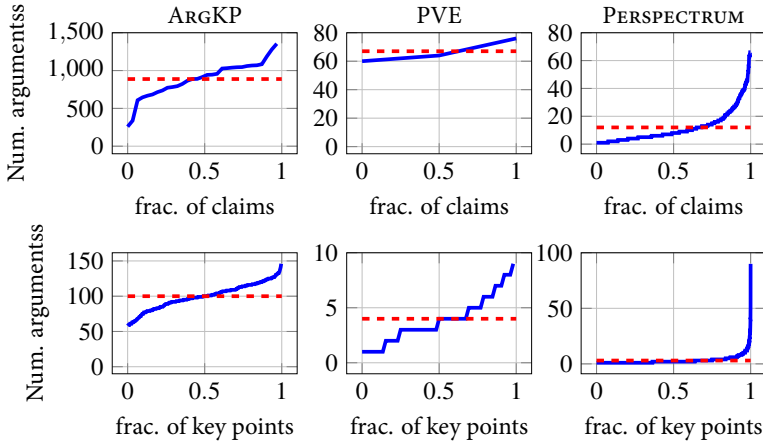


Figure 2.1: Number of arguments matched per claim (upper row) and key point (bottom row), sorted by frequency. The red dashed line shows the average number of arguments.

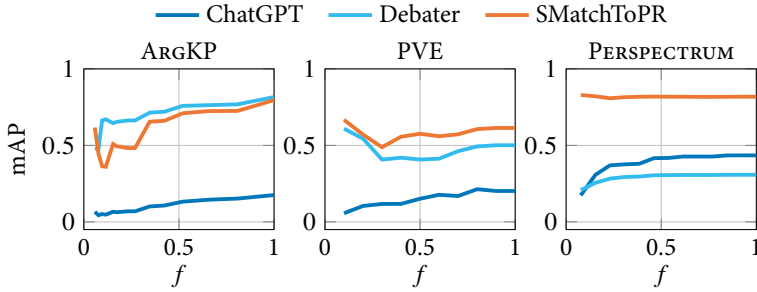


Figure 2.2: KPM performance when limiting data usage to a fraction f , starting with long tail first.

This allows us to investigate the approaches’ performance in the long tail.

When we limit data usage by taking long tail arguments first, the performance of the KPA approaches, mainly on ARGKP and PERSPECTRUM, decreases as shown in Figure 2.2. This shows that the ability to correctly match arguments is contingent on the frequency of the arguments. In some cases, the arguments associated with key points with the fewest matches can be matched, but there is a strong performance loss for low values of f . Across all datasets, ChatGPT suffers consistently in mAP when conditioning on low-frequency key points. For SMatchToPR on PERSPECTRUM, there is almost no effect, showing that representation learning may positively impact the matching of key points to arguments even with low amounts of data. Performing the same experiment for KPG results in similar results: key points with a low number of matched arguments are harder to represent well.

Next, we investigate whether the arguments in the long tail are different from the majority. Here, the long tail consists of arguments for key points that see less than the median number of arguments per key point. We examine whether the sets of lexical items—noun phrase chunks (NPs) and entities—mentioned in the long tail arguments are included in the

Left (long tail)	Right (majority)	NP		Entity		NP- τ	Ent- τ
		left-right	right-left	left-right	right-left		
ARGKP	ARGKP	0.168	0.234	0.191	0.273	0.216*	0.373*
PVE	PVE	0.638	0.787	0.719	0.809	0.521*	0.389
PERSPECTRUM	PERSPECTRUM	0.397	0.807	0.401	0.797	0.361*	0.427*

Table 2.5: Fraction of NPs and Entities in **Left** that are not in **Right** & vice-versa. * indicates Kendall τ with $p < 0.05$.

Approach	PVE		PERSPECTRUM	
	r	p	r	p
ChatGPT	0.030	0.687	0.039	0.469
Debater	0.163	0.029	-0.051	0.013
SMATCH-base	0.097	0.195	0.093	0.215
SMATCH-large	0.207	0.005	-0.03	0.123

Table 2.6: Pearson r correlation scores between predicted match scores and the annotator agreement per sample.

majority and vice versa. We also inspect the relative frequency of the shared lexical items via Kendall τ correlation on the NP and entity frequency rankings. Table 2.5 shows these results.

We see a large overlap of NPs and entities for ARGKP between the long tail and the frequent key points. We attribute this to the filtering of low-frequency data during dataset construction. For the other two datasets, we observe much less overlap—in most cases, more than half of the noun phrases and entities are unique to either part of the dataset. The only exception here is PERSPECTRUM, where roughly 40% of the NPs and entities in the long tail are unique. When comparing the ranks of the intersecting lexical items, we observe moderate (but significant) rank correlation scores. Thus, the overlapping NPs and entities may not be in different frequencies in the two parts of the datasets. However, there is a strong indication of unique items in the long tail, in at least two of our datasets, showing that the long tail may contain novel insights.

Annotator agreement Due to subjectivity in the annotation procedures, we expect annotators to rate argument-key point matches differently. We investigate whether the performance of KPA models reflects this subjectivity. That is, we test if match scores x correlate with the agreement between annotators. Intuitively, when annotators agree, an argument and key point should be considered to match more objectively and thus may be easier to score for a model. From the two datasets that have a per-sample agreement score, we measure the Pearson r correlation between the annotator agreement percentage (as obtained from data) and each approach’s match score $M(d, k)$. Results are shown in Table 2.6.

For all approaches, the correlations are negligible or weak at best [339]. This shows that the predictions made by the models fail to identify which matches are interpreted differently among annotators. Hence, these models are not able to represent the diversity stemming from annotation accurately [302].

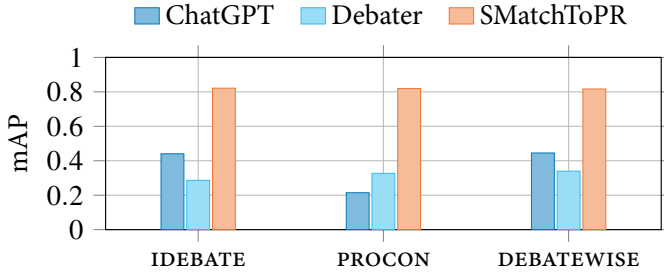


Figure 2.3: KPM performance for all approaches on the different data sources in PERSPECTRUM.

Data sources The KPG and KPM evaluations (Sections 2.5.1 and 2.5.2) indicate how the methods perform when applied to different datasets. The performance is dataset- and task-specific; no single approach performs both tasks best on any dataset. We further investigate the data sources in the PERSPECTRUM dataset, which was constructed using three distinct sources. Figure 2.3 shows the performance on each source separately. Although ARGKP and PERSPECTRUM share a data source, we find no overlapping claims and little repetition in content between the two (App. A.1.1). The SMatchToPR and Debater approaches are not sensitive to data source shift, but ChatGPT performance differs depending on the source data used, dropping considerably for the *procon* source. We find two factors that influence why these arguments are harder to match: (1) *procon* contains about 10 times fewer claims than the other two sources, and (2) *procon*’s arguments are copied verbatim from various cited sources, leading to large stylistic and argumentative differences.

2.6 Conclusion

We perform a novel diversity exploration of different KPA approaches on three distinct datasets. By splitting KPA into two subtasks (KPG and KPM), we investigate each subtask, independently.

First, we find that an LLM-based approach works well for generating key points, but fails to match arguments to key points reliably. Conversely, smaller fine-tuned models are better at matching arguments to key points but struggle to find good key points consistently. Second, using a single training set yields poor generalization across datasets, showing that data source impacts a KPA approach’s ability to generalize. Diversification of training data leads to promising results. Third, across all datasets, we see that existing methods for KPA are insensitive to long tail diversity, decreasing performance for key points supported by few arguments. Finally, all models are insensitive to differences between individual annotators, disregarding subjective interpretations of arguments and key points.

We showed how multiple aspects of diversity, a core principle when interpreting opinions, are not evaluated using the standard set of metrics. Our analysis revealed interesting complementary strengths of the KPA approaches. Future efforts could focus on addressing diversity, either by mining for minority opinions directly [425], or by identifying possibly subjective instances using socio-demographic information [43]. Further, models can be enhanced with subjective understanding [322], or work together with humans to jointly address some of the diversity issues [19, 397].

Limitations

We identify five limitations of our work.

Diversity definition Our definition of diversity is specific to three dimensions, but there may be additional dimensions. For example, our unit of analysis is at the *argument* level. Diversity may also be analyzed for the opinion holders or those affected by decisions in policy-making contexts.

Novel key points Our evaluation of KPG and KPM employs existing key points. However, KPA methods may generate novel or unseen key points. Evaluating such novel key points is nontrivial and it may require experiments involving human subjects.

Resource limitations KPA approaches are resource intensive. We limited some approaches where (1) it would become too expensive to run KPA because of the complexity of the number of comparisons (e.g., Debater approach), or (2) the models do not support a big enough window to fit all arguments (e.g., ChatGPT context window is limited). While there are alternatives (e.g., GPT-4), they drastically increase the cost.

Dataset diversity The arguments in our data are in English, and limited to data gathered from online sources. Further, the users involved in collecting the datasets we employ may not be demographically representative of the global population. We conjecture that increasing the diversity of the data sources would make our conclusions stronger. However, publicly available datasets, especially non-English sources, for this task are scarce. We make our code and experimental data public to incentivize further research in this direction.

Data exposure We cannot verify whether the data from the test sets have been used when training the LLMs. This would make the model familiar with the vocabulary and have a more reliable estimation of the arguments' semantics. That likelihood is the smallest for PVE since it is the most recent dataset, gathered with new crowd workers.

Ethical Considerations

There are growing ethical concerns about NLP (broadly, AI) technology, especially, when the technology is used in sensitive applications. Argument summarization can be used in sensitive applications, e.g., to assist in public policy making. An ethical scrutiny of such methods is necessary before their societal application. Our work contributes toward such scrutiny. The outcome of our analysis shows how KPA methods fail to handle diversity. Potential technological improvements may lead to better results, but due diligence is required before applying such methods to real-world use cases.

We do not collect new data or involve human subjects in this work. Thus, we do not introduce any ethical considerations regarding data collection beyond those that affect the original datasets. A potential concern is that reproducing our results may involve using (possibly paid) services for running KPA. However, we aimed to make the analyses feasible with limited budget and resources.

3

3

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

This Chapter describes our contributions to the Shared Task of the 9th Workshop on Argument Mining (2022). Our approach uses Large Language Models for the task of Argument Quality Prediction. We perform prompt engineering using GPT-3 and investigate the training paradigms of multi-task learning, contrastive learning, and intermediate-task training. We find that a mixed prediction setup outperforms single models. Prompting GPT-3 works best for predicting argument validity, and argument novelty is best estimated by a model trained using all three training paradigms.

3.1 Introduction

As debates are moving increasingly online, automatically processing and moderating arguments becomes essential to further fruitful discussions. The research field of automatic extraction, analysis, and relation detection of argument units is called Argument Mining [AM, 224]. The shared task of the 9th Workshop on Argument Mining (2022) focuses on argument quality [416]. Argument quality can be broken down into multiple dimensions, each with its own purpose, or be extended to *deliberative quality* [414]. In this work, we consider two aspects of the *logical* argument quality dimension: *validity* and *novelty*. Given a premise and a conclusion, a valid relationship indicates that sound logical inferences link the premise and conclusion. A novel relationship indicates that new information was introduced in the conclusion that was not present in the premise.

Prediction of an argument’s validity and novelty can be either through binary classification (Task A) or by explicit comparison between two arguments (Task B). We focus on Task A. A system that is able to estimate validity and novelty could be a building block in AM for online deliberation. For instance, in assisting humans to detect arguments in online deliberative discussions [121, 398] or presenting diverse viewpoints to users in a news recommendation system [318]. We address the task of validity and novelty prediction through a variety of approaches ranging from prompting, contrastive learning, intermediate task training, and multi-task learning. Our best-performing approach is a mix of a GPT-3 model (through prompting) and a contrastively trained multi-task model that uses NLI as an intermediate training task. This approach achieves a combined Validity and Novelty F_1 -score of 0.45.

3.2 Related Work

Given the two related argumentation tasks (novelty and validity), a Multi-Task Learning (MTL) setup [83] is a natural approach. Multi-task models use training signals across several tasks, and have been applied before in argument-related work with Large Language Models (LLMs) [73, 222, 389]. We use shared encoders followed by task-specific classification heads. The training of these encoders was influenced by the following two lines of work.

First, intermediate task training [309, 430] fine-tunes a pre-trained LLM on an auxiliary task before moving on to the final task. This can aid classification performance, also in AM [351]. Second, contrastive learning is shown to be a promising approach [10, 301] in a previous AM shared task [131]. Contrastive learning is used to improve embeddings by forcing similar data points to be closer in space and dissimilar data points to be further away. Such an approach may cause the encoder to learn dataset-specific features that help in downstream task performance.

In addition to MTL, we look at prompt engineering for LLMs, which has shown remarkable progress in a large variety of tasks in combination with [58] or without few-shot learning [364]. For this task we draw inspiration from ProP [8], an approach that ranked first in the “Knowledge Base Construction from Pre-trained Language Models” challenge at ISWC 2022.¹ ProP reports the highest performance with (1) larger LLMs, (2) shorter prompts, (3) diverse and complete examples in the prompt, (4) task-specific prompts.

¹LM-KBC, <https://lm-kbc.github.io/>

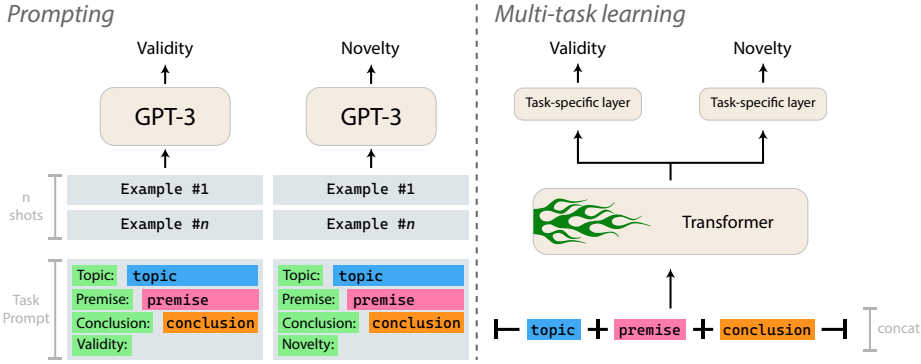


Figure 3.1: The two argument quality prediction setups used in our approach. At inference time, predictions from different setups may be mixed.

Split	Size	Distribution	Topics	Topic Overlap	
				w. train	w. dev
train	750	331/18/296/105	22	–	0
dev	202	33/44/87/38	8	0	–
test	520	110/96/184/130	15	0	8

Table 3.1: Shared task data overview. **Distribution** indicates the class distribution of {non-valid, non-novel}/{non-valid, novel}/{valid, non-novel}/{valid, novel} counts. The red count indicates a severe data imbalance in the training set.

3.3 Data and Training Paradigms

3.3.1 Data

The task data is in American English and consists of Premise, Conclusion, Topic, and a Novel and Validity label. As highlighted in Table 3.1, arguments that are both non-valid and novel are underrepresented in the data. We use the original training and validation distribution as provided and do not use any over- or undersampling strategies. Instead, we opt to resolve the data imbalance by adopting different training paradigms (see Section 3.3.2).

The content included in the dataset concerns common controversial issues popular on debate portals [144], with topics varying from “TV Viewing is Harmful to Children” to “Turkey EU Membership.” The training data also contains classes labeled “defeasibly” valid and “somewhat” novel, which are not in the development or test set. We map these to negative labels (i.e. not novel or not valid) to refrain from discarding data. However, we do not measure the effect of this decision on performance.

3.3.2 Training Paradigms

In our work, we mix different training paradigms to obtain our final approach. A schematic overview is given in Figure 3.1. Below, we outline each of the paradigms individually.

Multi-task Learning Since both validity and novelty are related, a shared encoder is used to process the text input into an embedding, which is fed to task-specific layers. We do not use any parameter freezing, allowing gradients from either task to pass through the entire encoder. During training, a single task is sampled uniformly at random, and a batch is sampled containing instances for that task.

Intermediate task training In our case, we use two related tasks for intermediate task training: Natural Language Inference (NLI) and argument relation prediction. For NLI, we use a released RoBERTa model [248] trained on the MNLI corpus [433], predicting whether two sentences show logical entailment. This is related because making sound logical inferences plays a role in validity. The released argument relation RoBERTa model [327] was trained on the relationship (inference, contradiction, or unrelated) between two sentences in a debate [415]. This is related to novelty and validity. For instance, unrelated arguments may be novel but not valid, and vice versa.

Contrastive Learning We use SimCSE's [134] supervised setting to further fine-tune the previously mentioned RoBERTa MNLI model in a contrastive manner. To train the model we take triples of premises and conclusions in the form of premise, conclusion with a positive novelty rating, and conclusion with a negative novelty rating.

3.4 Approach

Approach 1: GPT-3 Prompting In our prompt-engineering approach, we use OpenAI's GPT-3² [58] for few-shot classification of novelty and validity labels. We construct a prompt by concatenating the topic, premise, and conclusion in a structured format, and request either a validity or novelty label in separate prompts. In addition, we show four static examples before asking for a label from the model, selected from short, difficult examples (i.e. those with the lowest annotation agreement) in the training dataset.

Approach 2: NLI as Intermediate-task, Contrastive learning and Multi-Task Learning This model consists of a shared encoder with task-specific classification heads. We initialize the shared encoder using a pretrained RoBERTa model on the MNLI corpus. We then perform contrastive learning with a triplet loss. Afterward, the model is fine-tuned using MTL on the shared task training data. During training, we switch uniformly at random during training between the novelty and validity tasks.

Approach 3: Mixing Approach 1 (GPT-3) & Approach 2 (NLI+contrastive+MTL) Our Mixed Approach uses Approach 1 (prompt engineering) for validity labels, and Approach 2 (fine-tuned model) for novelty labels.

Approach 4: ArgRel as Intermediate-task and Multi-Task Learning This model uses intermediate-task training on the argument relation prediction task followed by Multi-Task Learning in the same set-up as in Approach 1, but without contrastive learning.

²<https://beta.openai.com/playground>

Model	F1		
	Validity	Novelty	Combined
SVM (TF-IDF + stemming)	0.60	0.08	0.21
GPT-3 (CLTeamL-1)	0.75	0.46	0.35
NLI+ contrastive +MTL (CLTeamL-2)	0.65	0.62	0.39
GPT-3 & NLI+contrastive+MTL (CLTeamL-3)*	0.75	0.62	0.45
ArgRel+MTL (CLTeamL-4)	0.57	0.59	0.33
GPT-3 & ArgRel+MTL (CLTeamL-5)	0.75	0.59	0.43

Table 3.2: Test set performance. CLTeamL- n indicates an official submission to the Shared Task with n corresponding to the Approach number also in Section 3.4. Bold scores indicate the best-performing approach in the shared task. “Combined” indicates the Shared Task organizer’s scoring metric for both tasks.

Approach 5: Mixing Approach 1 (GPT-3) & Approach 4 (ArgRel+MTL) This approach uses Approach 1 (prompt engineering) for validity and Approach 4 (ArgRel+MTL) for novelty labels.

Baseline: SVM Support Vector Machines (SVMs) are strong baselines for argument mining tasks with relatively small multi-topic datasets [319]. We train an SVM separately for validity and novelty as a competitive baseline.

3.4.1 Implementation details

We use Python3 and the HuggingFace transformers [436] framework for training our models. The SVM baseline instead uses sklearn [299]. Our code is publicly available.³ All models trained use RoBERTa (large) [248] as the base model, and the intermediate task trained models are obtained directly from the HuggingFace Hub.⁴ We provide hyperparameters for fine-tuned trained models in Appendix B.1. Model selection was done based on the combined (validity and novelty) F_1 performance on the development set. All experiments were run for 10 epochs, after which the best-performing checkpoint was selected for use in creating predictions on the test set. The training was performed on machines including either two GTX2080 Ti GPUs, or four GTX3090 GPUs.

3.5 Experiments and Results

We compare our approaches’ performance on the test set with the shared task’s metric: Combined F_1 of Validity and Novelty [165]. This combined score is the macro F_1 for predicting validity and novelty in four combinations (valid and novel, valid and not novel, not valid and novel, not valid and not novel). Additionally, we analyze our approaches’ errors and their connection to labels, annotator confidence, and topic. See Table 3.2 for performance on the test set. We also present an SVM-based approach as a baseline.

³<https://github.com/m0re4u/argmining2022>

⁴<https://huggingface.co/>

Model	F1 Validity		F1 Novelty	
	valid	non-valid	novel	non-novel
GPT-3	0.81	0.68	0.26	0.66
MTL	0.80	0.50	0.48	0.75

Table 3.3: Per-label performance on the test set.

	Predicted		
	-	+	
True	-	237	57
	+	184	42

(a) GPT-3

	Predicted		
	-	+	
True	-	265	29
	+	145	81

(b) MTL

Table 3.4: Confusion matrices for the novelty labels.

3.5.1 Error Analysis

We perform additional error analysis on three approaches (Approach 1, 2, and 3). We analyze errors in terms of (1) label-specific performance, (2) annotator confidence, and (3) topics. Additional results are in Appendix B.2.

Per-label performance We observe complementary strengths for the GPT-3 model and our MTL approach in Tables 3.3. The MTL model is remarkably stronger than GPT-3 at identifying *novel* arguments, even when considering this is a low-frequency class. We see a similar trend in terms of misclassifications (Table 3.4), as the MTL model has a 40% lower error rate for the novelty label.

Annotator confidence See Figure 3.2 for the relationship between annotator confidence and classification error. Surprisingly, examples labeled as very confident (easy for human annotators) are not consistently correctly classified by any approach. For novelty, GPT-3 gets about half of these examples wrong.

Topics The 3 topics with the highest error rates differ between approaches and tasks. For validity, GPT-3 struggles with “Was the Iraq War Worth it?” (44.8%), while MTL with “Vegetarianism” (40%). For novelty, GPT-3 also struggles with “Vegetarianism” (60%), and MTL with “Withdrawing from Iraq” (44.7%) and “Vegetarianism” (44%).

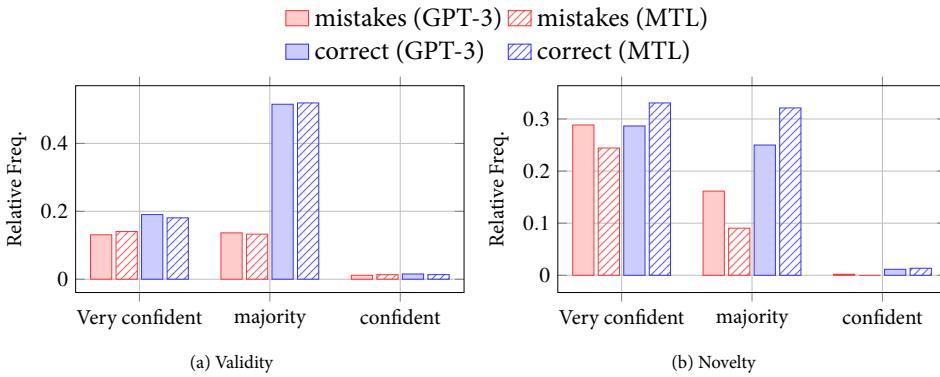


Figure 3.2: Relative accuracy rates divided over label confidence scores.

3.6 Conclusion

We highlight two main conclusions. First, different models have different strengths relating to the two tasks. A prompting approach with a generative model worked best for validity, while contrastive supervised learning worked best for novelty. The two tasks are related enough to be able to effectively use one multi-task learning model, but merging predictions from multiple heterogeneous models leads to the best score. Second, specific intermediate tasks before fine-tuning work well for low-resource argument mining tasks. NLI seems clearly related to validity prediction. For the novelty tasks, other tasks related to argument similarity [315] might be equally informative.

3.7 Access and Responsible Research

A core consideration in NLP research when sharing results is the accessibility and reproducibility of the solution. While our code is openly available, the approaches including GPT-3 require access to commercially trained models. We used free trial OpenAI accounts (allowing \$18 of free GPT-3 credit), but larger datasets and additional tasks can quickly make this approach infeasible. We also considered the freely accessible BLOOM model.⁵ BLOOM does not require payment but does require more GPU memory than what was available to us – making it inaccessible. Ultimately, GPT-3 and related LLMs have several biases and risks of use, including the generation of false information [379] and the fact that their training on internet language leads to a very limited set of language, ideas, and perspectives represented [46], with even racist, sexist, and hateful views [137]. This is especially important to mention, as the task description mentions a future use case of generating new arguments.

⁵<https://huggingface.co/bigscience/bloom>

II

3

Hybrid Intelligence for NLP

Introducing Part II: Hybrid Intelligence for NLP

We have seen how automated approaches to analyzing online discussions suffer from various limitations when applying them to realistic analysis scenarios. We saw how the generalization of LLMs is contingent on the diversity of training data, and how opinion frequency dictates how well a model can capture it, endangering alienating minority opinions. In Part II of this dissertation, we present our approach to incorporating humans and NLP methods for analyzing opinionated text data to address these limitations. First, we introduce a method for mining diverse arguments from citizen feedback in Chapter 4. Our method, HyEnA, finds more diverse arguments and improves the precision of the argument analysis compared to a manual and an automated approach. HyEnA guides human annotators across three distinct phases supported by LLMs for efficient selection of which opinions to analyze.

In Chapter 5, we further investigate how differences between annotators in subjective tasks, such as interpreting texts for extraction of arguments or personal values, can be modeled more efficiently. Our approach, Annotator-Centric Active Learning (ACAL), steers models to learn diverse label distributions by picking from a large pool of annotators. Central to our work, we create discussion analysis approaches that (1) select samples for human inspection that are interesting to annotate, (2) account for diversity, and (3) seek labels from multiple annotators. The hybrid nature of our methodology leads to **bidirectional gains**, serving the NLP system as well as the humans involved. For instance, we create approaches to capture more diverse interpretations of the arguments in discussions using a crowd of annotators. We observe that an active selection of diverse annotators can inform a model more quickly of the label distribution underlying subjective tasks in cases where the annotator pool is large. In Part III, we will apply our hybrid approach to capturing perspectives, and investigate the role of argumentation in constructing faithful opinion representations.

Part II focuses on the following research question:

Q2 How to combine human intelligence and NLP to effectively capture diverse perspectives?

4

A Hybrid Intelligence Method for Argument Mining

4

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.

4.1 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 [215, 226]. 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 [292].

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 [274]. 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) [224] 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 [367, 382, 405]. 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 [112, 315]. Second, although AM methods can identify argumentative content, they often do not compress the information [68, 93, e.g.]. That is, they struggle to recognize whether two arguments describe the same point of view, leaving the policymakers with the significant manual labor of aggregating arguments [209, 210]. Finally, naively relying on a small sample of labeled data might cause minority opinions to be ignored as they are not well represented [204], creating a bias toward popular (repeated) arguments, which can perpetuate echo chambers and filter bubbles [307, 342].

The *key point analysis* (KPA) task [32] 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 [33]. However, the reliance on a 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 4.1.

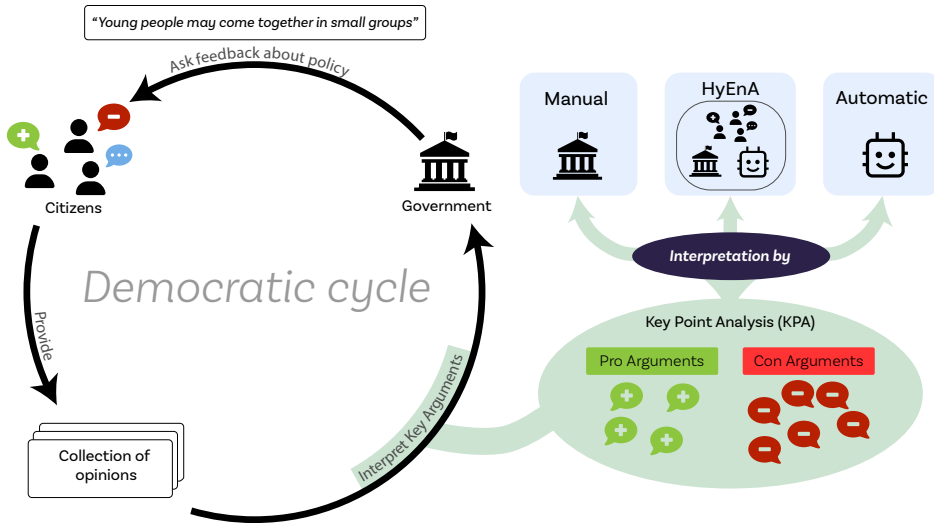


Figure 4.1: In a democratic cycle, citizens provide their opinions on options for governmental decision-making and their opinions need to be interpreted. Insights into the arguments embedded in their comments can be provided by Key Point Analysis (KPA). To perform KPA, most analysis is performed either manually or automatically. In our work, we propose HyEnA, a hybrid method.

HyEnA is evaluated on three corpora, each containing more than 10K public opinions on relaxing COVID-19 restrictions [274]. We compare HyEnA with an automated approach [33] performing the KPA task. In addition, we compare the key arguments generated by HyEnA with manually obtained insights identified by experts [274]. We find that HyEnA outperforms the automated baseline in terms of precision and diversity, specifically when confronted with a set of varied perspectives. 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 (1) We present a hybrid method for key argument extraction, which generates a diverse set of key arguments from a collection of opinionated user comments. (2) 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. (3) 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 Chapter, we provide details on an extended version of the HyEnA method [398, 403]. The original HyEnA method outputs argument clusters, and leverages manual annotations from the first two phases to select arguments from argument clusters. The extension introduces a method for selecting the most representative argument from each cluster.

ter through *argument selection*. The need to summarize argument clusters is not specific to HyEnA, as previous AM applications also retrieve clusters instead of singular arguments [54, 93, 417]. 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 into the HyEnA pipeline, and rerun the original evaluation to compare between HyEnA with and without the inclusion of argument selection. Finally, we perform additional analyses to derive further insights from annotators in HyEnA. We also provide our code, annotation guidelines, and experimental details in the supplementary materials [404].

4

Structure Section 4.2 provides background on Argument Mining for public opinions, and Section 4.3 introduces the HyEnA method for extracting arguments. We outline the experimental setup in Section 4.4 and provide extensive results in Section 4.5. A discussion of our work is given in Section 4.6 and we conclude with Section 4.7.

4.2 Related work

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

4.2.1 Computational Argument Analysis

Argument Mining (AM) methods [62, 224] 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 [407], 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 to reach moderate inter-rater agreement when annotating arguments [381]. Given argumentative texts, modern NLP models are reasonably good at recognizing argumentative discourse within specific contexts [110, 285, 315].

Typically, the first step of AM is to identify the elemental components of arguments (e.g., *claims* and *premises*) in text [296]. 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 [47, 92, 129, 418]. Nonetheless, a few characteristics have been recognized as important for recognizing arguments, namely that arguments (1) contain (informal) logical reasoning [365], (2) address a *why* question [50], and (3) have a non-neutral stance towards the issue being discussed [365].

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.

4.2.2 Summarization of Arguments

Automated methods have been proposed to derive high-level insights from large-scale argumentative content. For instance, these approaches focus on indexing and searching through arguments [366, 439], or creating visual overviews of argument structures [63, 197]. While these may provide access to argumentative content, they are limited in providing a single high-level overview of the arguments on a topic of discussion. Instead, we turn our focus to approaches that create a comprehensible text-based summary from a large corpus of individual comments [33, e.g.]. 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 [32]. Such approaches have been applied in multiple domains, showcasing their applicability across context [34] at varying levels of granularity [66]. While these approaches present high-level arguments, they struggle to capture diversity in opinions, which is important for accommodating multiple perspectives [405]. In this work, we evaluate the performance of these approaches on a novel domain of 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 [81, 314]. They have made a significant impact across a range of tasks [442, 453]. However, downstream applications often need additional fine-tuning [172] in order to perform a task well. They also capture generic aspects of semantic similarity and entailment, which may not be applicable to arguments [314], or overfit to spurious patterns in the data [262]. 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 [419]. 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 [378]. Yet, LLMs might be adequate for specific subtasks, as we showcase in the third phase of the HyEnA method.

4.3 Method

HyEnA is a hybrid method since it combines automated techniques and human judgment [5, 97]. HyEnA guides human annotators in synthesizing *key arguments* (i.e., high-level 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. Key arguments are high-level and summarize a group of arguments, similar to key points as introduced by [32]. We adopt the term key argument, to emphasize their argumentative nature, as opposed to more generic extractive summarization [346, e.g.].

HyEnA consists of three phases (Figure 4.2). 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

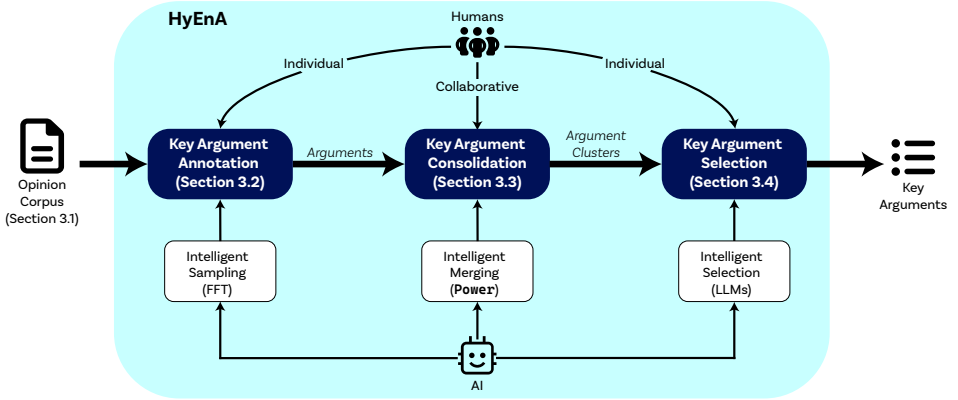


Figure 4.2: Overview of the HyEnA method.

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.

4.3.1 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 [274]. 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 explain 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 data used in our experiments concerns the COVID-19 regulations in the Netherlands during the height of the pandemic, in May 2020. We chose this scenario because (1) we had access to a unique dataset of citizen-provided comments on COVID-19 regulations, (2) we were able to run the study while the topic was still relevant, making it interesting for crowd workers, (3) a manual analysis had been performed over the exact same data, allowing for comparison to a human-only baseline, and (4) the data is reflective of real-world conditions, e.g. feedback was obtained in a matter of days but contains input from a broad group of citizens encompassing broad demographics. We analyze feedback from 26,293 Dutch citizens on three policy options, treating comments on each option as an opinion corpus. Table 4.1 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 [156], 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 chapter.

The original opinions were provided in Dutch. To accommodate a diverse set of anno-

Policy option	Example opinion	Num. 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 4.1: Example opinions in the COVID-19 corpora. The collection of opinions for a policy option forms an opinion corpus.

tators 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 [94, 111, 349]. Although all experiments are conducted in English, the link to the original Dutch text is preserved for future applications.

4.3.2 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) [314]. 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 [37]. 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 [144] to select one single clearest opinion related to the policy option. In this way, we aim to increase 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 [42, 341].

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

Topic Assignment We use a BERTopic [147] 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 a new set of human annotators, different from those in argument extraction, to associate the topics from the generated shortlist with each argument, resulting in an n-hot vector for each argument a per annotator. We obtain the final topic assignment T by summing over all annotators. 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.

4.3.3 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 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 [67], 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 4.2. 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.

Given the similarity scores, we construct a dependency graph G (as in the top-left part of Figure 4.3), where each key argument pair is a vertex in G and the edges indicate a Pareto dependency ($\succ$) between two pairs—the direction of the edge points to the argument pair with greater similarity. A Pareto dependency holds if one of the two scores is strictly greater,

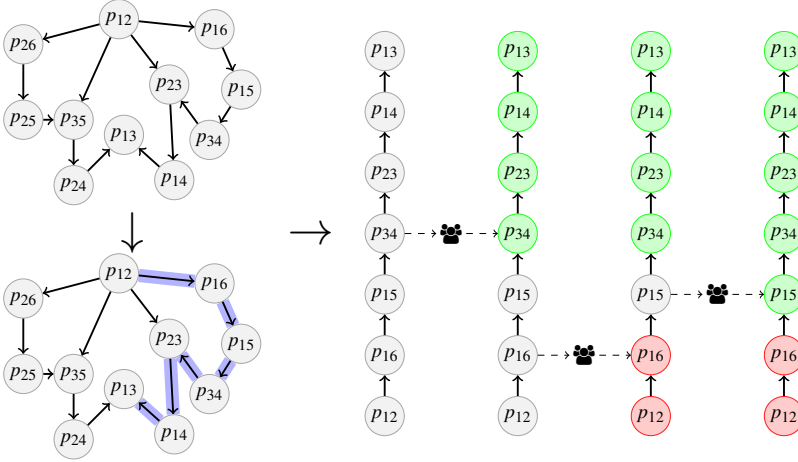


Figure 4.3: Pairwise annotation of the dependency graph, combining human and automatic judgments. Vertices indicate argument pairs; the edge direction points to the argument pair with greater similarity. The highlighted blue edges are a disjoint path selected by the POWER algorithm. Iteratively, vertices are annotated as similar (green) or non-similar (red).

with all others being at least equal between two arguments. We define the dependency as follows:

$$p_{ij} \succeq p_{i'j'} \quad \text{if} \quad \forall n \quad s_{ij}^n \geq s_{i'j'}^n \quad (4.1)$$

$$p_{ij} \succ p_{i'j'} \quad \text{if} \quad p_{ij} \succeq p_{i'j'} \quad \text{and} \quad \exists n \quad s_{ij}^n > s_{i'j'}^n \quad (4.2)$$

Next, we follow POWER to extract disjoint paths from G . The highlighted path in the bottom-left part of Figure 4.3 is an example disjoint path. For every path, we perform a pairwise annotation as in the right part of Figure 4.3. We select the vertex at the middle of the unlabeled portion of the path and ask up to seven humans to indicate whether the concepts described by the two arguments in the pair are similar on a binary scale. The arguments are similar when they are essentially bringing up the same point, i.e. provide the same reasoning. We select the label with the majority vote. Given the annotation, we can automatically label (1) all following pairs in the path as similar (yellow) 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 [67].

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 vertex 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*.

4.3.4 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 [144].

Prompting We prompt an LLM to synthesize a single argument out of the arguments provided in the argument cluster [58]. 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.

4.4 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 [404].

Table 4.3 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

Task	Option	Num. Items	Num. Annotators	Num. Annotators per item
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 4.3: Overview of the tasks in the experiment. Items to be annotated can be opinions (O), arguments (A), topics (T), or combinations. [†] denotes expert annotators.

of the task instances annotated by multiple annotators to investigate the agreement between annotators.

4.4.1 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 annotation feasible within a maximum of one hour. We instantiate the S-BERT model M_S using the Huggingface Model Hub¹. Since our opinion corpus stems from the PVE procedure, we have explicit labels denoting whether a comment was left in favor (*pro*) or opposing (*con*) a proposed policy, which we leverage for the argument stance label suggestion. For obtaining argument quality scores, we use the IBM API [35] to avoid having to retrain a new model.

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

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

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 4.4: Argument selection algorithms.

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

4.4.2 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 [52] uses network modularity to identify groups of vertices based on a resolution parameter r . (2) Self-tuning spectral clustering [446] 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. 4.3. 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. We base E on the homogeneity metric [323], although we do not have access to the ground truth cluster assignments for each argument. Instead, we assume that if all manually labeled arguments are considered similar, they would have been assigned to a single cluster, resulting in a homogenous cluster.

$$E = \frac{1}{|K|} \sum_{k \in K} \frac{\sum_{a_i, a_j \in k} \mathbb{1}_{y(a_i, a_j)=0}}{\binom{|k|}{2}} \quad (4.3)$$

4.4.3 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 4.4. 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.

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 [293]. *Llama* lacks instruction-tuning, and thus requires prompts designed for next-token generation [387]. For the *ChatGPT* model, we instruct it with Prompt 1. For *Llama*, we use Prompt 2.

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 4.3) 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 [449], 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) [120]. 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.

4.4.4 Baselines

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

4

Comparison to Automated Baseline

We use the **ArgKP** argument matching model [33] 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 [144]. 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 [33], 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 [248]. 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 [33]. 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:

$$C_H = \frac{|O_H^{ann}|}{|O_H^{obs}|} \quad (4.4)$$

$$C_A = \frac{|O_A^{ann}|}{|O_A^{obs}|} \quad (4.5)$$

Comparing the coverage scores as defined above naively may not be fair since the set of observed opinions (i.e., the denominators of Equations 4.4 and 4.5) 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:

$$C_H^{common} = \frac{|O_H^{ann} \cap O_A^{obs}|}{|O_H^{obs} \cap O_A^{obs}|} \quad (4.6)$$

$$C_A^{common} = \frac{|O_A^{ann} \cap O_H^{obs}|}{|O_H^{obs} \cap O_A^{obs}|} \quad (4.7)$$

We add the same term to both denominator and numerator in Equations 4.6 and 4.7 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 [33]. 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:

$$P_H^{common} = \frac{\sum_{o_i \in O_H^{ann} \cap O_A^{ann}} z(o_i, \mathcal{A}_H(o_i))}{|O_H^{ann} \cap O_A^{ann}|} \quad (4.8)$$

$$P_A^{common} = \frac{\sum_{o_i \in O_H^{ann} \cap O_A^{ann}} z(o_i, \mathcal{A}_A(o_i))}{|O_H^{ann} \cap O_A^{ann}|} \quad (4.9)$$

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:

$$D_H = \frac{\mathcal{A}_H}{|O_H^{obs} \cap O_A^{obs}|} \quad (4.10)$$

$$D_A = \frac{\mathcal{A}_A}{|O_H^{obs} \cap O_A^{obs}|} \quad (4.11)$$

Comparison to Manual Baseline

A manual analysis involving six experts examined a portion of the feedback stemming from the PVE procedure. This analysis included a sample of participants (2,237 out of 26,293) for whom key arguments were identified [274]. Each expert generated a list of arguments for and against each of the relaxation measures based on the opinion text. A single participant could leave multiple opinions, and the analysis does not report the exact number of opinions analyzed. Since we have access to 36,781 opinions for the three options (Table 4.1), we estimate the number of opinions the six experts would have analyzed to be 3,129 across the three options (following each participant entering ± 1.4 opinions), and at least 2,237 (at

least one opinion per participant). In contrast, HyEnA annotators analyze 765 intelligently selected opinions across the three options.

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.

4.5 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.

4

4.5.1 Annotator Agreement

Table 4.5 shows the inter-rater reliability (IRR) for four steps with overlapping human annotations. We didn'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 $ICC(3, k)$ [353] 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) [357], which adjusts Fleiss' κ for prevalence and bias resulting from small or skewed distribution of ratings.

In Topic Generation, the main source of the disagreement stems 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 an 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.

We obtained the lowest reliability scores for the last two annotation tasks, Key Argument Consolidation and Key Argument Evaluation. 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

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

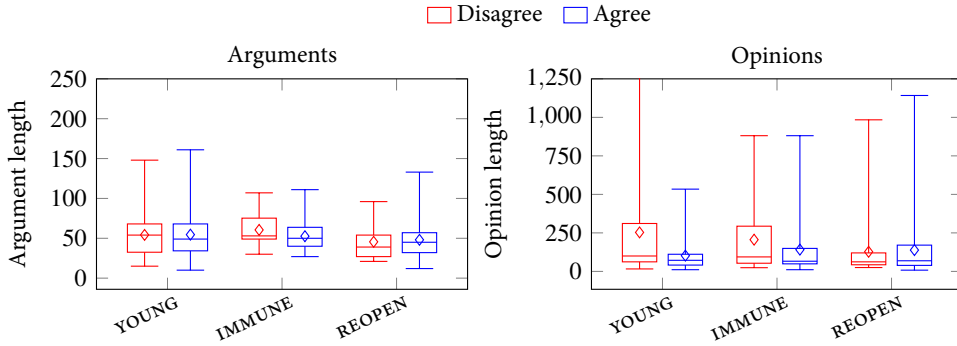


Figure 4.4: Disagreement analysis for the Key Argument Evaluation phase. On the left, argument lengths are the same whether annotators agree or disagree. However, on the right, annotators disagree on match labels in long opinions.

may influence their judgment. Thus, the low IRR scores may indicate a combination of task difficulty and the relatively subjective nature of the task [21]. Similar reasoning holds for the task of evaluating the match between the extracted argument and the original opinions.

Focusing on the evaluation phase, we compare argument–opinion pairs where large disagreement was observed (DISAGREE) to pairs with low disagreement (AGREE) in Figure 4.4. Specifically, we compared the lengths of the arguments and opinions. We find that the lengths of the arguments–opinion pairs with large inter-rater disagreement did not differ from those with low disagreement. However, we found considerably longer opinions on average when annotators disagreed. Possibly, long opinions contain multiple arguments, which in turn may cause the annotator to fail to identify the provided argument.

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 [352].

4.5.2 Phase 1: Key Argument Annotation

In Phase 1, individual annotators were guided through 51 opinions each and asked to annotate the observed arguments. Table 4.6 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,

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

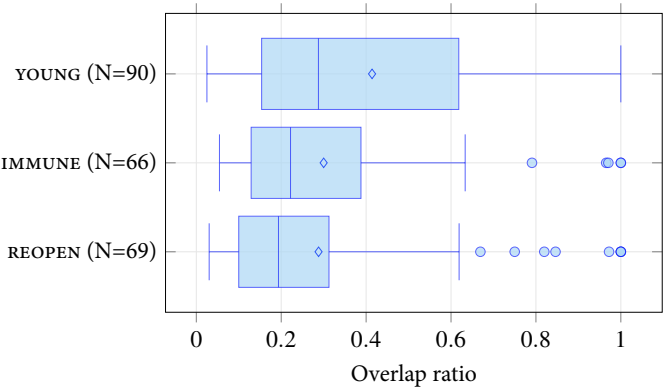


Figure 4.5: Distribution of argument overlap ratio for arguments generated by Key Argument Annotation in Phase 1.

4

IMMUNE, and REOPEN, respectively.

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.

Our instructions did not include an explicit mention of whether copying from the opinion text was allowed, but we observed that annotators often paraphrased arguments from opinions. To examine the behavior of the annotators, we measured the amount of text that was literally copied from the opinions. To do so, we take the largest common substring on the character level between opinion text and argument and divide it by the length of the argument. In Figure 4.5, 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. 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 4.7 shows some examples of arguments extracted with different overlap ratios.

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

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 4.7: Examples of extracted arguments in Phase 1 of HyEnA. Overlapping character sequences are highlighted in green.

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 4.8: Expert topic generation statistics in Phase 1.

feasible. Our experts eliminated one, two, and zero topics as duplicates in the three options (Table 4.8). 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 4.9 shows examples from the final list of topics, with low-scoring topics removed.

4.5.3 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 4.6 (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 [283] measure the extent to which transitivity holds among the similarity labels of argument pairs. The low τ scores indicate the need for subsequent clustering, given that there are no clear graph components in which all arguments are similar.

Figure 4.6 compares Louvain and spectral clustering for extracting argument clusters.

Option	Clarity Rating	Topic words
YOUNG	4.5	immune entertainment hospitality restrictions
	4	infection immunity risk infected
	4	<i>virus susceptible spread transmit</i>
	4.5	<i>schools reopen education students</i>
	5	<i>risk limited low dangerous</i>
	5	<i>group risk target least</i>
IMMUNE	4.5	homes nursing care vulnerable
	4	netherlands country provinces dutch
	5	<i>risk contamination danger dangerous</i>
	4.5	<i>work companies home economy</i>
	5	entertainment hospitality catering industry
REOPEN	5	homes nursing care vulnerable
	4	netherlands friesland groningen dutch
	5	risk hospitality entertainment dangerous
	3	<i>mental health income decrease</i>
	3	<i>measures rules these should</i>

Table 4.9: Examples of topics generated in Phase 1, including the top 4 words and the average clarity rating. Option-specific topics are *emphasized*.

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

4.5.4 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 4.4.3 for additional details). Then, we evaluate the quality of the arguments that are selected to represent clusters.

Odd-one-out task Figure 4.8 shows the results of the odd-one-out evaluation. We perform pairwise statistical analysis by employing McNemar’s test [101] with Holm-Bonferroni correction on multiple tests [4]. The test results indicate whether methods significantly differ

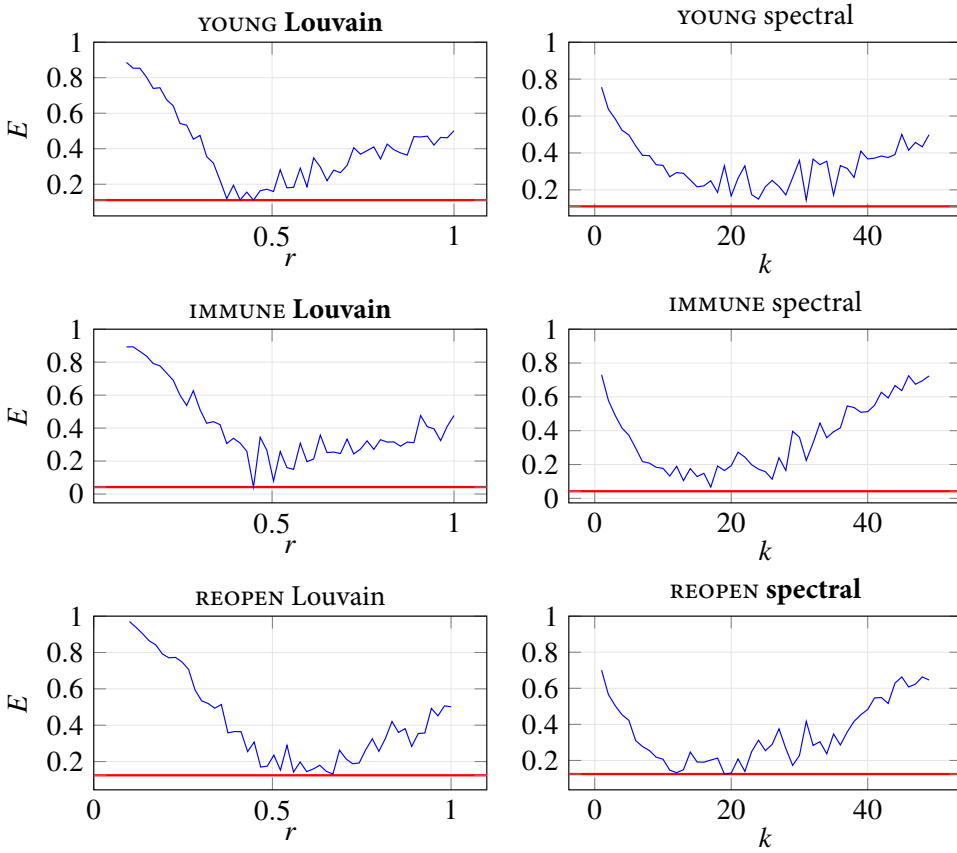


Figure 4.6: Error rate E for different parameters per clustering method (resolution parameter r for Louvain, k clusters for spectral) for each corpus in Phase 2.

in their misclassifications. We observe that only *Llama-random* does not have a significant difference in error proportions and can thus be assumed to perform similarly to each other. Conversely, two out of three 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.

Evaluating Argument Selection To select the best-performing Argument Selection method, we compare BERTScores in Figure 4.9. We use the Kruskal-Wallis test (a non-parametric alternative to ANOVA since the scores are not normally distributed) to test whether all medians are equal at a 5% significance level [212]. Since we obtain a score well below our

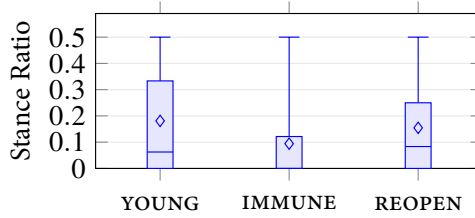


Figure 4.7: Stance distribution for clusters extracted for each corpus in Phase 2. A ratio of 0.5 denotes an equal number of pro and con arguments inside a cluster.

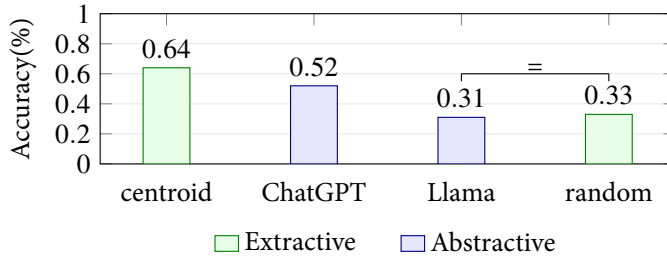


Figure 4.8: Accuracy on the odd-one-out task per method. Key Argument Selection methods marked with = do not significantly differ ($p < 0.05$) in their error proportions.

threshold, we conduct a post-hoc follow-up to identify pairs of significantly different Key Argument Selection methods. We employ Dunn’s multiple comparisons of mean rank sums [107] with Holm-Bonferroni correction on multiple tests [4].

All extractive methods have a higher standard deviation than the generative methods. Some selected representative arguments likely caused the high maxima for extractive methods, since they are copied verbatim from opinions in the corpus. Conversely, the low minima are due to the extractive methods’ inability to find representatives from the cluster (since there may be noisy clusters, see Figure 4.8). For the abstractive methods, the lower bound is higher, showing how rephrasing the selected argument makes it more related to all arguments inside a cluster. Between the abstractive methods, ChatGPT has a higher standard deviation than Llama. Since we did not perform extensive prompt engineering, there is room for improvement in both methods with better-crafted prompts.

The only significantly different method is Llama, with all others achieving similar BERT-Score performance. 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 because the argument clusters are sufficiently coherent. However, in the final evaluation, humans will be judging the match between selected arguments and individual opinions. Here, 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 (the abstractive methods) and higher median score (ChatGPT outperforms Llama significantly), 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

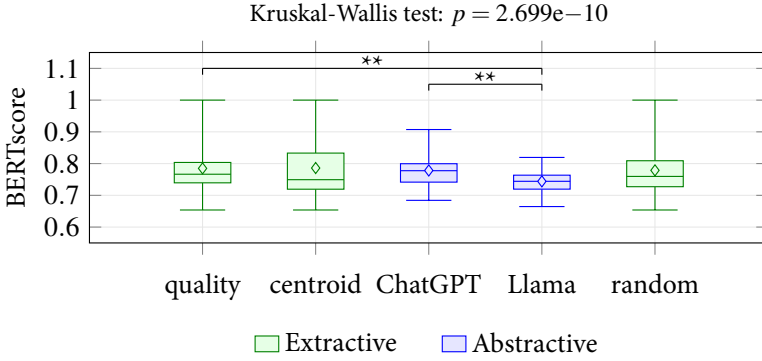


Figure 4.9: Aggregated BERTScore for the different Key Argument Selection methods across all corpora and argument clusters (Phase 3). Method pairs indicated by ** differ significantly from each other in median performance ($p < 0.05$).

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

was manual. In particular, we take the extractions from Phase 1 and re-evaluate them using a new set of annotators. In Table 4.10, we show the difference in Precision (Equation 4.8).

We find that the addition of Argument Selection on average has a slight negative impact on the ability of annotators to match opinions and arguments. 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 makes annotating the match for some opinions to selected key arguments easier while making others more difficult. Selecting arguments using ChatGPT generates key arguments that are representative of the entire cluster, which can be more general than the arguments extracted by annotators from individual opinions. On the one hand, this can cause external annotators to not recognize the specific argument from a given opinion. On the other hand, it may result in annotators matching opinions and arguments on a more abstract level.

4.5.5 Comparison with Automated Baseline

Figure 4.10 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 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 4.1)

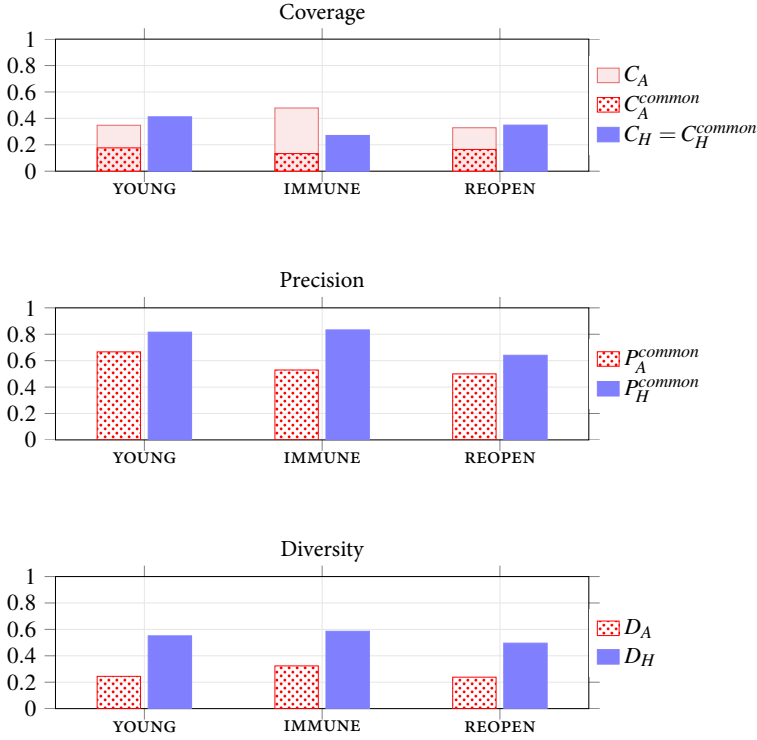


Figure 4.10: Comparing HyEnA and ArgKP.

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).

4.5.6 Comparison with Manual Baseline

Table 4.11 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

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

Table 4.11: A confusion matrix comparing the key argument lists generated by HyEnA and manual baseline. The complete mapping is given in Appendix C.3.

the lower human effort, the HyEnA lists largely overlap with the expert lists.

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 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.

4.6 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, can 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, Llama). Those models may be employed for KPA by using prompting techniques [245]. The capabilities of these models seem to imply that they have access to higher order argumentation knowledge [223], 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, and the danger of models synthesizing ungrounded arguments exists [185]. 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. [28], Van Der Meer et al. [402]) 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. Furthermore, future versions of HyEnA could benefit from investigating why annotators disagreed on labels in each phase, as it can lead to possible improvements in the annotation task.

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 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 All parts of the HyEnA pipeline are open to adjustment and can be performed by humans, machines, or a combination. In this work, we presented a specific version of this pipeline, but other ways of combining humans and AI are possible. However, the impact of choosing specific components remains unclear for parts of the pipeline, since we experimented with a single algorithm in some cases (e.g., the use of POWER in Key Argument Consolidation, or the LLMs in Key Argument Selection).

HyEnA presents a general framework that allows individual phases to be supported by different types of technologies and different groups of crowd/expert annotators. Within this hybrid framework, we considered the following criteria when deciding to allocate tasks to humans or 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 evaluation (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 [5]. 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 swapping out components is not straightforward and requires considerable amounts of work. We envision research to come up with similar use cases where HI can make a significant impact.

4.7 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 [238, 400], sentiment, emotion, and attribution [411]. 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.

Ethical Considerations

This chapter 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.

5

Annotator-Centric Active Learning for Subjective NLP Tasks

5

Active Learning (AL) addresses the high costs of collecting human annotations by strategically annotating the most informative samples. However, for subjective NLP tasks, incorporating a wide range of perspectives in the annotation process is crucial to capture the variability in human judgments. We introduce Annotator-Centric Active Learning (ACAL), which incorporates an annotator selection strategy following data sampling. Our objective is two-fold: (1) to efficiently approximate the full diversity of human judgments, and (2) to assess model performance using annotator-centric metrics, which value minority and majority perspectives equally. We experiment with multiple annotator selection strategies across seven subjective NLP tasks, employing both traditional and novel, human-centered evaluation metrics. Our findings indicate that ACAL improves data efficiency and excels in annotator-centric performance evaluations. However, its success depends on the availability of a sufficiently large and diverse pool of annotators to sample from.

5.1 Introduction

A challenging aspect of natural language understanding (NLU) is the variability of human judgment and interpretation in subjective tasks (e.g., hate speech detection) [302]. In a subjective task, a data sample is typically labeled by a set of annotators, and differences in annotation are reconciled via majority voting, resulting in a single (supposedly, true) “gold label” [393]. However, this approach has been criticized for treating label variation exclusively as noise, which is especially problematic in sensitive subjective tasks [21] since it can lead to the exclusion of minority voices [228].

Subjectivity can be addressed by modeling the full distribution of annotations for each data sample instead of employing gold labels [302]. However, resources for such approaches are scarce, as most datasets do not (yet) make fine-grained annotation details available [61], and representing a full range of perspectives is contingent on obtaining costly annotations from a diverse set of annotators [28].

One way to handle a limited annotation budget is to use Active Learning [350, AL]. Given a pool of unannotated data samples, AL employs a sample selection strategy to obtain maximally informative samples, retrieving the corresponding annotations from a ground truth oracle (e.g., a single human expert). However, in subjective tasks, there is no such oracle. Instead, we rely on a set of available annotators. Demanding all available annotators to annotate all samples would provide a truthful representation of the annotation distribution, but is often unfeasible, especially if the pool of annotators is large. Thus, deciding *which annotator(s)* should annotate is as critical as deciding which samples to annotate.

In most practical applications, annotators are randomly selected. This results in an annotation distribution insensitive to outlier annotators—most annotations reflect the majority voices and fewer reflect the minority voices. This may not be desirable in applications such as hate speech, where the opinions of the majority and minority should be valued equally. In such cases, a more deliberate annotator selection is required. To ensure a balanced representation of majority and minority voices, we leverage strategies inspired by Rawls’ principle of fairness [313], which advocates that a fair society is achieved when the well-being of the worst-off members of society (the minority annotators, in this case) is maximized.

We introduce Annotator-Centric Active Learning (ACAL) to emphasize and control who annotates which sample. In ACAL (Figure 5.1), the sample selection strategy of traditional AL is followed by an *annotator selection strategy*, indicating which of the available annotators should annotate each selected data sample.

Contributions (1) We present ACAL as an extension of the AL approach and introduce three annotator selection strategies aimed at collecting a balanced distribution of minority and majority annotations. (2) We introduce a suite of annotator-centric evaluation metrics to measure how individual and minority annotators are modeled. (3) We demonstrate ACAL’s effectiveness in three datasets with subjective tasks—hate speech detection, moral value classification, and safety judgments.

Our experiments show that the proposed ACAL methods can approximate the distribution of human judgments similar to AL while requiring a lower annotation budget and modeling individual and minority voices more accurately. However, our evaluation shows how the task’s annotator agreement and the number of available annotations impact ACAL’s effectiveness—ACAL is most effective when a large pool of diverse annotators is available.

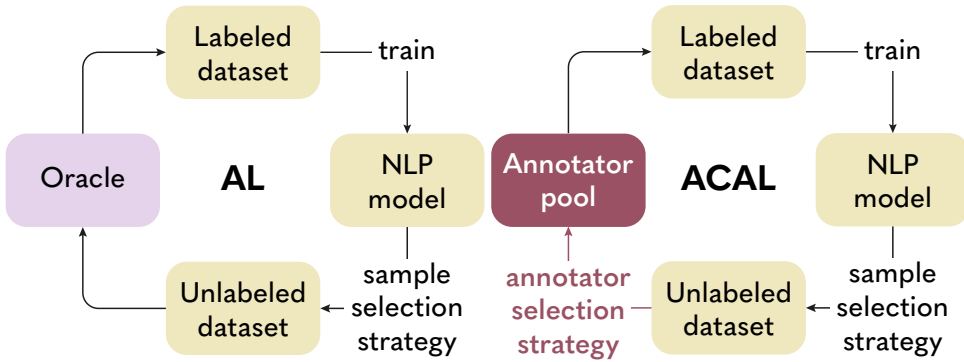


Figure 5.1: Active Learning (AL) approaches (left) use a sample selection strategy to pick samples to be annotated by an oracle. The Annotator-Centric Active Learning (ACAL) approach (right) extends AL by introducing an annotator selection strategy to choose the annotators who annotate the selected samples.

Importantly, our experiments show how the ACAL framework controls how models learn to represent majority and minority annotations. This is crucial for subjective and sensitive applications such as detecting human values and morality [203, 239], argument mining [405], and hate speech [198].

5

5.2 Related work

5.2.1 Learning with annotator disagreement

Modeling annotator disagreement is garnering increasing attention [21, 61, 302, 393]. Changing annotation aggregation methods can lead to a fairer representation than simple majority [171, 380]. Alternatively, the full annotation distribution can be modeled using soft labels [79, 277, 300]. Other approaches leverage annotator-specific information, e.g., by including individual classification heads per annotator [89], embedding annotator behavior [269], or encoding the annotator’s socio-demographic information [44]. Yet, modeling annotator diversity remains challenging. Standard calibration metrics under human label variation may be unsuitable, especially when the variation is high [24]. Trade-offs ought to be made between collecting more samples or more annotations [149]. Further, solely measuring differences among sociodemographic traits is not sufficient to capture opinion diversity [291]. Instead, we represent diversity based on *which* annotators annotated *what* and *how*. We experiment with annotator selection strategies to reveal what aspects impact task performance and annotation budget.

5.2.2 Active Learning

AL enables a supervised learning model to achieve high performance by judiciously choosing a few training examples [350]. In a typical AL scenario, a large collection of unlabeled data is available, and an oracle (e.g., a human expert) is asked to annotate this unlabeled data. A *sampling strategy* is used to iteratively select the next batch of unlabeled data for annotation [316]. AL has found widespread application in NLP [451]. Two main strategies are employed, either by selecting the unlabeled samples on which the model prediction is

most uncertain [450], or by selecting samples that are most representative of the unlabeled dataset [116, 452]. The combination of AL and annotator diversity is a novel direction. Existing works propose to align model and annotator uncertainties [39], adapt annotator-specific classification heads in AL settings [421], or select texts to annotate based on annotator preferences [192]. These methods ignore a crucial part of learning with human variation: the diversity among annotators. We focus on selecting annotators such that they best inform us about the underlying label diversity.

5.3 Method

First, we define the soft-label prediction task we use to train a supervised model. Then, we introduce the traditional AL and the novel ACAL approaches.

5.3.1 Soft-label prediction

Consider a dataset of triples $\{x_i, a_j, y_{ij}\}$, where x_i is a data sample (i.e., a piece of text) and $y_{ij} \in C$ is the class label assigned by annotator a_j . The multiple labels assigned to a sample x_i by the different annotators are usually combined into an aggregated label $\hat{y}_i$. For training with soft labels (i.e. non-binary class assignment), the aggregation typically takes the form of maximum likelihood estimation [393]:

$$\hat{y}_i(x) = \frac{\sum_{i=1}^N [x_i = x][y_{ij} = c]}{\sum_{i=1}^N [x_i = x]} \quad (5.1)$$

In our experiments, we use a passive learning approach that uses all available $\{x_i, \hat{y}_i\}$ to train a model f_θ with cross-entropy loss as a baseline.

5.3.2 Active Learning

AL imposes a sampling technique for inputs x_i , such that the most *informative* sample(s) are picked for learning. In a typical AL approach, a set of unlabelled data points U is available. At every iteration, a sample selection strategy $\mathcal{S}$ selects samples $x_i \in U$ to be annotated by an oracle $\mathcal{O}$ that provides the ground truth label distribution $\hat{y}_i$. The selected samples and annotations are added to the labeled data D , with which the model f_θ is trained. Alg. 1 provides an overview of the procedure.

Algorithm 1: AL approach.

input: Unlabeled data U , Data sampling strategy $\mathcal{S}$, Oracle $\mathcal{O}$
 $D_0 \leftarrow \{\}$
for $n = 1..N$ **do**
 sample data points x_i from U using $\mathcal{S}$
 obtain annotation $\hat{y}_i$ for x_i from $\mathcal{O}$ $D_{n+1} = D_n + \{x_i, \hat{y}_i\}$
 train f_θ on D_{n+1}
end

In the sample selection strategies, a batch of data of a given size B is queried at each iteration. Our experiments compare the following strategies:

Random ($\mathcal{S}_R$) selects a B samples uniformly at random from U .

Uncertainty ($\mathcal{S}_U$) predicts a distribution over class labels with $f_\theta(x_i)$ for each $x_i \in U$, and selects B samples with the highest prediction entropy (the samples the model is most uncertain about).

5.3.3 Annotator-Centric Active Learning

ACAL builds on AL. In contrast to AL, which retrieves an aggregated annotation $\hat{y}_i$, ACAL employs an annotator selection strategy $\mathcal{T}$ to select one annotator and their annotation for each selected data point x_i . Alg. 2 describes the ACAL approach.

Algorithm 2: ACAL approach.

```

input: Unlabeled data  $U$ , Data sampling strategy  $\mathcal{S}$ , Annotator sampling strategy  $\mathcal{T}$ 
 $D_0 \leftarrow \{\}$ 
for  $n = 1..N$  do
    sample data points  $x_i$  from  $U$  using  $\mathcal{S}$ 
    sample annotators  $a_j$  for  $x_i$  using  $\mathcal{T}$ 
    obtain annotation  $y_{ij}$  from  $a_j$  for  $x_i$ 
     $D_{n+1} = D_n + \{x_i, y_{ij}\}$ 
    train  $f_\theta$  on  $D_{n+1}$ 
end

```

5

We propose three annotator selection strategies to gather a distribution that uniformly contains all possible (majority and minority) labels, inspired by Rawls' principle of fairness [313]. The strategies vary in the type of information used to represent differences between annotators, including *what* or *how* the annotators have annotated thus far. Our experiments compare the following strategies:

Random ($\mathcal{T}_R$) randomly selects an annotator a_j .

Label Minority ($\mathcal{T}_L$) considers only information on *how* each annotator has annotated so far (i.e., the labels that they have assigned). The minority label is selected as the class with the smallest annotation count in the available dataset D_n thus far. Given a new sample, x_i , $\mathcal{T}_L$ selects the available annotator that has the largest bias toward the minority label compared to the other available annotators, i.e., who has annotated other samples with the minority label the most.

Semantic Diversity ($\mathcal{T}_S$) considers only information on *what* each annotator has annotated so far (i.e., the samples that they have annotated). Given a new sample x_i selected through $\mathcal{S}$, $\mathcal{T}_S$ selects the available annotator for whom x_i is semantically the most different from what the annotator has labeled so far. To measure this difference for an annotator a_j , we employ a sentence embedding model to measure the cosine distance between the embeddings of x_i and embeddings of all the samples annotated by a_j . We then take the average of all semantic similarities. The annotator with the lowest average similarity score is selected.

Representation Diversity ($\mathcal{T}_D$) selects the annotator that has the lowest similarity on average with all other annotators available for that item. We create a representation for each annotator by averaging the embeddings of samples annotated by a_j together with their respective labels, followed by computing the pair-wise cosine similarity between all annotators.

5.4 Experimental Setup

We describe the experimental setup for the comparisons between ACAL strategies. In all our experiments, we employ a TinyBERT model [187] to reduce the number of trainable parameters. Appendix D.1 includes a detailed overview of the computational setup and hyperparameters. We make the code for the ACAL strategies and evaluation metrics available via GitHub.¹

5.4.1 Datasets

We use three datasets which vary in domain, annotation task (in *italics*), annotator count, and annotations per instance.

The **DICES Corpus** [22] is composed of 990 conversations with an LLM where 172 annotators provided judgments on whether a generated response can be deemed safe (3-way judgments: yes, no, unsure). Samples have 73 annotations on average. We perform a multi-class classification of the judgments.

The **MFTC Corpus** [169] is composed of 35K tweets that 23 annotators annotated with any of the 10 moral elements from the Moral Foundation Theory [142]. We select the elements of *loyalty* (lowest annotation count), *care* (average count), and *betrayal* (highest count). Samples have 4 annotations on average. We create three binary classifications to predict the presence of the respective elements. As most tweets were labeled as non-moral (i.e., with no moral element), we balanced the datasets by subsampling the non-moral class.

The **MHS Corpus** [328] consists of 50K social media comments on which 8K annotators judged three hate speech aspects—*dehumanize* (low inter-rater agreement), *respect* (medium agreement), and *genocide* (high agreement)—on a 5-point Likert scale. Samples have 3 annotations on average. We perform a multi-class classification with the annotated Likert scores for each task.

The datasets and tasks differ in levels of annotator agreement, measured via entropy of the annotation distribution. DICES and MHS generally have medium entropy scores, whereas the MFTC entropy is highly polarized (divided between samples with very high and very low agreement). Appendix D.1.5 provides details of the entropy scores.

5.4.2 Evaluation metrics

The ACAL strategies aim to guide the model to learn a representative distribution of the annotator’s perspectives while reducing annotation effort. To this end, we evaluate the model both with a traditional evaluation metric and a metric aimed at comparing predicted and annotated distributions:

Macro F_1 -score (F_1) For each sample in the test set, we select the label predicted by the model with the highest confidence, determine the golden label through a majority agreement aggregation, and compute the resulting macro F_1 -score.

Jensen-Shannon Divergence (JS) The JS measures the divergence between the distribution of label annotation and prediction [286]. We report the average JS for the samples in the test set to measure how well the model can represent the annotation distribution.

¹<https://github.com/m0re4u/acal-subjective>

Further, since ACAL shifts the focus to annotators, we introduce novel annotator-centric evaluation metrics. First, we report the average among annotators. Second, in line with Rawls' principle of fairness, the result for the worst-off annotators:

Per-annotator F_1 (F_1^a) and JS (JS^a) We compute the F_1 (or JS) for each annotator in the test set using their annotations as golden labels (or target distribution), and average it.

Worst per-annotator F_1 (F_1^w) and JS (JS^w) We compute the F_1 (or JS) for each annotator in the test set using their annotations as golden labels (or target distribution), and report the average of the lowest 10% to mitigate noise.

These metrics allow us to measure the trade-offs between modeling the majority agreement, a representative distribution of annotations, and accounting for minority voices. In the next section, we describe how we obtained the results.

5.4.3 Training procedure

We test the annotator selection strategies proposed in Section 5.3.3 by comparing all combinations of the two sample selection strategies ($\mathcal{S}_R$ and $\mathcal{S}_U$) and the four annotator selection strategies ($\mathcal{T}_R$, $\mathcal{T}_L$, $\mathcal{T}_S$, and $\mathcal{T}_D$). At each iteration, we use $\mathcal{S}$ to select B unique samples from the unlabeled data pool U . We select B as the smallest between 5% of the number of available annotations and the number of unique samples in the training set. For each selected sample x_i , we use $\mathcal{T}$ to select one annotator and retrieve their annotation y_{ij} .

We split each dataset into 80% train, 10% validation, and 10% test. We start the training procedure with a warmup iteration of B randomly selected annotations [451]. We proceed with the ACAL iterations by combining $\mathcal{S}$ and $\mathcal{T}$. We select the model checkpoint across all AL iterations that led to the best JS performance on the validation set and evaluate it on the test set. We repeat this process across three data splits and model initializations. We report the average scores on the test set.

We compare ACAL with traditional oracle-based AL approaches ($\mathcal{S}_R\mathcal{O}$ and $\mathcal{S}_U\mathcal{O}$), which use the data sampling strategies but obtain all possible annotations for each sample as in Alg. 1. Further, we employ a passive learning (PL) approach as an upper bound by training the model on the full dataset, thus observing all available samples and annotations. Similar to ACAL, the AL and PL baselines are averaged over three seeds.

5.5 Results

We start by highlighting the benefits of ACAL over AL and PL (Section 5.5.1). Next, we closely examine ACAL on efficiency and fairness (Section 5.5.2). Then, we select a few cases of interest and dive deeper into the strategies' behavior during training (Section 5.5.3). Finally, we investigate ACAL across varying levels of subjectivity (Section 5.5.4).

5.5.1 Highlights

Our experiments show that ACAL can have a beneficial impact over using PL and AL. Figure 5.2 highlights two main findings: (1) ACAL strategies can more quickly learn to represent the annotation distribution with a large pool of annotators, and (2) when agreement between annotators is polarized, ACAL leads to improved results compared to learning from aggregated labels. In the next sections, we provide a deeper understanding of the conditions in which ACAL works well.

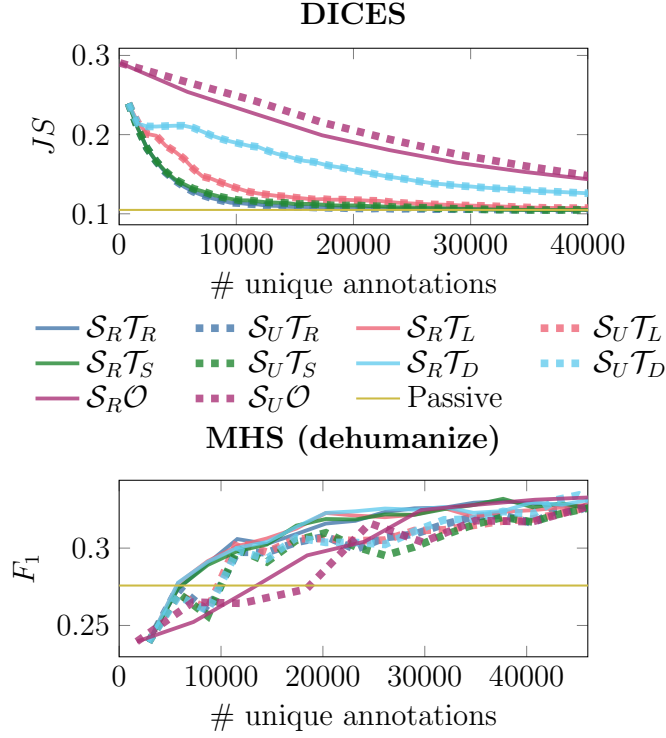


Figure 5.2: Learning curves showing model performance on the validation set. On DICES (upper), ACAL approaches are quicker than AL in obtaining similar performance to passive learning. On MHS (lower), ACAL surpasses passive learning in F_1 when data has high disagreement.

5.5.2 Efficiency and Fairness

Table 5.1 presents the results of evaluating the best models (those with the highest JS scores on the validation set) on the test set. We analyze the results along two dimensions: (a) *efficiency*: what is the impact of the different strategies on the trade-off between annotation budget and performance? (b) *fairness*: do the selection strategies that aim for a balanced consideration of minority and majority views lead to better performance in the human-centric evaluation metrics? For MFTC we focus on *care* because it has an average number of samples available, and for MHS we focus on *dehumanize* because it has high levels of disagreement. Appendix C.3 presents the remainder of the results.

Efficiency We discuss the performance on F_1 and JS to measure how well the proposed strategies model label distributions and examine the used annotator budget. Across all tasks and datasets, ACAL and AL consistently yield comparable or superior F_1 and JS with a lower annotation budget than PL. When comparing ACAL with AL, the results vary depending on the task and dataset. For DICES, there is a significant benefit to using ACAL, as it can save up to $\sim 40\%$ of the annotation budget while yielding better scores across all metrics than AL. With AL, we observe only a small reduction in annotation cost. For MFTC, AL with

	App.	F_1	JS	Average		Worst-off		$\Delta\%$
				F_1^a	JS^a	F_1^w	JS^w	
DICES	$\mathcal{S}_R \mathcal{T}_R$	53.2	.100	43.2	.186	16.7	.453	-36.8
	$\mathcal{S}_R \mathcal{T}_L$	55.5	.101	42.4	.187	15.5	.450	-32.7
	$\mathcal{S}_R \mathcal{T}_S$	61.0	.103	44.2	.186	16.4	.447	-35.5
	$\mathcal{S}_R \mathcal{T}_D$	58.9	.142	43.1	.203	16.9	.370	-30.0
	$\mathcal{S}_U \mathcal{T}_R$	53.2	.100	43.2	.186	16.7	.453	-36.8
	$\mathcal{S}_U \mathcal{T}_L$	55.5	.101	42.4	.187	15.5	.450	-32.7
	$\mathcal{S}_U \mathcal{T}_S$	63.1	.098	43.9	.187	18.4	.447	-38.2
	$\mathcal{S}_U \mathcal{T}_D$	58.9	.142	43.1	.203	16.9	.370	-30.0
	$\mathcal{S}_R \mathcal{O}$	59.1	.112	41.4	.191	13.3	.425	-0.1
	$\mathcal{S}_U \mathcal{O}$	46.2	.110	38.4	.192	11.7	.427	-0.1
MFTC (<i>care</i>)	PL	59.0	.105	37.1	.211	12.3	.479	-
	$\mathcal{S}_R \mathcal{T}_R$	78.9	.038	61.1	.141	37.7	.247	-1.6
	$\mathcal{S}_R \mathcal{T}_L$	78.5	.037	61.6	.142	39.2	.249	-0.4
	$\mathcal{S}_R \mathcal{T}_S$	78.1	.039	60.0	.145	35.1	.248	-1.7
	$\mathcal{S}_R \mathcal{T}_D$	76.6	.040	60.4	.144	35.7	.243	-1.7
	$\mathcal{S}_U \mathcal{T}_R$	79.4	.038	61.2	.143	37.7	.252	-5.6
	$\mathcal{S}_U \mathcal{T}_L$	80.7	.037	58.9	.142	42.3	.248	-2.5
	$\mathcal{S}_U \mathcal{T}_S$	79.1	.037	60.8	.143	39.9	.258	-1.1
	$\mathcal{S}_U \mathcal{T}_D$	78.1	.040	58.6	.145	35.7	.253	-2.5
	$\mathcal{S}_R \mathcal{O}$	79.0	.037	58.6	.141	39.2	.255	-0.2
MHS (<i>dehumanize</i>)	$\mathcal{S}_U \mathcal{O}$	79.4	.037	58.3	.144	35.7	.253	-12.7
	PL	81.1	.032	51.2	.179	37.7	.251	-
	$\mathcal{S}_R \mathcal{T}_R$	33.6	.081	31.5	.394	0.0	.489	-50.0
	$\mathcal{S}_R \mathcal{T}_L$	33.1	.081	32.2	.397	0.0	.478	-62.5
	$\mathcal{S}_R \mathcal{T}_S$	30.5	.079	31.3	.397	0.0	.480	-62.5
	$\mathcal{S}_R \mathcal{T}_D$	32.4	.081	31.8	.398	0.0	.479	-62.5
	$\mathcal{S}_U \mathcal{T}_R$	32.4	.080	32.2	.389	0.0	.508	-7.8
	$\mathcal{S}_U \mathcal{T}_L$	33.1	.080	32.8	.388	0.0	.507	-7.8
	$\mathcal{S}_U \mathcal{T}_S$	33.6	.080	32.6	.388	0.0	.506	-7.8
	$\mathcal{S}_U \mathcal{T}_D$	33.0	.079	32.6	.384	0.0	.513	-3.0
	$\mathcal{S}_R \mathcal{O}$	32.8	.077	33.9	.387	0.0	.496	-60.1
	$\mathcal{S}_U \mathcal{O}$	33.3	.080	33.1	.390	0.0	.497	-24.7
	PL	28.0	.075	20.2	.424	0.0	.547	-

Table 5.1: Test set results on the DICES, MFTC (*care*), and MHS (*dehumanize*) datasets. Results report the average test scores from the best-performing model checkpoint on the validation set (lowest JS), evaluated across three data splits and model initializations. $\Delta\%$ denotes the reduction in the annotation budget with respect to passive learning. In bold, the best performance per column and per dataset (higher F_1 are better, lower JS are better).

$\mathcal{S}_U$ leads to the largest cost benefits ($\sim 12\%$ less annotation budget), but at a cost in terms of absolute JS and F_1 . ACAL slightly outperforms AL but does not lead to a decrease in annotation budget. For MHS, both AL and ACAL significantly reduce the annotation cost ($\sim 60\%$) while yielding better scores than PL—however, AL and ACAL do not show substantial performance differences. Overall, when looking at F_1 and JS which are aggregated over

Fairness We investigate the extent to which the models represent individual annotators fairly and capture minority opinions via the annotator-centric evaluation metrics (F_1^a , JS^a , F_1^w , and JS^w). We observe a substantial improvement when using AL or ACAL over PL. Further, we observe no single winner-takes-all approach: high F_1 and JS scores do not consistently co-occur with high scores for the annotator-centric metrics. This highlights the need for a more comprehensive evaluation to assess models for subjective tasks. Yet, we observe that ACAL slightly outperforms AL in modeling individual annotators (JS^a and F_1^a). This trend is particularly evident with DICES, again likely due to the large pool of annotators available per data sample. Lastly, ACAL is best in the worst-off metrics (JS^w and F_1^w), showing the ability to better represent minority opinions as a direct consequence of the proposed annotator selection strategies on DICES and MFTC. However, all approaches score 0 for F_1^w on MHS. This is due to the high disagreement in this dataset: the 10% worst-off annotators always disagree with a hard label derived from the predicted label distribution. In conclusion, our experiments show that, when a large pool of annotators is available, a targeted sampling of annotators requires fewer annotations and is fairer. That is, minority opinions are better represented without large sacrifices in performance compared to the overall label distribution.

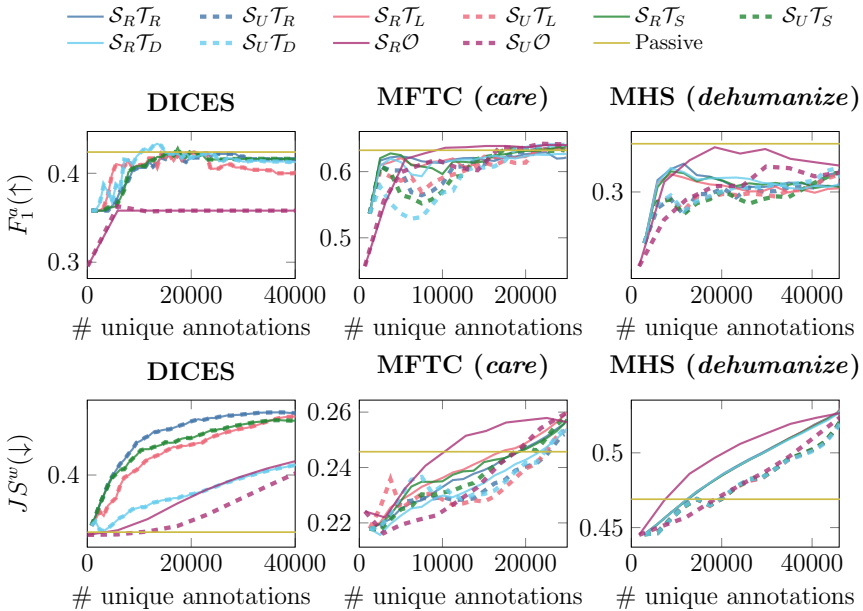


Figure 5.3: Selected plots showing the F_1^a and JS^w performance on the validation set during the ACAL and AL iterations for DICES, MFTC (*care*), and MHS (*dehumanize*). Higher F_1^a is better, lower JS^w is better. Y-axes are scaled to highlight the relative performance to PL.

5.5.3 Convergence

The evaluation on the test set paints a general picture of the advantage of using ACAL over AL or PL. In this section, we assess how different ACAL strategies converge over iterations. We describe the major patterns across our experiments by analyzing six examples of interest with F_1^a and JS^w (Figure 5.3). We select F_1^a because it reveals how well individual annotators are modeled on average, and JS^w to measure how strategies deviate from modeling the majority perspective. Appendix D.2.2 provides an overview of all metrics.

First, we notice that the trends for F_1^a and JS^w are both increasing—the first is expected, but the second requires an explanation. As the model is exposed to more annotations over the training iterations, the predicted label distribution starts to fit the true label distribution. However, here we consider each annotator individually: JS^w reports the average of the 10% lowest JS scores per annotator. The presence of disagreement implies the existence of annotators that annotate differently from the majority. Since our models predict the full distribution, they assign a proportional probability to dissenting annotators. Thus, learning to model the full distribution of annotations leads to an increase in JS^w .

Second, we notice a difference between ACAL and AL. On MFTC and MHS, ACAL, compared to AL, yields overall smaller JS^w at the cost of a slower convergence in F_1^a , showing the trade-off between modeling all annotators and representing minorities. However, with DICES the trend is the opposite. This is due to AL having access to the complete label distribution: it can model a balanced distribution, leading to lower worst-off performance. With a large number of annotations, ACAL requires more iterations to get the same balanced predicted distribution.

Third, we observe differences among the annotator selection strategies ($\mathcal{T}$). $\mathcal{T}_D$ shows the most differences—both JS^w and F_1^a increase slower than for the other strategies. This suggests that selecting annotators based on the average embedding of the annotated content strongest emphasizes diverging label behavior.

Finally, we analyze the impact of the sample selection strategies ($\mathcal{S}$, dotted vs. solid lines in Figure 5.3). For DICES, $\mathcal{S}_R$ and $\mathcal{S}_U$ lead to comparable results, likely due to the low number of samples. Using $\mathcal{S}_U$ in MFTC leads to F_1^a performance decreasing at the start of training. The strategy prioritizes obtaining annotations for already added samples to lower their entropy, while the variation in labels is irreconcilable (since there are limited labels available, and they are in disagreement). We see a similar pattern for MHS.

These results further underline our main finding that ACAL is effective in representing diverse annotation perspectives when there is a (1) heterogeneous pool of annotators, and (2) a task that facilitates human label variation.

5.5.4 Impact of subjectivity

We further investigate ACAL strategies on (1) label entropy, and (2) cross-task performance.

Alignment of ACAL strategies during training We want to investigate how well the ACAL strategies align with the overall subjective annotations: do they drive the model entropy in the right direction? We measure the entropy of the samples in the labeled training set at each iteration and compare it to the entropy of all annotations of those samples. Higher entropy in the labeled training set than the actual entropy suggests that the selection strategy overestimates uncertainty. Lower entropy indicates that the model may not sufficiently account

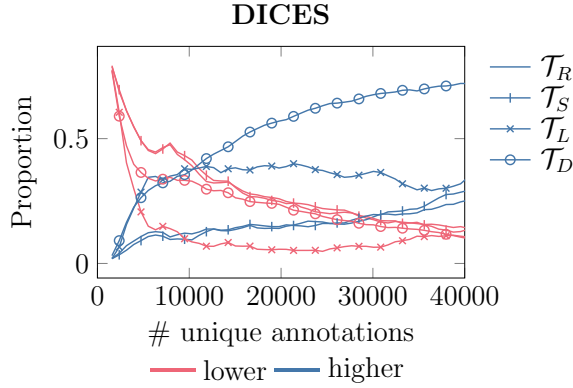


Figure 5.4: Proportion of data samples that result in higher or lower entropy than the target label distribution per ACAL strategy.

5

for disagreement. When the entropy matches the true entropy, the selection strategy is well-calibrated to strike a healthy middle ground between sampling diverse labels and finding the majority class. We focus on DICES as a case study due to the wide range of entropy scores. We group each sample based on the true label entropy into low (< 0.43), medium ($0.43 - 0.72$), and high (> 0.72). We apply the same categorization at each training iteration for samples labeled thus far. Subsequently, we plot the proportion of data points for which the selection strategy results in excessively high or excessively low entropy.

Figure 5.4 visualizes the proportions. At the beginning of training, entropy is generally low because samples have few annotations. Over time, the selected annotations better align with the true entropy. At the start (at 10K unique annotations), roughly only a third of the samples have aligned entropy scores ($T_R = 27\%$, $T_S = 27\%$, $T_L = 33\%$, $T_D = 32\%$). Further towards the end of the ACAL iterations, this has increased for all ACAL strategies except T_D ($T_R = 64\%$, $T_S = 62\%$, $T_L = 57\%$, $T_D = 17\%$). When and how much the strategies succeed in matching the true label distribution differs: T_S and T_R take longer to increase label entropy than the other two strategies. They are conservative in adding diverse labels. T_L and T_D increase the proportion of well-aligned data points earlier in the training process, achieving a balanced entropy alignment sooner. However, both strategies start to overshoot the target entropy, whereas the others show a more gradual alignment with the true entropy. This effect is strongest for T_D . This finding suggests that minority-aware annotator-selection (T_L and T_D) strategies achieve the best results in the early stages of training—that is, they are effective for quickly raising entropy but can lead to overrepresentation.

Cross-task performance Figure 5.5 compares the two annotator-centric metrics on the three tasks of MFTC and MHS—the datasets for which we have seen the least impact of ACAL over AL and PL. We select a data sampling (S_R) and annotator sampling strategy (T_S), based on its strong performance on DICES for comprehensive comparison.

When evaluating MFTC *loyalty*, which has the highest disagreement, JS^w is more accurately approximated with PL. Similarly, ACAL is outperformed by AL on F_1^a for the *de-humanize* (high disagreement) task. However, for the less subjective task *genocide*, ACAL

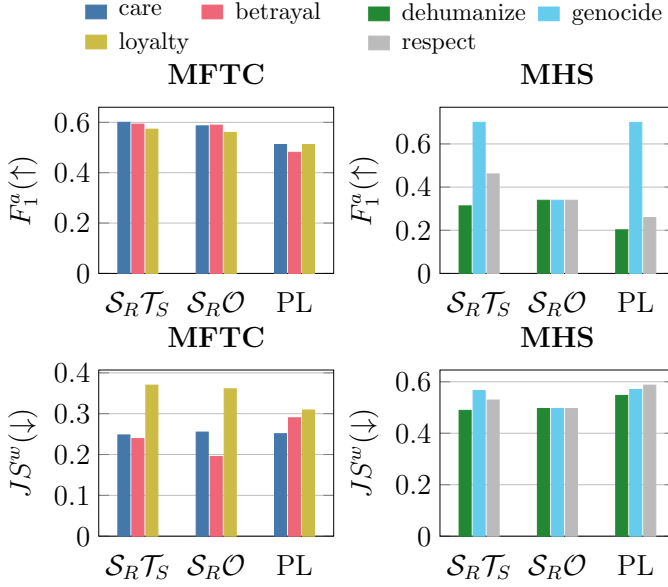


Figure 5.5: Comparison of ACAL, AL, and PL across different MFTC and MHS tasks. Higher F_1^a is better, and lower JS^w is better.

leads to higher F_1^a . This suggests that the effectiveness of annotation strategies varies depending on the task's degree of subjectivity *and* the available pool of annotators. The more heterogeneous the annotation behavior, indicative of a highly subjective task, the larger the pool of annotators required for each sample selection. We also observe that there is a trade-off between modeling the majority of annotators equally (F_1^a) and prioritizing the minority (JS^w).

5.6 Conclusion

We present ACAL as an extension of AL to emphasize the selection of diverse annotators. We introduce three novel annotator selection strategies and four annotator-centric metrics and experiment with tasks across three different datasets. We find that the ACAL approach is especially effective in reducing the annotation budget when the pool of available annotators is large. However, its effectiveness is contingent on data characteristics such as the number of annotations per sample, the number of annotations per annotator, and the nature of disagreement in the task annotations. Furthermore, our novel evaluation metrics display the trade-off between modeling overall distributions of annotations and adequately accounting for minority voices, showing that different strategies can be tailored to meet different goals. Especially early in the training process, strategies that are aggressive in obtaining diverse labels have a beneficial impact in accounting for minority voices. However, we recognize that gathering a distribution that uniformly contains all possible (minority and majority) labels can be overly sensitive to small minorities or noise. Future work should integrate methods that account for noisy annotations [426]. Striking a balance between utilitarian and egalitarian

tarian approaches, such as between modeling aggregated distributions and accounting for minority voices [229] is crucial for inferring context-dependent values [242, 400].

Limitations

The main limitation of this work is that the experiments are based on simulated AL which is known to bear several shortcomings [261]. In our study, a primary challenge arises with two of the datasets (MFTC, MHS), which, despite having a large pool of annotators, lack annotations from every annotator for each item. Consequently, in real-world scenarios, the annotator selection strategies for these datasets would benefit from access to a more extensive pool of annotators. This limitation likely contributes to the underperformance of ACAL on these datasets compared to DICES. We emphasize the need for more datasets that feature a greater number of annotations per item, as this would significantly enhance research efforts aimed at modeling human disagreement.

Since we evaluate four different annotator selection strategies and two sample selection strategies across three datasets and seven tasks, the amount of experiments is high. This did not allow for further investigation of other methods for measuring uncertainty such as ensemble methods [218], different classification models, the extensive turning of hyperparameters, or even different training paradigms like low-rank adaptation [173]. Lastly, a limitation of our annotator selection strategies is that they rely on a small annotation history. This is why we require a warmup phase for some of the strategies, for which we decided to take a random sample of annotations. Incorporating informed warmup strategies, incorporating ACAL strategies that do not rely on annotator history, or making use of more elaborate hybrid human-AI approaches [403] may positively impact its performance and data efficiency.

Ethical Considerations

Our goal is to approximate a good representation of human judgments over subjective tasks. We want to highlight the fact that the *performance* of the models differs a lot depending on which metric is used. We tried to account for a less majority-focussed view when evaluating the models which is very important, especially for more human-centered applications, such as hate-speech detection. However, the evaluation metrics we use do not fully capture the diversity of human *judgments*, but just that of *labeling behavior*. The selection of metrics should align with the specific goals and motivations of the application, and there is a pressing need to develop more metrics to accurately reflect human variability in these tasks.

Our experiments are conducted on English datasets due to the scarcity of unaggregated datasets in other languages. In principle, ACAL can be applied to other languages (given the availability of multilingual models to semantically embed textual items for some particular strategies used in this work). We encourage the community to enrich the dataset landscape by incorporating more perspective-oriented datasets in various languages, ACAL potentially offers a more efficient method for creating such datasets in real-world scenarios.

III

Social Science with Hybrid Intelligence

Introducing Part III: Social Science with Hybrid Intelligence

In Part III, we make use of our hybrid setup described in Part II to investigate the perspectives of participants in online discussions at scale. We leverage insights into how to strategically incorporate human input with LLMs and design a collaborative process where humans can contribute to the discussion analysis. We make use of human annotators to provide a nuanced understanding of their motivations behind online opinions while using LLMs for processing large-scale data. The HI setup enables us to obtain a deeper understanding of the multifaceted nature of online discussions, and also explore the different capabilities of humans and LLMs when extracting high-level insights from discussions. In Chapter 6, we acquire deep representations using the Perspective Hierarchy from online social media comments and examine the connection between value conflicts and disagreements on societally relevant topics. We observe that value conflicts lead to disagreements in cases where values are likely to be relevant and diverse. In other cases, we need additional information to create a complete perspective. Our approach shows that arguments are a crucial component in productively revealing the rationale behind opinions.

Part III focuses on the following research question:

Q3 How to construct a perspective hierarchy based on diverse opinions in a discussion?

6

Do Differences in Values Influence Disagreements in Online Discussions?

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

6

6.1 Introduction

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

We want to ensure that a discussion incorporates many perspectives and that those are actively communicated. For this reason, we turn to *value conflicts*, a potential root cause for disagreement. We consider the hypothesis that when users with conflicting values engage in a discussion, diverging views come up. Perspective and value clashes are at the heart of disagreement [371]. In collaborative teams, value conflicts are linked to disagreement [182]. Specifically, values are said to be an effective way to make conflict explicit among participants in a discussion [41]. To evaluate our hypothesis, we construct value profiles based on user comments on Reddit, a social media platform. A value profile captures the relative importance a user ascribes to values. We employ ten values, e.g., stimulation, universalism, and security, from the well-known Schwartz theory of basic values [344]. Then, we compare the similarities among profiles to the disagreement among users on different topics. This allows us to investigate the association between value conflict (low similarity) and disagreement. Figure 6.1 shows an overview of our approach.

We gather 11.4M comments from 19K users on Reddit to construct value profiles. We perform up to 200 tests with different settings to investigate our hypothesis. We further experiment with replacing estimated value profiles with self-reported ones. To do so, we collect 572 judgments from 26 annotators in combination with self-reported value profiles. Selecting conversation partners based on their profile to manage value conflicts and influence the level of disagreement in a discussion could be a tool for moderators to balance conversations. To provide support for moderators, we investigate the impact of adding profile information to the agreement analysis task [305]. Since the contextual implications of values are usually unknown, connecting user concerns to values [11] opens up human-machine collaboration opportunities for a more constructive conversation [5, 158, 238].

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

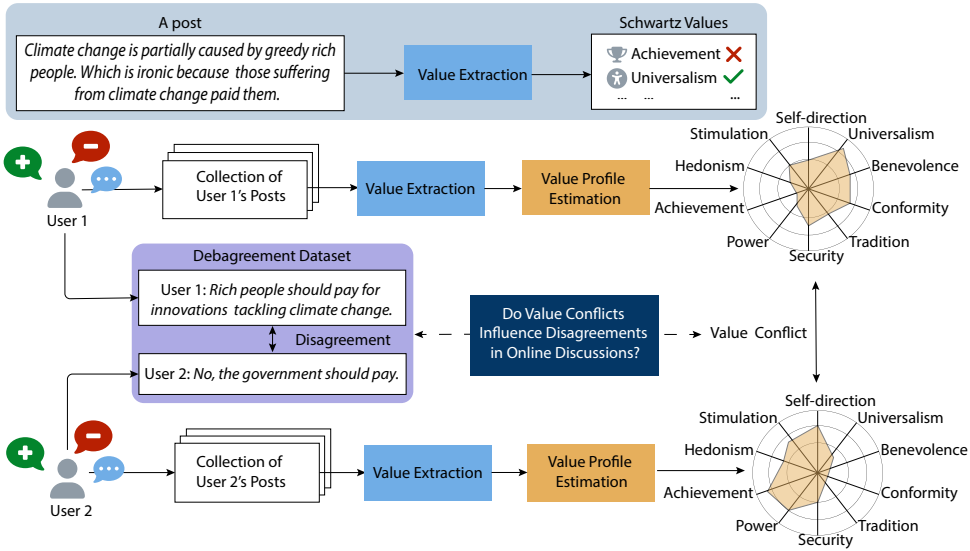


Figure 6.1: Setup of measuring value conflicts by means of Value Profile Estimation (VPE).

6.2 Related Work

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

6.2.1 (Dis)-agreement and discussion analysis

Detecting whether people agree or disagree with given statements is commonly framed as stance classification [e.g., 7]. Recently, more effort has been put into exploring various aspects of the task [9, 161, 246]. However, little work is done in adjusting the task to detect stances among users within online discussions. To this end, **agreement analysis** focuses on detecting (dis-)agreement in data that (1) represents realistic online discussions, (2) provides contextual information (post authors, timestamps, etc.), (3) contains diverse writing styles, (4) touches on multiple topics [305].

Existing work on agreement analysis is aimed at (1) identifying language that indicates disagreement [e.g., 126, 284, 434], (2) leveraging stylistic choices like sarcasm for detecting disagreement [139], (3) finding stance and target pairs, followed by the traditional stance classification [e.g., 71, 96], and (4) mixing detailed opinion information using e.g., logic of evaluation [104]. Recently, adding social role context to textual comments was shown to have a positive impact on the agreement analysis task [255], which indicates the usefulness of background information. In this work, we focus on capturing the implicit motivations underlying opinions using *personal values*, which have been known to drive individual opinions and actions across cultures [344].

6.2.2 Value models

Values explain ideological beliefs underlying actions and opinions and may guide the design of applications [130]. Two leading value models have been used in NLP research: Moral Foundations [143] and the Schwartz Value model [344]. Each of these models includes a set of general values. The Moral Foundation Theory (MFT) includes five foundations, each a vice–virtue dichotomy (e.g., *harm–care*). However, MFT does not stipulate any relationship among the foundations. In contrast, the Schwartz model includes ten basic values organized as a circumplex (right-hand side of Figure 6.1), where similar values are placed close to each other. Further, Schwartz values can be grouped into four classes: *openness to change*, *conservation*, and *self-transcendence*, *self-enhancement*. Since the Schwartz model has more values and a structure among the values, it is better suited than MFT for comparing the value profiles of individuals. Thus, we employ Schwartz values in our work.

6.2.3 Value estimation

Most works based on representing an individual’s value priorities (value profiles) use explicit preference elicitation, such as self-reporting and questionnaires [e.g., 57]. However, a promising behavior-based approach focuses on analyzing textual motivations [70]. To this end, dictionary-based approaches can be used for finding value mentions in texts [141, 304]. Using such lexicons shows promising results in large-scale value estimation applications [356].

Recently, datasets annotated with personal values for training NLP methods have been released. In this chapter, we use two recent datasets annotated with Schwartz values: (1) ValueNet [311] is a dataset containing textual scenarios related to moral decision-making that have been annotated with relevant Schwartz values. (2) ValueArg [201] contains user-submitted arguments that relate to specific Schwartz values. There are some datasets on MFT values, e.g., [169, 253, 388]. These datasets include value annotations for messages but do not include a link between the messages and users. Thus, estimating value profiles from such datasets is not possible.

Applications include dialogues about moral scenarios [311], review texts [288], and value-laden arguments [11, 207]. However, both the annotation and extraction of values remain difficult, with specific questions relating to the granularity of the value labels [201], their transfer to new domains [237], and how classifiers understand morality in language [239]. Moreover, large variances exist between the frequency of values across domain [196], and even the relevance of values differs depending on the domain [55, 235]. However, users can still be represented inside each domain by examining relative frequencies inside value profiles, as stipulated by Schwartz [344].

6.3 Method

Figure 6.1 shows an overview of our method. We collect posts from users in online discussions. Using a trained value estimation model, we aggregate predictions over the collection to form a value profile. Then, to evaluate our hypothesis, we compare the value profiles for users known to be in disagreement based on an existing dataset. Our code¹ and data [401] is available online.

¹<https://github.com/m0re4u/value-disagreement>

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

Table 6.1: List of subcorpora gathered in Debagreement.

6.3.1 Data

We use **Debagreement** [305] as the dataset containing (dis-)agreement labels. This dataset contains user-submitted post pairs in English from five topics (Table 6.1), with post pairs annotated as {agree, neutral, disagree} by at least three crowd annotators.

We gather additional posts through the Reddit API using the usernames available in the Debagreement dataset. For each user still active, we collect up to 1000 most recent posts, which can be in any subreddit. The resulting posts range from September 2015 to April 2022. Subreddits host content on a variety of topics, not all of which encourage users to provide opinions based on their values. We are interested in finding preferences among values with respect to widespread societal issues, such as climate change. Thus, we filter out posts that are not likely to be of relevance to such issues. We (1) exclude Not Safe For Work and entertainment-related subreddits, removing 1.4M posts, (2) filter out noisy low-frequency subreddits (those with less than 50 collected posts), removing an additional 850K posts, and (3) retain only English text posts, removing 377K posts. Table 6.1 shows the amount of data collected after filtering.

6.3.2 Value Extraction

We formulate the value estimation task as recognizing whether a comment is related to a value by means of binary classification per value, matching the setup of Qiu et al. [311]. Our training data comprises general texts annotated for the presence of values across multiple domains. We combine data from two sources.

- (1) **ValueNet** [311]: We collapse non-neutral labels (1 and -1) into a single positive class and take the neutral labels (0) as a negative class. A non-neutral utility means that annotators considered the value to be relevant to the scenario, whereas the neutral class indicates that the value plays no apparent role.
- (2) **ValueArg** [201]: Their annotation scheme uses an updated (20) Schwartz values [345], which we map back to the original 10 Schwartz values to allow joint training with the ValueNet dataset.

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

6.3.3 Value Profile Estimation

Using a trained model, we construct a value profile v per user by summing over value estimations of all individual messages. We assume relative frequencies of value mentions to be indicative of value preference similar to Siebert et al. [354].

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

Kendall τ We sort value mentions by frequency and assign a rank label to each value. Kendall's rank correlation metric τ is a robust measure of correlation [85], and considers the ranks of all pairs of values. If a pair of values is ranked differently in v than in w , the pair is considered discordant. Low scores indicate value conflict.

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

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

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

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

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

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

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

6.4 Experiments and Results

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

6.4.1 Training Models for Value Estimation

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

Method	Training	Test		
		ValueNet	ValueArg	Both
All-ones	–	0.40	0.11	0.26
Value Dict.	–	0.45	0.64	0.57
Kiesel et al. [201]*	ValueArg	0.15	0.37	0.28
Qiu et al. [311]*	ValueNet	0.59	0.52	0.57
BERT _{VE}	ValueNet	0.66	0.57	0.65
	ValueArg	0.46	0.76	0.67
	Both	0.63	0.81	0.79
RoBERTa _{VE}	ValueNet	0.62	0.59	0.63
	ValueArg	0.46	0.76	0.67
	Both	0.63	0.78	0.78

Table 6.2: Macro-averaged F_1 scores of the value estimation approaches on the value datasets. Methods marked with * are adapted for our comparison.

comment (“All-ones”) to examine label imbalance, (2) predict values based on mentions of value words from the **Schwartz Value Dictionary** [304], (3) the multi-label approach from Kiesel et al. [201], which uses an expanded label set, and (4) the utility model from Qiu et al. [311]. The latter two baselines are BERT-based models. For Kiesel et al. [201], we use their multi-label setup to make predictions and map to the 10 Schwartz values at inference time (*humility* and *face* are not mapped to any value). Similarly, we map the rounded ternary utility labels from Qiu et al. [311] into binary value relevance labels at inference.

Table 6.2 shows the F_1 scores for the value extraction methods for different combinations of training and test datasets. We outperform all our baselines, including those from previous work. BERT_{VE} and RoBERTa_{VE} yield similar F_1 scores, and they perform best when trained on both datasets. We use our best-performing BERT_{VE} model, trained on *both* datasets, to construct the value profiles in the rest of the experiments.

6.4.2 Value Profile Estimation

Table 6.3 shows the top two frequent values in each domain. We observe that the distribution of values is specific to discussion contexts. For example, although stimulation is a common and frequent value, it is not the most frequent value in the BREXIT subcorpus. We aggregate the values extracted for each user into their value profile. Table 6.3 (last column) shows the mean pairwise τ distance (Equation 6.1) among the value profiles in each domain. We observe that the BLM subcorpus has the most diversity among the five subcorpora.

Next, to qualitatively assess the estimated value profiles, we normalize profiles (by the total number of value mentions) and compute covariance between profiles. Then, we perform metric Multi-Dimensional Scaling (MDS) of the covariance matrix similar to Ponizovskiy et al. [304]. Figure 6.2 shows a visualization of the first two dimensions after MDS. We observe that values that are close to each other in the Schwartz circumplex [344], e.g., achievement and power, also tend to be closer in the MDS visualization.

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

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

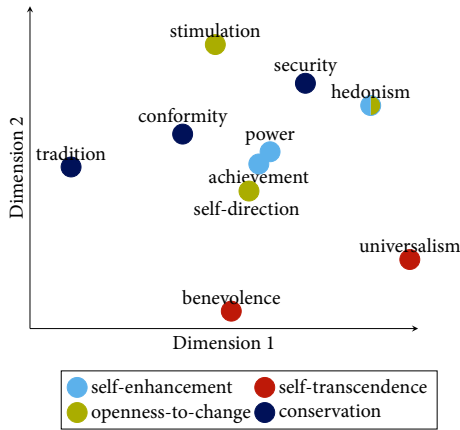


Figure 6.2: Visualization of the covariance between values in estimated profiles.

6.4.3 Value Conflicts and Disagreement

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

H_0 The mean value profile similarity score between user pairs that disagree is equal to the mean value profile similarity score between user pairs that agree.

H_a The mean value profile similarity score between user pairs that disagree is lower than the mean value profile similarity score between user pairs that agree.

We report the Bayes’ Factor (BF_{10})² to assess the relative increase in odds for assuming the alternative over the null hypothesis after observing data [23]. BF_{10} scores in $[3^{-1}, 3]$ are considered to indicate evidence for neither hypothesis, whereas more extreme values favor one hypothesis over the other, allowing us to make conclusions in either direction [195].

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

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

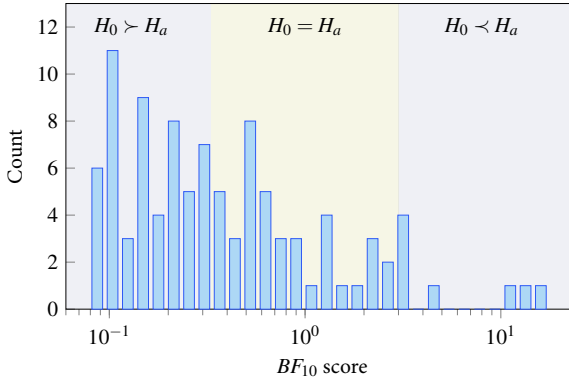


Figure 6.3: BF_{10} scores obtained for the combinations of data, value estimation methods, and scoring metrics.

Profiles from VPE

We split Disagreement based on *agree* and *disagree* labels (and drop all pairs with a neutral label), obtaining respectively G^+ and G^- . For each group, we compute the profile similarity scores using each method mentioned in Section 6.3.2. We do this per subreddit and observe the differences in score distributions. The alternative hypothesis is defined as the mean similarity scores in G^- being lower³ than the mean for G^+ :

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

$$H_0 : \theta_{G^-} = \theta_{G^+} \quad (6.6)$$

$$H_a : \theta_{G^-} < \theta_{G^+} \quad (6.7)$$

We report the BF_{10} for all combinations of similarity methods and parameters. We run 100 tests, considering 5 subreddits, 4 similarity scores, and 5 value profile thresholds $l = \{1, 10, 50, 200, 500\}$. Figure 6.3 provides an overview of the BF_{10} scores.

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

Thus, given specific settings and domains, we can trace disagreement between users to value conflicts. Table 6.4 shows the tests where $BF_{10} > 3$. In all cases, the filter l was 10 or more, stipulating that populated value profiles are required for measuring value conflicts reliably. We observe that BLM, the subcorpus with the highest profile diversity (Table 6.3), is frequent among these positive cases. Thus, having diverse profiles increases the likelihood of finding a link between values and disagreement. One positive test result is observed for the BREXIT subcorpus for a high profile threshold (500). Brexit includes the smallest number of

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

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

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

user profiles; the high profile threshold further removes several profiles. Thus, the positive result for BREXIT, based on a low number of profile comparisons, may not be reliable.

Mixing with Self-reported Profiles

Given that we use a novel method for estimating value profiles, we compare the results from the previous experiment with one that uses self-reported value profiles. Self-reported profiles mitigate the noise stemming from the value estimation step. The setup is identical to Section 6.4.3, but now we compute similarities between an estimated profile and a self-reported profile, obtained from a value survey.

We run a user study to obtain (1) self-reports of value profiles using an established value survey [PVQ-21, 343], and (2) agreement labels on posts in Disagreement. We obtained an IRB approval (exempt status) for our study.

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

The results are shown in Figure 6.4. We observe that deciding between the two hypotheses is not possible, in a majority of cases, as most evidence attributed both as equally likely. However, it is interesting to notice that using self-reported value profiles shifts the majority of results from favoring the null hypothesis to the undecidable range. In combination with the results from the previous section, this indicates that VPE methods need careful evaluation with respect to self-reported profiles as both may contain errors stemming from different sources and may have complementary merits. VPE suffers from errors made by the value estimation model but has the potential to use large amounts of data. In contrast, although self-reports yield a profile directly, they may be prone to biases.

Two tests still show evidence in favor of accepting H_a (see Table 6.5). They are on two task instances in the same domain, DEMOCRATS, and are measured for the τ and MD metrics. Here, our results differ from the previous experiment, and different subreddits result in high BF_{10} scores. In this case, one user’s value profile is constructed using self-reports, which are obtained without reference to discussions (i.e. not estimated from posts on Reddit). This may cause other factors to influence the diversity of profiles stemming from the PVQ. Furthermore, the task instances contained a call for action (e.g., *Please just vote [...]*

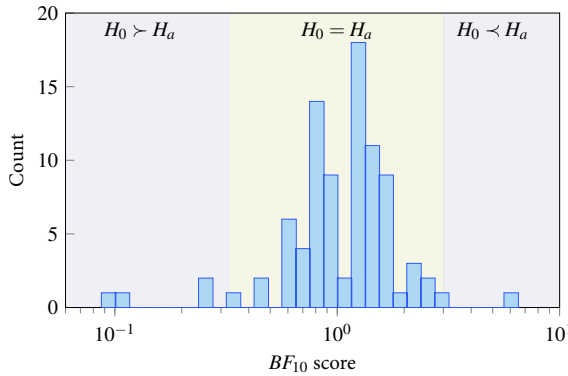


Figure 6.4: BF_{10} scores for all similarity scores and task instances comparing VPE and self-reported profiles.

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

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

and *The gloves should come off [...]*). The values embedded in the call to action may be one of the reasons why annotators felt inclined to disagree or agree.

Qualitative Assessment

To better understand when value conflicts influence disagreement, we perform a qualitative analysis of some instances (comment pairs) from the dataset that follow our hypothesis and some that do not (Figure E.6 in Appendix E.2 shows such examples).

We identify five trends in misaligned instances. (1) **Not enough information** in a value profile (i.e., low-frequency value mentions). This means that the user posted little value-laden content or that the value extraction method erroneously ignored some value-laden comments. (2) **No apparent value-based reasoning** involved in the comments, e.g., factual answers to a question. (3) **(Dis-)agreement happens on a content level** since profiles do not dictate individual utterances. This occurs when users disagree that a decision is “for good,” but fail to motivate their motivations for what is “good.” (4) The **target** of disagreement can be **partial**, whereas value conflicts are measured between two users. (5) In a few cases, the label given in Disagreement is **faulty** (e.g., annotators misinterpreting sarcasm or the text is vague).

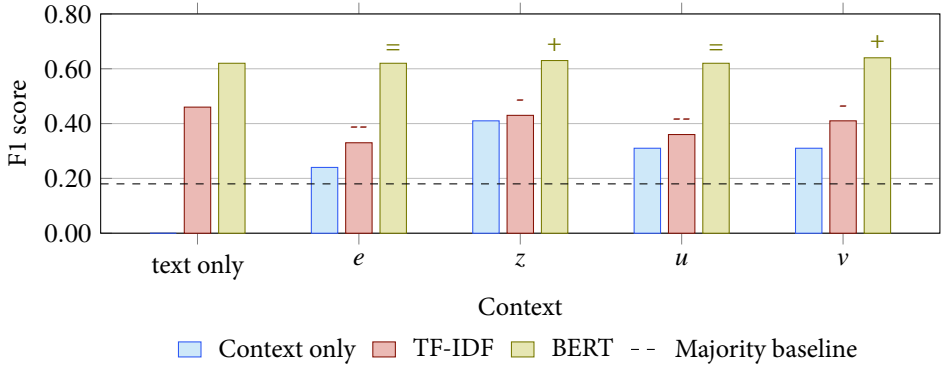


Figure 6.5: F_1 scores when adding extra context information. Symbols above bars show changes with respect to text-only: -- for $\Delta F_1 < -0.1$; - for $-0.1 < \Delta F_1 < 0$; = for $\Delta F_1 = 0$; and + for $\Delta F_1 > 0$.

6.4.4 Use Case: Predicting (Dis-)agreement

We assume that users' value profiles (in addition to the content of users' posts) play a role in predicting the agreement between users. We adopt the setup from Pougué-Biyong et al. [305], where an agreement label is predicted between parent p and child comments c . We add extra information to p and c using four methods.

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

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

Explicit user features (u) Nine features commonly extracted for representing users on Reddit (e.g., [74, 184]) to add extra contextual information.

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

We create embeddings (TF-IDF or BERT) for p and c and concatenate them to the user-specific context [151]. We standardize the user-specific context information to avoid raw values having a large impact, similar to the value profiles (v). When training with user profiles, we subsample Debagreement to include only those (p, c) pairs in which we have background data for both p and c . This leaves 65% of the data (28K samples). We train our classifier on an 80/10/10 split, retaining the most recent 20% as validation and test sets to reflect a real-world training scenario on historical data [361].

Figure 6.5 shows the results. Classifiers using TF-IDF embeddings fail to use the information effectively. BERT outperforms both our baselines, in line with the results for [305]. In this setting, none of the additional information causes major changes in performance, but we see an improvement using the value profiles and centroids. Compared to other work, using user-specific information is surprisingly difficult [6]. Further inspection for BERT indicates that the *neutral* class is hard to predict, as information from the value profiles may not be relevant. Mixing background information using, e.g., GNNs [255] may make more effective use of the profile information.

6.5 Conclusion

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

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

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

Constructing profiles from behavioral cues, such as written opinions, is noisy. For future work, we hope to see the creation of resources that allow end-to-end evaluation by combining text posts with a consistent set of users that allows aggregation to ground truth profiles or self-reported profiles. However, gathering such profile information outside controlled lab settings is highly complex. Future experiments may incorporate more judgments and provide stronger evidence for one hypothesis. These can be retrofitted with our results through Bayesian updating [268].

Limitations

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

First, the value extraction methods we employ (see Table 2) may have unknown errors. Our work is not focused on optimizing value extraction, which is an emerging research direction [202]. Adding more annotated Reddit data would allow us to judge the performance of value extraction models better. A future direction is to employ other training paradigms like Multi-task Learning [e.g., 122] or techniques for mixing in general-purpose language models [e.g., 399].

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

Third, the reporting of our results is limited to the Bayes Factor (BF). Further, most of our results fall inside the neutral category (“cannot decide between H_0 and H_a ”). We require more data to decide which of the hypotheses is more likely. An estimation of the posterior odds of the hypotheses e.g., in the form of *Highest Density Intervals* (HDI) might yield more insights, and would involve deciding on a *region of practical equivalence* (ROPE), as well as picking a thus far unknown prior distribution over the values for $\mathcal{S}$ in our two hypotheses [211]. However, BF and HDI interpretations can be seen as complementary, respectively

quantifying evidence or beliefs [410].

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

Ethical Considerations

First, the dataset used to model online discussions, Debagreement, was sourced from online interactions between users on Reddit. Research conducted on Reddit data is biased to a WEIRD (Western, Educated, Industrialized, Rich, Democratic) demographic, and results may not generalize to a broader set of users [308]. However, our method outlines which data is required for performing the same analysis given the availability of richer data, not necessarily stemming from Reddit. Second, models for predicting values may be wrong, they may lead to harmful outcomes for particular groups or populations [265]. In any application, the incorporation of control mechanisms (i.e., providing users a way to influence the construction of their own value profile) is a requirement for making sure the value profiling is conducted in a transparent and accountable manner. Broadly, this work should further be situated in a system containing checks and balances, making sure any output stemming from automated classification is verified by human agents before having an effect on actual users.

IV

Conclusions

7

Contributions and Future Work

Diversity is an important factor in achieving high-quality outcomes from deliberations. Current Natural Language Processing (NLP) approaches for supporting deliberations fail to facilitate diversity, especially in the range of perspectives involved. Hybrid Intelligence (HI)—a synergistic approach that augments human intelligence with AI techniques—offers effective analysis methods that align with deliberative ideals. Our HI approaches require nuanced insights from humans but exploit the processing capacity of NLP for mining diverse perspectives to facilitate online dialogue **at scale**. We experiment with extracting perspective hierarchies to derive deep insights into human opinions on contemporary topics. We explore how modeling arguments in discussions can lead to **bidirectional gains** by connecting underlying motivations and expressed agreement. Structuring the opinions in a discussion in terms of evidence-based and argumentative discourse encourages participants to articulate their perspectives more clearly, support their claims with relevant evidence, and engage critically with counterarguments. Fostering a culture of reasoned debate promotes a deeper understanding of complex issues while revealing the connection between deeply rooted personal values and the stance a person might adopt in a discussion.

Using NLP to analyze online discussions is a lively research area. The surge of LLM-based techniques has given a significant boost to understanding text-based human opinions across contexts. This dissertation critically examines these techniques by applying them to investigate how humans deliberate online. Our results reveal three error cases for existing LLM-based approaches to summarizing arguments: (1) generating and matching high-level arguments remains difficult for LLMs, (2) performance is dataset-dependent, and (3) low-frequency arguments are often missed in the summary. Thus, mining human subjective opinions with LLMs remains an open challenge, especially for sensitive and controversial topics. Further, aiming for a single ground truth in a discussion defeats the purpose of engaging with an opinionated audience. Being sensitive to the pluralistic nature of the opinions and values involved is a core capacity for making responsible NLP methods.

We also identify a strength of LLMs, that of **continued interaction**, as a spearhead for driving insights into the deliberative process at scale. Continuous interaction leads to iterative refinement, where users steer model responses and obtain desired outcomes. Interaction with models and other humans in a discussion requires active participation from the users. We offer a first step in this direction: leveraging large-scale feedback from individuals combined with input from language models. Nonetheless, measuring bidirectional gains remains challenging, as established benchmarks typically rely on large manual annotation studies.

Structure

The rest of this Chapter is structured as follows. In Section 7.1, we dive into the concrete findings related to the individual research questions. We summarize our cross-cutting findings and outline the contributions to the field in Section 7.2. We describe the limitations of our research in Section 7.3. Lastly, we provide an outlook for future work in Section 7.4.

7.1 Research Findings

We set out to investigate the practical issues of interpreting large-scale opinionated feedback from citizens with LLMs, and how to create hybrid methods to improve the diversity of opinion representations. We divided this goal into three research questions and examined each question separately. What follows is a description of our findings per question.

7.1.1 NLP for Perspective Analysis

Q1 *What are the fundamental issues in using NLP to analyze perspectives?*

Our work provides insights into the behavior of state-of-the-art NLP models for discussion analysis. LLMs are becoming a core tool for such purposes, and are capable of extracting high-level insights from large-scale text data. However, we empirically observe numerous error cases for LLMs, including poor out-of-domain generalization performance and an inability to saliently represent infrequent opinions. These error cases impose practical limits on how to design and use LLMs in sensitive situations, such as interpreting citizen feedback [287, 394]. In this section, we further break down our findings on the fundamental limitations of using NLP for discussion analysis along four main threads.

Heterogeneous models Deciding which model out of the rapidly developing number of models to use has become increasingly difficult [65]. Our findings show that for analyzing perspectives, no clear dominant NLP model or approach exists. Choosing between zero-shot LLMs and fine-tuned classification setups relies on context-specific knowledge and extensive experimentation. Our work shows that all parts of the NLP pipeline, including the scraping, preprocessing, and annotation of data, as well as model training and evaluation setup, greatly impact how a model behaves for downstream tasks. Creating an NLP tool for capturing diverse perspectives in online discussions requires considering every aspect of this pipeline carefully. Much like how humans utilize their diverse skills and perspectives to achieve optimal outcomes, we can rely on empirical approaches that recognize and leverage the complementary strengths of different methodologies. This paves the way for robust, inclusive, and nuanced NLP solutions, but puts considerable strain on the experimental setup used for assessing the efficacy of the developed approaches. This is exacerbated by stochastic behavior from LLMs, leading to uncertainty in the reproducibility of results [254].

Out-of-context generalization A key factor informing us on which model to use is the ability of an NLP model to learn from data in one domain and generalize to another [177]. Even complex tasks under severe data constraints, like argument quality prediction, can be modeled effectively. The diversity of the data used during training drives cross-domain performance. Most opinionated data stems from online platforms, which is hampered by their fundamental deficits outlined in Chapter 1, in particular in the limited inclusion of under-represented voices. Therefore, we stipulate that improved representation of diverse users, as is the goal in this dissertation, can ultimately lead to greater model performance: through improved representation, we promote participation from those users previously alienated from discussions, which in turn drives the data diversity of the training data for our models.

Level of abstraction This dissertation encompasses different tasks for extracting information from an individual's opinion. Generally, tasks that capture low-level linguistic phenomena are easier to model, making approaches to such tasks easier to interpret and evaluate by humans. However, low-level constructs only reveal crude information about a person's opinion. Hence, we shift focus to tasks of a higher level of abstraction. Our work shows that NLP models are capable of performing highly abstract tasks like argument extraction

but do so with considerable error. For instance, models can miss up to 50% of arguments with low frequency. Extracting perspectives requires reasoning over implicit and common-sense knowledge [71]. While LLMs seem to fluently deal with abstract tasks by interpolating missing information, even the largest models struggle with theory of mind tasks [406], a key capacity for performing perspective-taking [306]. We use this observation as a guide to develop our hybrid approaches by incorporating tasks of various levels of abstraction into the Perspective Hierarchy. The higher the level of abstraction, the more difficult the task, and the more human oversight we require. This strategy not only accounts for failures of the model but also allows us to deal with implicit signals that are involved in reasoning over high-level abstract information. By involving humans in the loop, the abstract information can be made more explicit, which we can, in turn, leverage as additional training data.

Human disagreement alignment Our experiments show that the errors made by LLMs are often unlike those of humans and that NLP models do not always align with human uncertainty. This misalignment means that models require calibration before they can accurately reflect human judgment. Therefore, reasoning over LLM capabilities according to human standards is unwarranted. This problem is further compounded by the rampant anthropomorphization of AI models in modern applications, which can lead to unrealistic expectations of their abilities [1]. We see that current benchmarks for evaluating the performance of LLMs often fail to capture the diversity of perspectives, instead prioritizing the majority opinion. This is problematic as it can lead to the marginalization of minority voices and the perpetuation of biased viewpoints. To address this, we adopt a perspectivist approach, learning from a distribution of subjective interpretations instead of aiming for a single ground truth [61]. LLMs are highly sensitive to context and will vary depending on the prompt. By exploring the variance of LLM responses, we can start to uncover some of the disagreements in how opinions may be perceived between humans, too. This brings about the integration of machine and human judgments to create systems that can accurately and fairly represent the full range of perspectives present in a given discussion.

7.1.2 Hybrid Intelligence for NLP

Q2 *How to combine human intelligence and NLP to effectively capture diverse perspectives?*

Our experiments focus on improving the representation of infrequent voices in online discussions when using NLP to extract perspectives. We do so by incorporating humans in the loop and designing pipelines that lead to an increase in the diversity of arguments. In this section, we highlight our three main contributions to designing HI using NLP.

Sample efficiency A significant challenge for discussion analysis is the human inability to manually examine the entirety of the data in-depth, due to time and cognitive constraints. While automated tools are an attractive solution that can process entire datasets quickly, we show that tools often fail to extract all perspectives from the data, particularly in the case of nuanced human opinions. To address these limitations, we develop a novel approach that leverages the sample efficiency of human understanding. Humans have the unique ability to extract a wealth of information from a single stated opinion and require fewer examples

than modern NLP models to derive meaningful insights into diverse perspectives. However, this human involvement may introduce biases, as they may fill in implicit information, project their personal views when interpreting others, and have biased background knowledge. Therefore, the task of selecting which data samples should be examined, and who to choose for annotation, becomes a critical one. We show the adoption of active sampling strategies can dynamically assign diverse opinions to humans. The analysis of large-scale data necessitates a balance between the speed of automation and the depth of human insight, which is answered by our integration of sampling diverse opinions (HyEnA) for diverse annotators (ACAL).

Advancing benchmark-based evaluation The integration of human and artificial intelligence in hybrid approaches presents a new challenge for evaluation. Traditional methods of measuring the performance of NLP systems obtain gold labels manually. For hybrid systems, this is insufficient, as hybrid systems can provide important insights that may be missed in manual analyses, such as in Chapter 4. Instead, a three-way setup, where a hybrid approach is pitched against manual and automated ones directly provides a fairer comparison. Further, common high-level performance statistics, such as a single F_1 score per benchmark, do not provide information about how the model behaves for particular samples and annotators. This information is essential for designing context-specific hybrid approaches, creating user-specific instructions in using LLMs, and setting realistic expectations [189]. Instead, fine-grained evaluation metrics such as those focusing on individual annotators, are crucial for understanding how various approaches deal with diversity for different types of annotators. These findings show the importance of considering fair evaluation setups and the characteristics of the data when creating context-specific applications. Using humans in a hybrid approach may offer additional benefits beyond the primary task. For instance, the annotation procedure can foster understanding and empathy among annotators, as they report an increase in sympathy and recognition of the issues raised in the comments they annotate. Capturing this in a multi-objective evaluation setup presents an interesting avenue for future research, where the goal is to create synergy between the different parts of the hybrid approach, such that the cumulative gain outweighs the sum of its parts.

Measuring diversity A core goal of the approaches developed in this dissertation is to improve the representation of diverse perspectives. To measure diversity, we often assume that a fixed pool of opinions is at our disposal for analysis. Within this pool, diversity can be well-defined and measurable, for instance, by counting all unique items in a collection of arguments. The use of HI systems, such as those developed in this dissertation, can be particularly effective in this context, as they have been shown to achieve higher coverage and precision than state-of-the-art automated methods when compared to a common set of diverse opinions. Similarly, annotator-centric evaluation provides valuable insights into how different methods deal with disagreement and diversity on an individual level. For instance, large gaps between average, individual, and worst-off evaluations hint toward tradeoffs between representing the majority versus focusing on the minority. However, fixed pools of opinions obtained from online social media platforms already contain skewed opinion distributions. This underscores how data and annotation characteristics are key factors in measuring diversity, even when benchmark data is available.

7.1.3 Perspective Hierarchy

Q3 *How to construct a perspective hierarchy based on diverse opinions in a discussion?*

Our Perspective Hierarchy model illustrates how different levels of abstraction interplay when interpreting diverse opinions with Hybrid Intelligence. We discuss our experiments, showing that argumentation forms a core ingredient of the hierarchy, and highlight that obtaining perspectives from text should be done using hybrid intelligence approaches.

Importance of argumentation Our analysis reveals a nuanced relationship between value profile similarity and disagreement in online discussions. While a general lack of correlation is observed, specific cases emerge where value dissimilarity aligns with disagreement. The lack of a general correlation points towards the importance of incorporating arguments in our perspective hierarchy and the relevance of creating HI approaches to capturing arguments. This uncovers how values drive opinions. The cases that revealed a strong correlation were those where values matter most and were diverse. This suggests that value conflicts, though not directly correlated, signal underlying motivational diversities that contribute to disagreements. Such signals can be leveraged to find opinions that differ from the majority.

Hybrid hierarchies Constructing value profiles based on automated judgments over texts is noisy. Involving a human in the loop helps infer values relevant in a context [240]. In our experiments, we estimate value profiles by analyzing text-based opinions and through self-reporting. Our findings show that these two approaches differ considerably, indicating that a mix of methods is required to represent individuals' perspectives. Hybrid approaches support such combinations of methods. Through interaction, misrepresentations can be corrected [332]. How individuals correct models may also drive further insights into the difference between behavior-based opinion analysis and self-reported preferences.

7.2 Contributions

Each Part of this dissertation provides an answer to an individual research question. In this section, we combine our findings to provide answers to the question of how humans and NLP can improve their understanding of diverse perspectives in online discussions.

7.2.1 Scientific Relevance

Deliberation process In most of our experiments, we lack access to the original participants of a discussion to further probe their perspective, since we primarily rely on historical user-generated data. This makes it impossible to verify the original intent with the author. Traditional NLP often relies on ad-hoc annotation procedures that combine interpretations from a crowd of annotators for creating training and evaluation data. During the execution of our hybrid approaches, we also employ crowds of annotators but invite them to provide more productive information. We actively account for the annotator's point of view when requesting additional labels. This provides insights for the formation and diversity of opinions in subjective tasks beyond investigating demographic characteristics post-hoc. Furthermore, by making annotators observe diversified opinions we encourage the exploration of

novel ideas from a multitude of viewpoints. We find that this approach is beneficial to the faithful representations of opinions, and improves the facilitation of a constructive and inclusive discussion. Hence, we conclude that hybrid approaches can play a crucial role in facilitating deliberative discussions by promoting active perspective-taking.

Interactive AI for HI Hybrid methods are effective because they *iterate*. It is crucial to engage in an interactive and continuous process of correction, particularly when seeking to acquire opinions from a diverse range of individuals. The conventional approach in the fields of NLP often involves single isolated interactions, such as a human providing a set of labels at a specific point in time, or a model providing a single prediction. However, it is important to consider NLP methods within the respective contexts they are applied. Designers and developers constantly refine their algorithms to enhance performance, while improving the evaluation procedures to obtain a more accurate assessment of the model capabilities. Similarly, instructing humans is not a one-time event, but rather a continuous process of receiving and integrating multimodal feedback from the environment. The interaction between AI and developers, or AI and users, is complex and rich, and by turning to HI we can guide this interaction in mutually beneficial ways. Our work demonstrates this by leveraging LLMs to sample from large pools of data but letting humans read them, thereby uncovering unique perspectives from a large and imbalanced set of opinions.

Fundamental limits for representing minorities We find that LLMs are suited for representing majority opinions since these constitute frequent and salient signals in training data. Further, LLMs can be steered in their alignment, rendering objectivity problematic. The sensitivity of LLMs to prompts and the lack of a faithful representation of the dynamic context of real-world applications leads to irreproducible research. Carefully crafting experimental designs and training procedures can mitigate this behavior, but LLMs remain brittle when confronted with novel infrequent opinions. HI addresses this shortcoming by exploiting the complementary strengths of humans and LLMs in interpreting opinions.

Explicit communication and deliberation The elicitation of explicit communication is a critical aspect of the development of HI for analyzing online discussions. NLP methods can benefit from explicit communication from humans since it leads to additional training data or labels. Humans can also benefit from a more rationale-based discussion since engagement in a discussion develops the understanding among participants. Coaching the argumentative motivation of opinions is an effective facilitation move that encourages individuals to articulate the reasoning behind their beliefs and opinions. This approach can be particularly effective in situations where there is likely to be consensus on a particular issue, but where disagreement arises due to conflicts in values. For instance, in a discussion about vaccination, there is often agreement on the need to protect children from harm, but disagreement arises due to differing beliefs about the safety and efficacy of vaccines, and the trustworthiness of the scientific and medical establishments. Explicit communication can acknowledge the common ground, and progress a discussion by shifting focus to the underlying beliefs. Furthermore, while both NLP methods and human annotators can deal with implicit information, they do so differently. NLP methods are likely to insert majority opinions based on their training data, while humans are likely to contribute their personal opinions. This

underscores the importance of a hybrid approach that combines the strengths of both NLP methods and human annotators: promoting more explicit, rationale-based communication ensures that a diversity of perspectives is represented.

HI benchmarks The development of new benchmarks is a critical aspect of the evaluation of HI systems (HIS). However, experimenting with and evaluating LLM predictions can be resource-intensive. Obtaining labels from crowd workers requires annotation guidelines, annotation platforms, and monetary compensation. Even after spending such resources, there is a significant strain on the reproducibility of experiments. All this makes it attractive to reuse existing datasets. However, benchmarking hybrid approaches requires careful consideration of the task context. Measuring additional behavioral signals that objectively capture the interaction in the HIS, or breaking apart overall performance into the contributions of its components through ablations can, either quantitatively or qualitatively, reveal why methods are effective. This dissertation proposes a mechanism for benchmarking HI using an iterative approach. We break apart tasks into elementary phases, which we can evaluate both intrinsically and extrinsically. We capture performance on an overall task (e.g., Argument Extraction), but also evaluate smaller steps in the procedure (e.g., Pairwise Argument Similarity Scoring). Such a breakdown allows for the flexible reuse of data across tasks to investigate their interaction.

7.2.2 Societal Relevance

Our findings show that it is possible to address the fundamental limitation of capturing diversity with NLP approaches using Hybrid Intelligence. In this section, we highlight how our approach to opinion analysis might achieve broader societal impact.

7

Citizen feedback data Our work is focused on interpreting textual comments in the form of citizen feedback for deriving insights into their opinions. In particular, we do so on contemporary topics, such as COVID-19 regulation [236, 274, 403]. Our work can be extended to feedback on other issues, such as transportation [275] or environmental issues [276]. Next to interpreting direct citizen feedback, numerous existing online platforms are already packaged as datasets, such as the Wikipedia Discussion Pages [125], UN debate corpus [340], and Kialo [359]. Mining the insights from them by, e.g., extracting the key arguments can help in furthering the discussion. Cross-topic application of the hybrid analysis procedures can lead to higher-level insights into opinion formation. For instance, in helping to distinguish what aspects of facilitating a diversity of perspectives are related to the discussion contexts, and what aspects transcend a particular topic.

Enhancing participation Incorporating diversity is a driving factor of the quality of discussions online, but also a requirement for legitimate policy-making. By making the analysis hybrid, we actively involve humans in the process, enhancing participation. For instance, requesting citizens to participate in analysis procedures such as HyEnA offers them the opportunity to contribute to the analysis while developing their personal views on the subject. After, the annotators can be approached for inclusion in future deliberation, as they have had the opportunity to familiarize themselves with the most important arguments in the matter.

This would progress the deliberation where ideas are based on each other's arguments. Targeted recruitment campaigns can help in finding a representative demographic, taking care to create inclusive samples of the population.

Other application areas The analysis of opinionated text has a wide range of potential applications beyond the interpretation of citizen feedback for policy-making purposes. For instance, a broader thematic analysis for qualitative data with HyEnA could be useful for deriving insights for product feedback [188] or education [90]. Uncovering the main concerns using argument extraction, and distinguishing them from deeper value-driven criteria is useful for all organizations looking to improve their products or services.

7.3 Limitations

Since we conduct empirical research, it is important to underline the limitations involved in the experiments, data, and analysis. In addition to the limitations mentioned in each Chapter, this section highlights cross-cutting aspects that influence the generalizability and conclusions derived in the previous sections. Addressing these limitations paves the way for future research that could contribute to a more nuanced understanding of our findings.

Perspective Hierarchy In the construction of the perspective hierarchy, we emphasized the reasoning behind the stances that individuals adopt, both at the communicative (arguments) and motivational (values) levels. The extracted hierarchy representations are specific to a particular human-generated opinion or proposed action, making it challenging to compare hierarchies across different claims or contexts. There are alternative approaches to modeling the target of a perspective. For instance, others extract perspectives for high-level claims [71], short free-form viewpoints [104], or events [412]. These alternatives can be ways to compare perspective representations across different discussions. Beyond the levels included in our hierarchy, other expressions or behavioral signals can be captured from text-based opinion data. Examples include sentiment [244], and emotion [2]. Incorporating these additional dimensions of human expression can provide valuable insights into an individual's feelings in a discussion. However, a high degree of analysis of these feelings may lead to a focus on affect over content or chilling effects, as individuals may feel monitored [59]. Furthermore, the introduction of additional levels increases the likelihood of generating incorrect predictions. Extracting further content-specific information may be beneficial for providing high-level overviews of the content in a discussion, such as resolving attribution of who holds what opinion, or the entities related to the topic of discussion.

Experimental constraints Empirical research is inevitably constrained by experimental conditions and design limitations. For instance, the participant sample that provides opinions in some of our experiments and the annotators we employ in them often reflect a WEIRD (Western, Educated, Industrialized, Rich, and Democratic) demographic. The concept of an ideal participant group is complex and multifaceted, but it is crucial to consider the potential biases that may arise from it, especially in the context of facilitating diverse perspectives. When humans provide their opinions in a discussion, we rely heavily on self-reporting, with the underlying assumption that participants are reporting their interpretations faithfully. We also assume that the discussion is largely free of malicious behavior,

such as trolling or other types of misconduct. These assumptions do not always hold in real-world use cases, and the potential for intentional misinformation or disruption in the discussion must be acknowledged. Further, as is standard practice NLP research, we often rely on third-party annotators to interpret opinions that they did not originally author. This approach missed the information the original author could provide, e.g., the context and intent of their message, which the third-party raters cannot provide. Improvements in LLMs or prompting techniques can directly benefit our work, but the need to rerun experiments can be costly. The choice of the LLM model can significantly impact the performance of a hybrid approach, but finding the best model for the task at hand requires extensive experimentation, incurring both human and computational costs. This feeds into the broader benchmarking problem, where the lack of standardized evaluation metrics can make it difficult to compare and contrast different versions of the same approach. A similar reasoning holds for the use of data in our experiments, which limits our ability to investigate how changes to dataset characteristics, e.g., increasing the number of annotations per sample, impacts our results.

Repeated interaction We emphasized the benefits of repeated interaction between NLP models and humans in the creation of high-level overviews of opinions and developed hybrid approaches that construct high-level overviews of opinions, such as summarizing arguments into key points. The main focus in these approaches has been collaboration between people and NLP models to iteratively refine the overview. However, our current efforts have not focused on continued deep interaction with a single human, which could be taken as an alternative design to HI. While we have not yet conducted experiments with continued interactions, we acknowledge that this approach could lead to complementary outcomes for the hybrid analysis of online discussions. For example, iteratively refining the perspective hierarchy through a conversation between LLM and a human could facilitate perspective-taking and improve the accuracy of the analysis. To demonstrate such improvements orthogonal experimentation is necessary. Some work has already begun in this direction, with research indicating that deliberation among annotators can be beneficial for reaching consensus on labels, although it depends on the characteristics of the discussion [338].

7.4 Future Work

In this final section, we present our vision for the future of research at the intersection of HI, NLP, and online deliberation. Through these suggestions, we hope to advance the state of the art in HI, NLP, and online deliberation, and to inspire contribution to the development of more inclusive, productive, and democratic online discussions. We outline four avenues.

Design of Hybrid Intelligence Integrating human and artificial work requires careful task balancing. In developing our hybrid approaches, we have cast this in a fixed process. However, dynamic task allocation and balancing are core capacities of effective teams. Knowing when and whom to ask, such as obtaining an automated judgment from an LLM or querying a pool of diverse human annotators enables successful collaboration [199]. Frameworks like learning to defer [259] or other active learning approaches [40] can be used to facilitate this. These examples touch on the integration of humans and AI, but a broader understanding of how to design HISs is lacking. There are general guidelines [413], but how to develop HIS for the field of NLP remains unclear. In our work, we identified that specific designs can

reshape human–AI interactions significantly. For instance, swapping the order in which humans and LLMs collaborate in HyEnA may decrease the precision but increase efficiency.

Evaluating HI Evaluation of HIs requires novel benchmarking paradigms. Existing benchmarks are usually annotated manually and composed out of many individual existing datasets, and therefore lack a faithful representation of the dynamic context of real-world applications [69]. Alternative approaches can instead incorporate interactive crowd-sourced benchmarks that develop over time [200], or turn to use-case-specific evaluation, leveraging objective behavioral cues to assess our methods. To target the desired capacities of language models, we identify them based on context and judge whether LLMs fulfill our requirements. This leads to the creation of a sort of “unit-test” for our use cases [369]. Versions of this context-specific evaluation for facilitating online discussions can directly target diverse opinions [425], or measure interaction structure to reveal the quality of a conversation [331].

Contextualizing HI for online deliberation We suggest several approaches for bringing HI to online deliberation. First, we suggest that the analysis of online deliberations results from a mix of self-reporting, machine interpretations of opinions, and crowd-sourced labels. This can result in a thorough understanding of the differences in interpretation between the intention of an author, and how it is perceived in an analysis. Second, we looked into how people conduct discussions but refrain from committing to a particular topic of discussion. However, context impacts the strategy for facilitation. Future work can start by taking a real-world use case, and design interventions based on the hybrid approaches developed in this work. The true impact of HI may only be known after engaging in long-term interaction between humans and AI. Lastly, our hybrid approach represents the citizens’ preferences from a societal discussion in one iteration. Nonetheless, societal problems are not solved with a single decision, and citizen consultation processes take place continually. In the long run, perspective hierarchies can support negotiations [317] among societal stakeholders, e.g., on which portfolio of choices to make to combat a pandemic [274].

Opinion shift We have adopted a hybrid approach to modeling perspectives, which involves the extraction of stances, arguments, and values based on human-provided opinions. First, it is important to consider that opinions are not formed in a vacuum, but are rather shaped by a myriad of factors, including the political, social, and personal context of the opinion holder. Consequently, the temporal aspect of when an opinion is expressed is an important aspect that enriches the understanding of a perspective [152]. However, extracting and placing events based on text-based opinion expressions is complex [310]. Hybrid approaches facilitate the engagement and interaction between participants, causing opinions to shift. Insights into how opinions change over time, for instance in the frequency of certain topics or arguments can subsequently serve as an indicator of changing consensus. Finally, the relevance of an analysis is often confined to a specific time frame, as opinions and perspectives change in response to world events. Therefore, to accurately contextualize and interpret perspectives for deriving insights into public opinion, it is essential to consider the state of the world at the time opinions were expressed.

V

Appendices



An Empirical Analysis of Diversity in Argument Summarization

A.1 Detailed Experimental Setup

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

A.1.1 Data

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

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

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

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

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

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

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

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

A.1.2 Per-approach Specifics

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

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

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

Approach name	Model
ChatGPT	gpt-3.5-turbo-16k
<i>ChatGPT (closed book)</i>	gpt-3.5-turbo
Debater	<i>closed-source</i>
SMatchToPR (base)	RoBERTa-base
SMatchToPR (large)	RoBERTa-large

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

Subtask	Parameter	Value
KPG	mapping_policy	<i>LOOSE</i>
	kp_granularity	<i>FINE</i>
	kp_relative_aq_threshold	0.5
	kp_min_len	0
	kp_max_len	100
	kp_min_kp_quality	0.5
KPM	min_matches_per_kp	0
	mapping_policy	<i>LOOSE</i>

Table A.5: API Configuration for Debater approach.

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

Prompt 1: ChatGPT closed book, KPG prompt

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

Prompt 2: ChatGPT open book, KPG prompt

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

Prompt 3: ChatGPT open book, KPM prompt

For the claim of {**claim**}, indicate for each of the following argument/key point pairs whether the argument matches the key point. Return a JSON object with just a "match" boolean per argument/key point pair.

ID: {**pair id**} Argument: {**argument**} Key point: {**key point**} (*up to B_{KPM} times*) ...

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

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

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

A.1.3 Evaluation metrics

For Key Point Generation, we resort to measuring lexical overlap and semantic similarity. To make our results reproducible we provide further details on the configuration of the ROUGE metrics [150]. Our evaluation uses the sacrerouge package that wraps the original ROUGE implementation¹. The full evaluation parameters can be seen in Table A.7.

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

¹<https://github.com/danieldeutsch/sacrerouge>

Parameter	Value
Porter Stemmer	<i>yes</i>
Confidence Interval	95
Bootstrap samples	1000
α	0.5
Counting unit	<i>sentence</i>

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

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

A.2 Additional results

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

A.2.1 Detailed ROUGE scores for Key Point Generation

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

A.2.2 Additional BERTScores for Key Point Generation

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

A.2.3 Long-tail experiment for KPG

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

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

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

A.2.4 ChatGPT generated key points for PVE

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

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

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

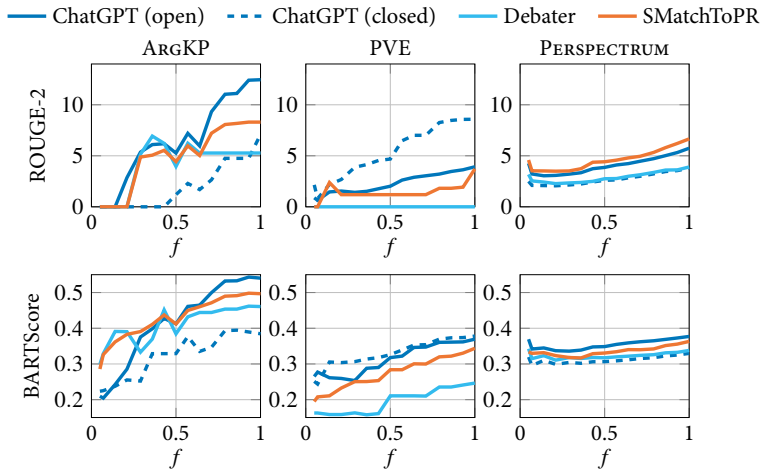


Figure A.1: KPG performance when limiting data usage to a fraction f , starting with the long tail first.

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

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

B

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

B.1 Hyperparameters

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

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

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

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

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

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

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

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

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

B.2 Additional results

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

B.2.1 Per-label Performance

See Tables B.2 and B.3.

B.2.2 Label confusion

See Tables 3.4 and B.4.

B.2.3 Seed Variance

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

		Predicted	
		-	+
True	-	131	75
	+	48	266

(a) GPT-3

		Predicted	
		-	+
True	-	75	131
	+	18	296

(b) MTL

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

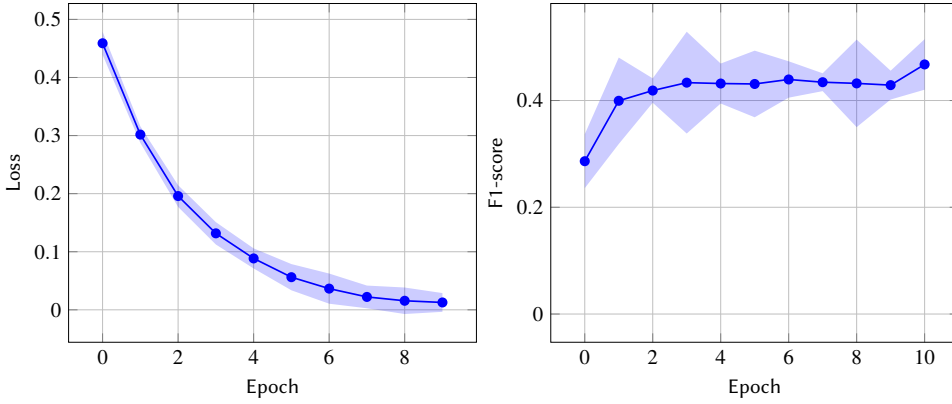


Figure B.1: Training loss and combined F1 score for multiple training runs of *CLTeamL-2* with different seeds.

B

B.2.4 Topics

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

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

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

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

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

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

C

C

A Hybrid Intelligence Method for Argument Mining

C.1 Experiment Protocol & Description

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

C.1.1 Preliminaries

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

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

C.1.2 Phase 1: Argument Annotation

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

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

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

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

C

C.1.3 Phase 2: Argument Consolidation

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

C.1.4 Comparison to Automated Baseline

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

C.1.5 Annotation platform

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

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

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

C.2 Method Details

C.2.1 Parallel Pairwise Annotation Algorithm

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

C

Algorithm 3: Parallel Pairwise annotation

Input: Dependency graph $G = \{V, E\}$

Output: Labeled vertices V

$B = \text{create bipartite graph}(G)$

$Y = \text{find maximal matching}(B)$

$P = \text{find disjoint paths}(Y)$

while *!fully_labeled*(G) **do**

for $p \in P$ **do**

$v = \text{find middle}(p)$

 label vertex(v) ;

▷ N humans

end

 automatically label paths(P , label)

end

C.2.2 Hyperparameters

HyEnA

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

ArgKP

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

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

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

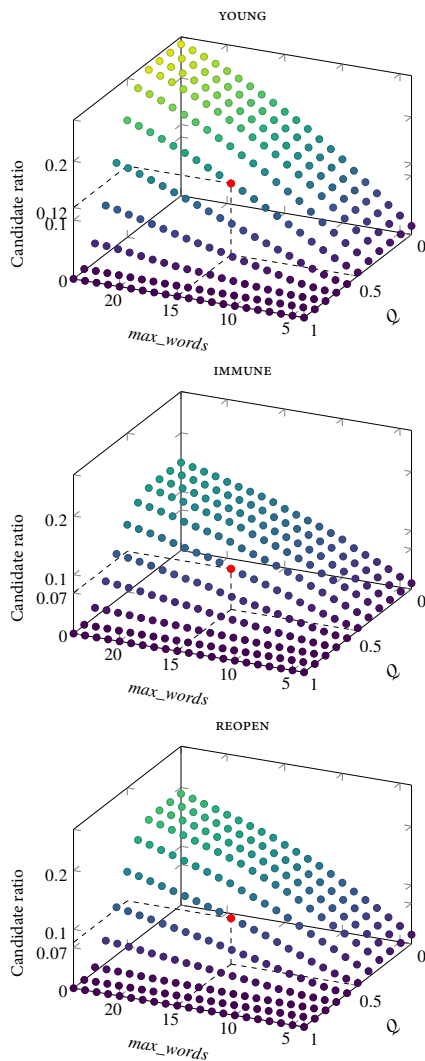
Table C.1: Quantative descriptive information for running ArgKP.

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

Table C.2: Hyperparameters used by HyEnA.

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

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



C

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

C.3 Detailed Results

C.3.1 Unclear Translation Actions

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

C.3.2 Clustering Arguments

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

C.3.3 Key Arguments

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

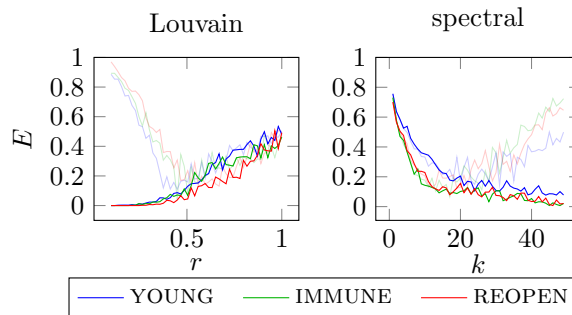


Figure C.2: Parameter tuning for argument clustering with $E = 0$ for argument clusters of size 1. Results are overlaid on Figure 4.6.

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

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

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

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

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

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

Option	ID	Stance	Argument cluster
REOPEN	0	pro	⟨ This will bring improvement in employment rate, This will improve the economy, This will help these industries recover, to support these sectors and to entertain and please us all, Killing the industry, This helps the economy ⟩
	1	con	⟨ will end up in another confinement, 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 attendees in the shows, will be controlled environment, With the necessary restrictive measures, cultural events must be able to be visited again as they are an important part of human life, Workers are well protected ⟩
	9	pro	⟨ No evidence that the lockdown works, A distinction should be made, some contact professions are basic service and others are not, Restriction of liberty is a violation of human rights ⟩
	10	pro	⟨ Excited to do things as before for preserving mental health, This will ensure freedom for the people, In order to save people's lives, we should be very careful and not relax too quickly, To support the churches and meet fellow believers again and pray and sing together ⟩
	11	con	⟨ It's not worth getting people sick, It's not safe yet , These are not vital industries ⟩
	12	pro	⟨ People need to let out pressure , People are tired and bored , Culture and entertainment is important in life, This will make people feel better ⟩
	13	pro	⟨ It will help everyone tremendously, This will help people go back to work, This will motivate people to be more active and healthy ⟩
	14	pro	⟨ Need freedom, It is best to know more of the virus before reopening these industries, This can be done following certain conditions, This will support small businesses recover ⟩
	15	pro	⟨ This will empower the people to be more responsible ⟩
	16	pro	⟨ Cannot be maintained, These places can't be maintained ⟩
	17	pro	⟨ It is easy to maintain social distancing in these industries. ⟩

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

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

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

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

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

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

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

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

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

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

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

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

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

D

Annotator-Centric Active Learning for Subjective NLP Tasks

D

D.1 Detailed Experimental Setup

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

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

D.1.1 Dataset details

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

D.1.2 Hyperparameters

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

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

Table D.2: Hyperparameters for the passive learning.

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

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

D

D.1.3 Training details

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

D.1.4 ACAL annotator strategy details

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

$\mathcal{T}_S$ uses a sentence embedding model to represent the content that an annotator has annotated. We use all-MiniLM-L6-v2¹. We select annotators that have not annotated yet (empty history) before picking from those with a history to prioritize filling the annotation history for each annotator.

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

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

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

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

D.1.5 Disagreement rates

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

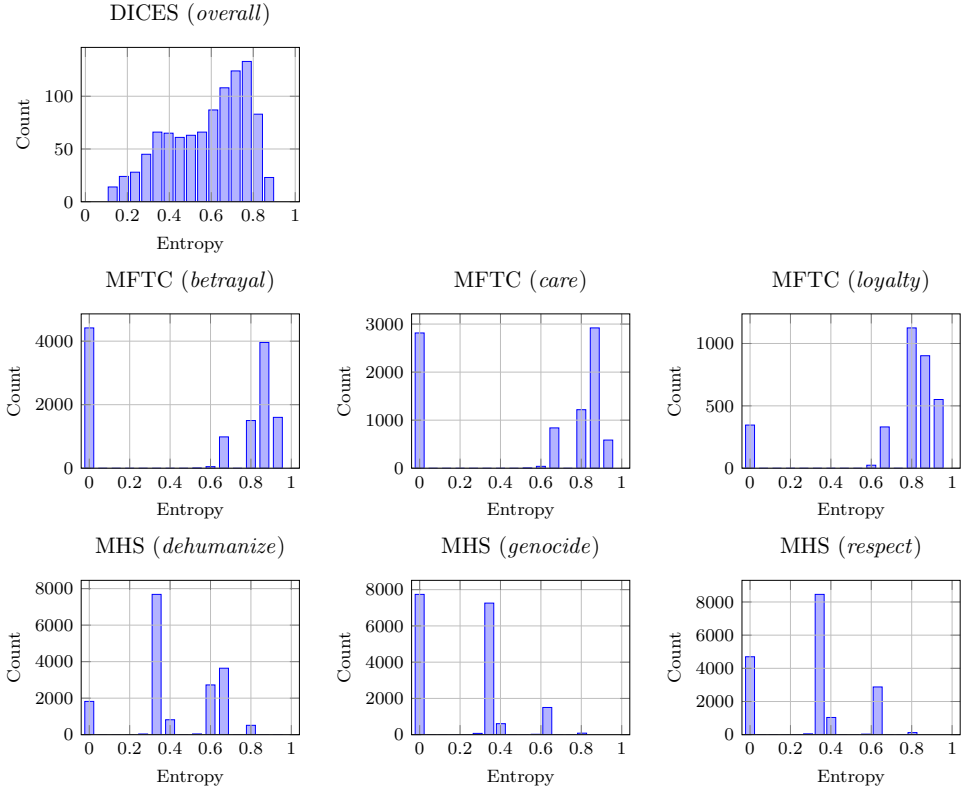


Figure D.1: Histogram of entropy score over all annotations per sample for each dataset and task combination.

D.2 Detailed results overview

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

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

D.2.2 Training process

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

D

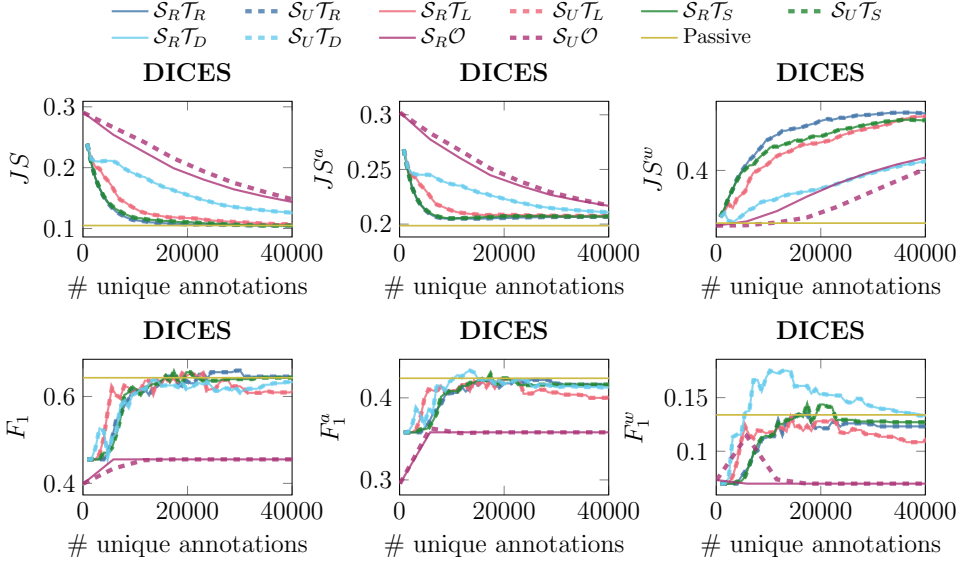


Figure D.2: Validation set performance across all metrics for DICES during training.

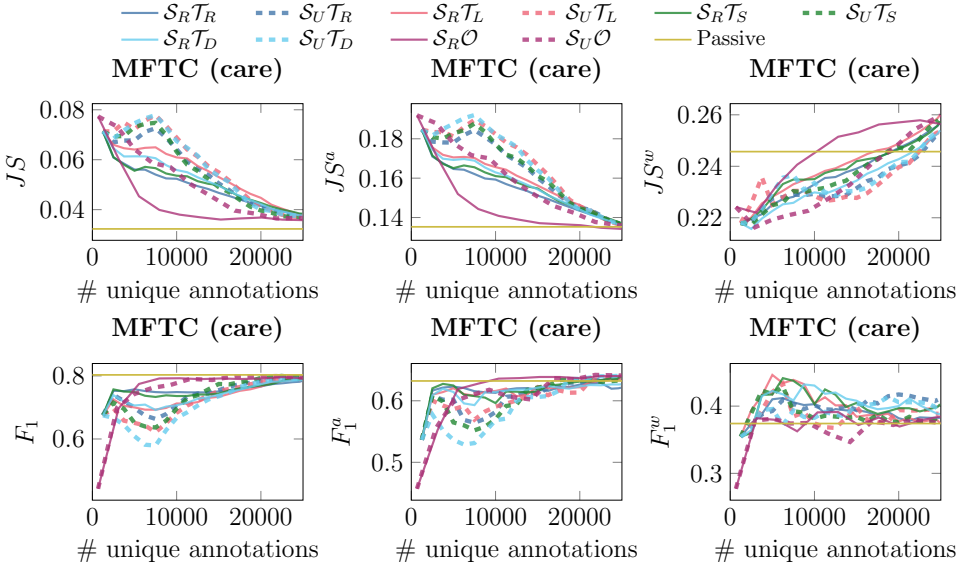


Figure D.3: Validation set performance across all metrics for MFTC (care) during training

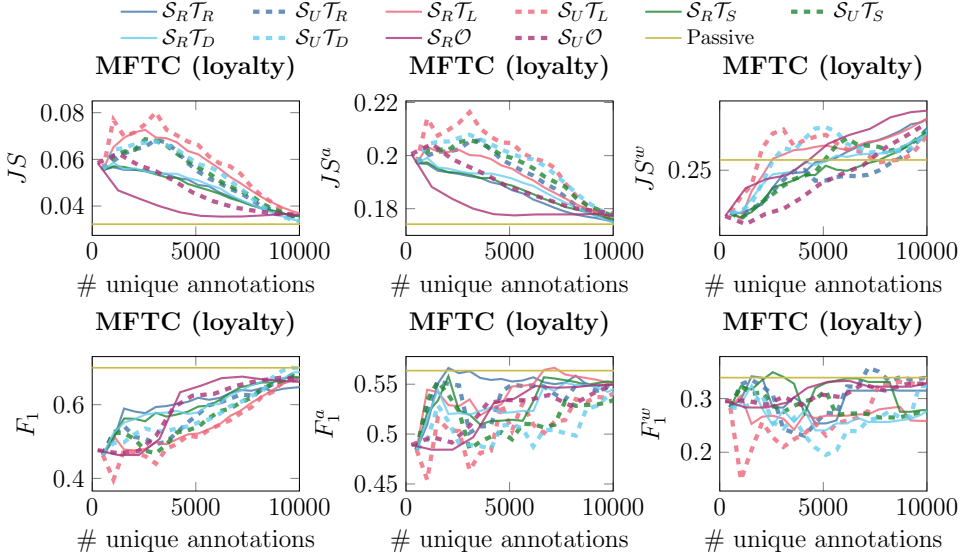


Figure D.4: Validation set performance across all metrics for MFTC (loyalty) during training

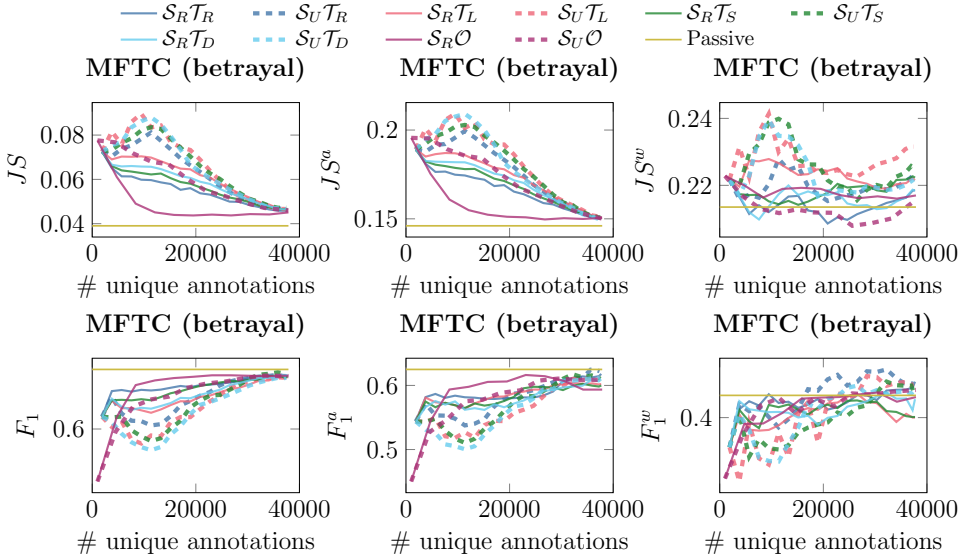


Figure D.5: Validation set performance across all metrics for MFTC (betrayal) during training

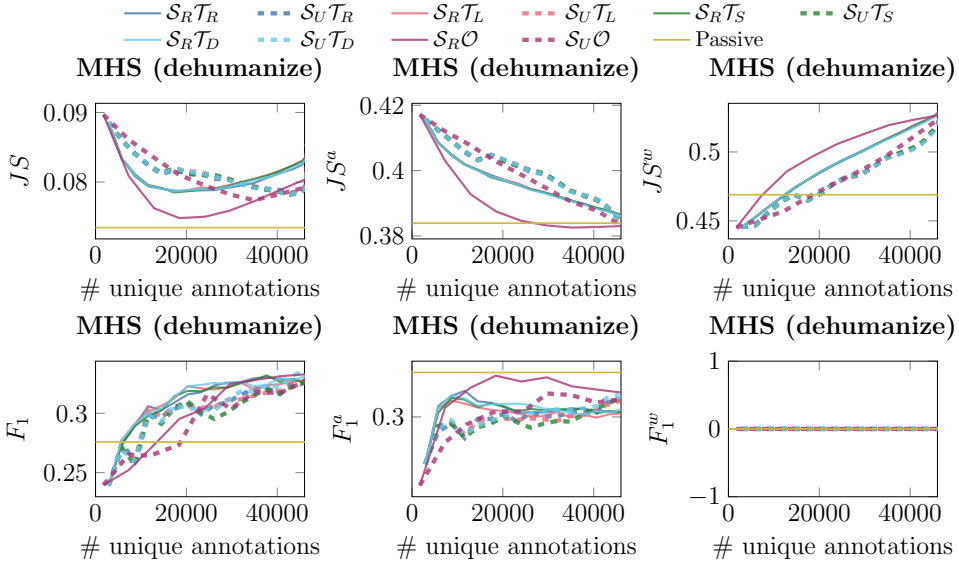


Figure D.6: Validation set performance across all metrics for MHS (dehumanize) during training

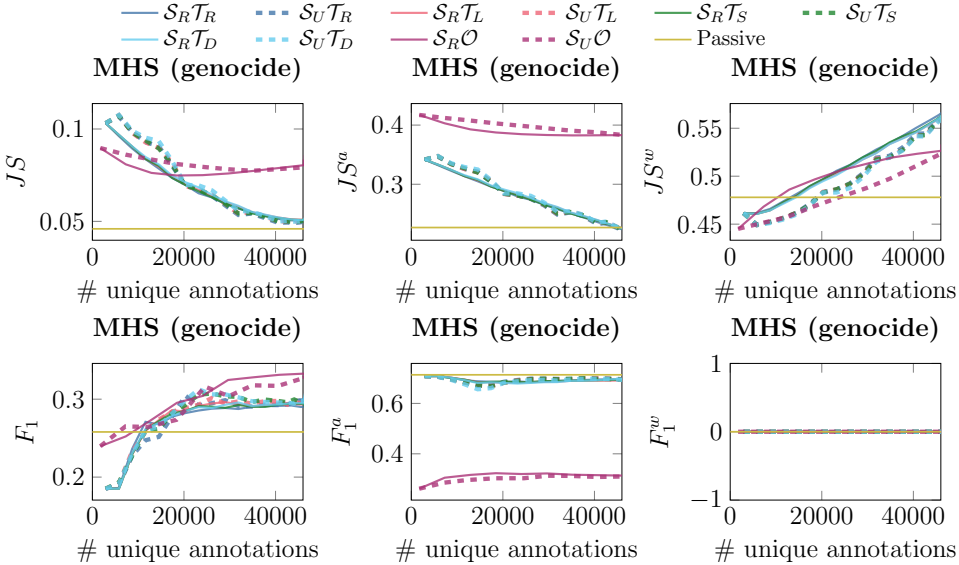


Figure D.7: Validation set performance across all metrics for MHS (genocide) during training

D

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

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

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

D

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

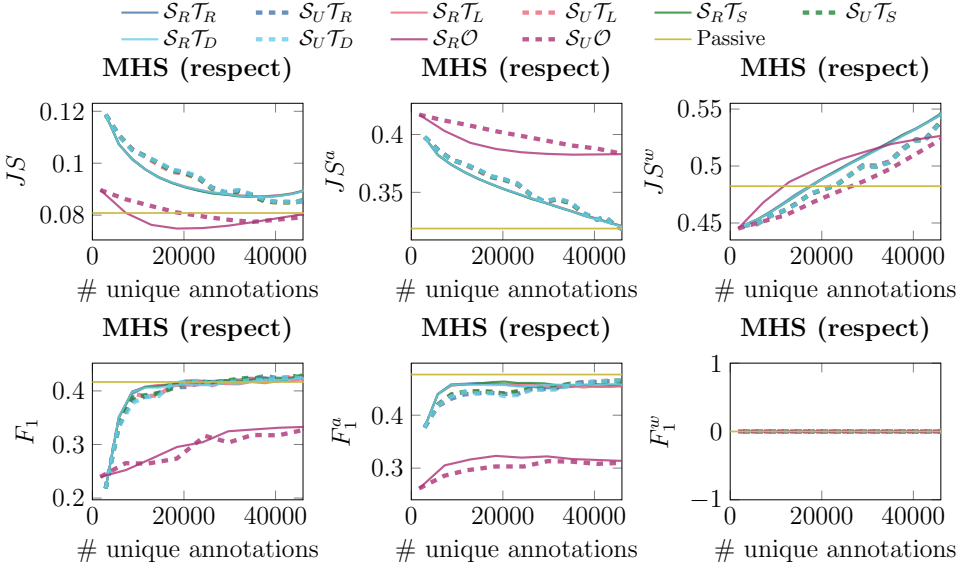


Figure D.8: Validation set performance across all metrics for MHS (*respect*) during training

E

Do Differences in Values Influence Disagreements in Online Discussions?

E

E.1 Methodological details

E.1.1 Training Value extraction methods

For training our Transformer-based NLP models, we turned to the Huggingface transformers Python package [436]. See Table E.1 for the hyperparameters used for training value extraction models. All computational experiments were run on machines containing up to 2x 3090 Nvidia RTX GPUs. Training a single value extraction model takes around 3 hours. Running VPE on background data takes significantly longer due to the number of inferences made, up to 7 days of computation.

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

Table E.1: Hyperparameters used for training models for value extraction

Filtering Reddit data We construct value profiles from the data scraped from Reddit, from which we filter posts not likely to be of relevance to discussing widespread societal issues. We remove posts from (1) NSFW subreddits¹, (2) gaming subreddits², (3) image-related subreddits³, (4) user subreddits, all subreddits starting with “u_”, (5) non-English posts (as detected

¹<https://www.reddit.com/r/ListOfSubreddits/wiki/nsfw>

²<https://www.reddit.com/r/gaming/wiki/faq>

³<https://www.reddit.com/r/ListOfSubreddits/wiki/sfwporn>

using the FastText [191] Language Identification model⁴), (6) and subreddits for which we could extract less than 50 posts.

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

1. Load words from Ponizovskiy et al. [304]. Some values have more words in the dictionary, and thus we introduce a weighting scheme to normalize over the number of words, such that a value v inside the profile with relatively few dictionary words has a higher weight w_v .
2. Replace URLs with a special [URL] token.
3. Apply lemmatization to all comments from a single user.
4. Classify individual comments for values. If a comment contains at least one term from the VD, classify the comment as being relevant for that value.
5. Aggregate over all comments.
6. Apply weighting $z = \text{count}(v) \times w_v$.
7. Apply normalization over the profile so it sums to 1.

E.1.2 Annotator experiment

We separated our annotator experiment into two phases: (1) the filling in of the PVQ-21, and (2) providing judgments on posts from Disagreement. The first phase was performed through Qualtrics questionnaire software. We provide screenshots of all steps (informed consent, annotation instructions) below. The second phase is hosted on Prodigy [270].

- **Informed consent** See Figure E.1. Shown to users before starting the experiment outlining the data protection and disclaimers of any risks.
- **Value Survey** See Figure E.2. Users fill in 21 items on a Likert scale.
- **Annotation instructions** See Figure E.3.
- **Annotation interface** See Figure E.4. Users were asked to fill in 25 task instances (five per subcorpus) on the annotation platform.

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

⁴<https://fasttext.cc/docs/en/language-identification.html>

Purpose of this research study: In this study, we aim at obtaining your preferences across a set of personal values, as well as your opinion on statements made in online discussions.

What you will do in the study: You will fill in a questionnaire where you indicate whether you identify with 21 statements. Optionally, you may be selected to fill in your opinion on a series of statements. Additional details are available in the annotation instructions.

Time required: It is dependent on you, as will be explained in the following instructions. The questionnaire takes an estimated 5 minutes to complete. If you are selected to provide your opinion on a series of statements, an additional 15 minutes is required.

Risks: There are no risks anticipated in this study. However, in case of doubts or concerns, do not hesitate to contact the researchers.

Privacy and confidentiality: Should you agree to take part, your participation will be completely confidential. All information gathered in the survey will be stored securely in compliance with the standards set by the European Union General Data Protection Regulation (GDPR). No one outside the research group will have access to the data during the research period. Background data will be kept by the research group until the analyses are finalized, at the latest in December 2023. No personal information is gathered by our platform. Upon analysis and publication, anonymized and aggregated information will be made available on open access for other researchers to analyze.

Right to withdraw from the study: Participation in the study is completely voluntary. If at any time you do not wish to continue your participation, you are welcome to withdraw from the survey without penalty.

How to withdraw from the study: You can end your participation by closing the browser window. If you want to withdraw your participation after completing a session, please contact us through email by sending a message to RESEARCHER NAME/EMAIL and mention your Prolific ID, or reach out on the Prolific platform. It is only possible to withdraw up to 2 months after the end of participation. It is not possible to withdraw after the publication of the data.

Questions? For questions, concerns, or complaints, please contact RESEARCHER NAME/EMAIL.

If you wish to participate in this study and agree with the informed consent, please select the "I am not a robot" box below.

E

Please select box. Click to write the question text

☐ I'm not a robot



Q14. Enter your Prolific ID

Your Prolific ID is required in order to process your contributions and provide you with the survey reward at the end.

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

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

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

Here we briefly describe different people. Please read each description and think about how much that person is or is not like you. Select the option that indicates how much the person described is like you.

Not like me at all

Not like me

A little like me

Moderately like me

Like me

Very much like me

Thinking up new ideas and being creative is important to him. He likes to do things in his own original way.

☐☐☐☐☐☐

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

E

E.1.3 Training agreement analysis models

Training models for agreement analysis takes around 4 hours for the BERT models on the subsampled Debagreement dataset. See Table E.2 for the hyperparameters used. Debagreement may be reused under the CC BY 4.0 license. For the implementation of the TF-IDF, we used the sklearn [299] Python package. All training involving TF-IDF embeddings takes under 1 hour.

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

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

- We constructed three types of extra user information for the agreement analysis task:
- Random noise** We sample a vector of size 768 from a random uniform distribution over $[0, 1)$.
 - User centroids** We stem the posts from users that contain at least one value term according to the value dictionary and transform comments to TF-IDF vectors. We restrict the vocabulary to the 768 most frequent terms. We then compute the average over all vectors for a single user.
 - Explicit user features** We construct user feature vectors for Reddit users through the Reddit PRAW API. See Table E.3 for the features used.

Opinion Experiment

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

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

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

Workflow

We suggest to use the following workflow.

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

Rules & Tips

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

Annotation Interface

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



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

prodigy ? ✓

PROJECT INFO

DATASET debugreement_small
SESSION test_user99
RECIPE debugreement.annotate
VIEW ID blocks

PROGRESS

THIS SESSION 7
TOTAL 7

ACCEPT 7
REJECT 0
IGNORE 0

HISTORY

Brexit Britain's Post-Brexit Identity Crisis - Boris Johnson has contradictory ideas for his country's future — and no clear paths for getting there

User 1: Britain had an identity centuries before the EU even existed. Britain doesn't need to be defined by Brexit. No nation has ever left the EU before, so we're in uncharted waters. Of course we don't know where we're going - we're the pioneers here. Thankfully, we have a sturdy vessel and an experienced statesman at the helm. Where we end up depends largely on the outcome of trade negotiations. Therefore, the priority now is those negotiations. We'll cross the other bridges when we come to them.

Please select your opinion with respect to what User 1 said:

☐ 🐱 I agree with User 1 1

☐ 🐱 I disagree with User 1 2

☒ 🐱 I am neutral towards User 1 3

☐ 🐱 I don't have enough information to decide 4

✓

© 2017-2023 Explosion (Prodigy v1.11.8)

Figure E.4: Annotation interface.

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

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

E.2 Additional Results

E.2.1 Value Extraction

For a complete overview of the performance of the value extraction models, including the standard deviation over 10 random seeds for the *VE* models, see Table E.4.

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

Table E.4: Macro-averaged performance of the value estimation approaches on the value datasets, showing averages and standard deviation for our own models over 10 different seeds. VN denotes ValueNet, VA denotes ValueArg. Methods marked with * are trained on a different objective than our VE task.

E.2.2 Value Survey

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

Reliability Since the PVQ has two questions for each personal value, we are able to compute internal consistency using Cronbach α per value. See the results in Table E.5. We observe a wide range of reliability scores, of which only conformity reaches above a score of 0.7. Most interestingly, we see that tradition is of very low reliability, possibly due to the demographic of some of our participants (students). Three task instances received mostly neutral or not-enough-information labels, and were disregarded in our analysis.

E.2.3 Qualitative Examples of Value Conflicts and (Dis-)agreement

We perform a qualitative analysis of some instances (comment pairs) from the dataset that follow our hypothesis and some that do not to gain a better understanding of when value conflicts influence disagreement. Table E.6 shows examples of the types of pairs we analyze.

E.2.4 Decomposition of BF_{10} results

We create overviews of the different tests performed in Sections 6.4.3 and 6.4.3. We decompose the aggregated scores into three separate figures, each showing how a single variable (either subreddit, similarity score, or profile threshold) impacts the obtained results. We

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

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

show the decomposition for the BF_{10} scores obtained for comparisons between two VPE-estimated profiles in Figures E.5 and for the comparison between VPE and self-reports in Figure E.6. In the latter case, since we picked samples from Disagreement with authors with populated value profiles, we do not need to test over multiple profile thresholds.

We show the highest and lowest BF_{10} scores and the test parameters in Tables E.7 and E.8 between two VPE profiles, and in Tables E.9 and E.10 for the experiments comparing VPE and self-reported profiles.

E.2.5 Kendall τ vs. Spearman ρ

We include a comparative overview of the tests that use the Kendall τ and add the BF_{10} scores for the same tests conducted with Spearman ρ . See Figure E.7. We see that generally, the ρ scores are similarly distributed as the τ scores. Two tests that for τ fall into the undecidable range, for ρ favor the null hypothesis H_0 . We attribute this to the size of our value profiles: since we have only 10 entries, ties are likely, and Spearman ρ does not explicitly account for them.

E.2.6 Agreement Analysis

For additional results (Precision, Recall, F_1 scores, accuracy, and the change w.r.t. a text-only baseline), see Table E.11.

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

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

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

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

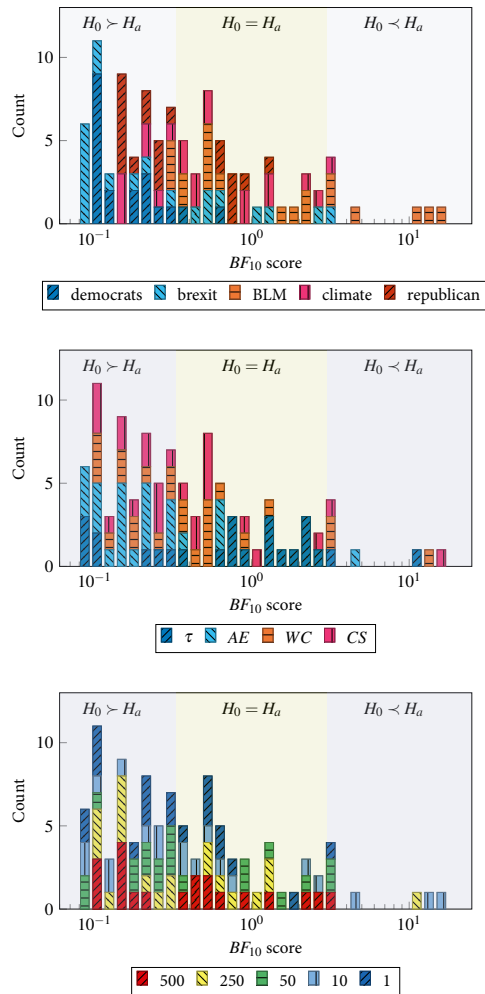


Figure E.5: BF_{10} scores when testing between two VPE-constructed profiles, obtained for all combinations of subreddits (top figure), similarity scores (middle figure) and profile thresholds (bottom figure).

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

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

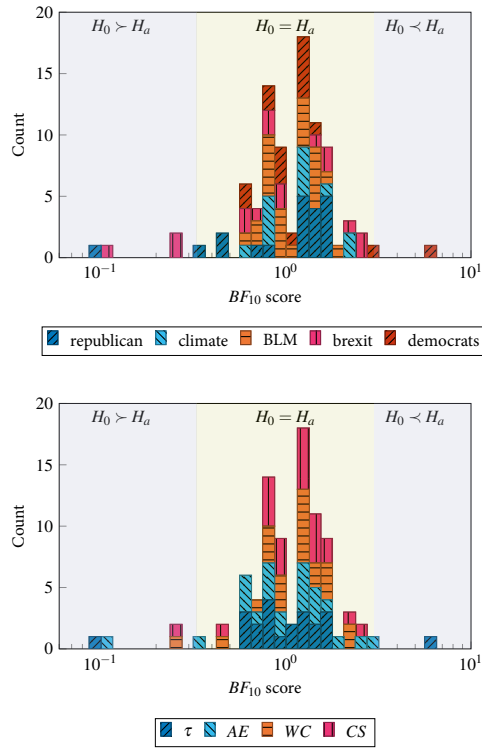
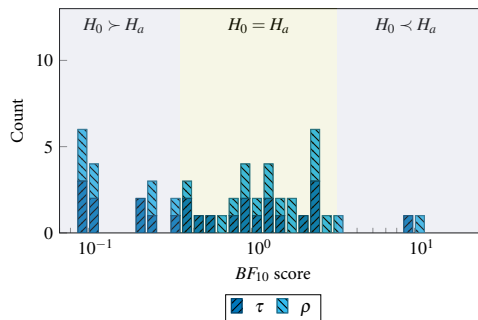


Figure E.6: BF_{10} scores when testing between a VPE-constructed profile and a self-reported profile, split into different subreddits (top figure) and different similarity scores (bottom figure).

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

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

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

Table E.10: The five tests between a VPE-constructed profile and a self-reported profile with the highest BF_{10} scores.Figure E.7: BF_{10} scores when testing between two VPE-constructed profiles, obtained for the similarity scores Kendall τ and Spearman ρ .

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

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

Bibliography

References

- [1] Gavin Abercrombie, Amanda Cercas Curry, Tanvi Dinkar, Verena Rieser, and Zeerak Talat. Mirages. on anthropomorphism in dialogue systems. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4776–4790, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.290.
- [2] Francisca Adoma Acheampong, Chen Wenyu, and Henry Nunoo-Mensah. Text-based emotion detection: Advances, challenges, and opportunities. *Engineering Reports*, 2(7):e12189, 2020.
- [3] G. Adomavicius and A. Tuzhilin. Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering*, 17(6):734–749, 2005. doi: 10.1109/TKDE.2005.99.
- [4] Mikel Aickin and Helen Gensler. Adjusting for multiple testing when reporting research results: the bonferroni vs holm methods. *American journal of public health*, 86(5):726–728, 1996.
- [5] Zeynep Akata, Dan Balliet, Maarten De Rijke, Frank Dignum, Virginia Dignum, Gusztai Eiben, Antske Fokkens, Davide Grossi, Koen Hindriks, Holger Hoos, et al. A research agenda for hybrid intelligence: augmenting human intellect with collaborative, adaptive, responsible, and explainable artificial intelligence. *Computer*, 53(8): 18–28, 2020. doi: 10.1109/MC.2020.2996587.
- [6] Khalid Al Khatib, Michael Völske, Shahbaz Syed, Nikolay Kolyada, and Benno Stein. Exploiting personal characteristics of debaters for predicting persuasiveness. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7067–7072, 2020. doi: 10.18653/v1/2020.acl-main.632.
- [7] Abeer ALDayel and Walid Magdy. Stance detection on social media: State of the art and trends. *Information Processing & Management*, 58(4):102597, 2021. ISSN 0306-4573. doi: <https://doi.org/10.1016/j.ipm.2021.102597>.
- [8] Dimitrios Alivanistos, Selene Báez Santamaría, Michael Cochez, Jan-Christoph Kalo, Emile van Krieken, and Thiviyan Thanapalasingam. Prompting as probing: Using language models for knowledge base construction, 2022.
- [9] Emily Allaway and Kathleen McKeown. Zero-Shot Stance Detection: A Dataset and Model using Generalized Topic Representations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8913–8931, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.717.

- [10] Milad Alshomary, Timon Gurcke, Shahbaz Syed, Philipp Heinisch, Maximilian Splithöfer, Philipp Cimiano, Martin Potthast, and Henning Wachsmuth. Key point analysis via contrastive learning and extractive argument summarization. In *Proceedings of the 8th Workshop on Argument Mining*, pages 184–189, 2021.
- [11] Milad Alshomary, Roxanne El Baff, Timon Gurcke, and Henning Wachsmuth. The moral debater: A study on the computational generation of morally framed arguments. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8782–8797, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.601.
- [12] Milad Alshomary, Jonas Rieskamp, and Henning Wachsmuth. Generating contrastive snippets for argument search. In *Computational Models of Argument*, pages 21–31. IOS Press, 2022.
- [13] Reinald Kim Amplayo, Stefanos Angelidis, and Mirella Lapata. Aspect-controllable opinion summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6578–6593, 2021.
- [14] Anitha Anandhan, Liyana Shuib, Maizatul Akmar Ismail, and Ghulam Mujtaba. Social media recommender systems: review and open research issues. *IEEE Access*, 6: 15608–15628, 2018.
- [15] Stefanos Angelidis and Mirella Lapata. Summarizing opinions: Aspect extraction meets sentiment prediction and they are both weakly supervised. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun’ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3675–3686, Brussels, Belgium, October–November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1403.
- [16] Stefanos Angelidis, Reinald Kim Amplayo, Yoshihiko Suhara, Xiaolan Wang, and Mirella Lapata. Extractive opinion summarization in quantized transformer spaces. *Transactions of the Association for Computational Linguistics*, 9:277–293, 2021.
- [17] Miguel Arana-Catania, Felix-Anselm Van Lier, Rob Procter, Nataliya Tkachenko, Yulan He, Arkaitz Zubiaga, and Maria Liakata. Citizen participation and machine learning for a better democracy. *Digital Government: Research and Practice*, 2(3):1–22, 2021.
- [18] Nikolay Archak, Anindya Ghose, and Panagiotis G Ipeirotis. Show me the money! deriving the pricing power of product features by mining consumer reviews. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 56–65, 2007.
- [19] Lisa P. Argyle, Christopher A. Bail, Ethan C. Busby, Joshua R. Gubler, Thomas Howe, Christopher Rytting, Taylor Sorensen, and David Wingate. Leveraging ai for democratic discourse: Chat interventions can improve online political conversations at scale. *Proceedings of the National Academy of Sciences*, 120(41):e2311627120, 2023. doi: 10.1073/pnas.2311627120.

- [20] Lisa P Argyle, Ethan C Busby, Nancy Fulda, Joshua R Gubler, Christopher Rytting, and David Wingate. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351, 2023. doi: 10.1017/pan.2023.2.
- [21] Lora Aroyo and Chris Welty. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1):15–24, Mar. 2015. doi: 10.1609/aimag.v36i1.2564.
- [22] Lora Aroyo, Alex S Taylor, Mark Diaz, Christopher M Homan, Alicia Parrish, Greg Serapio-Garcia, Vinodkumar Prabhakaran, and Ding Wang. Dices dataset: Diversity in conversational ai evaluation for safety. *arXiv preprint arXiv:2306.11247*, 2023.
- [23] Erfan Sadeqi Azer, Daniel Khashabi, Ashish Sabharwal, and Dan Roth. Not all claims are created equal: Choosing the right statistical approach to assess hypotheses. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5715–5725, 2020.
- [24] Joris Baan, Wilker Aziz, Barbara Plank, and Raquel Fernández. Stop Measuring Calibration When Humans Disagree. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1892–1915, 2022. doi: 10.18653/v1/2022.emnlp-main.124.
- [25] André Bächtiger, Simon Niemeyer, Michael Neblo, Marco R Steenbergen, and Jürg Steiner. Disentangling diversity in deliberative democracy: Competing theories, their blind spots and complementarities. *Journal of political philosophy*, 18(1), 2010.
- [26] André Bächtiger, John S Dryzek, Jane Mansbridge, and Mark Warren. Deliberative democracy. *The Oxford handbook of deliberative democracy*, pages 1–34, 2018.
- [27] Christopher A. Bail, Lisa P. Argyle, Taylor W. Brown, John P. Bumpus, Haohan Chen, M. B. Fallin Hunzaker, Jaemin Lee, Marcus Mann, Friedolin Merhout, and Alexander Volfovsky. Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences*, 115(37):9216–9221, 2018. doi: 10.1073/pnas.1804840115.
- [28] Michiel Bakker, Martin Chadwick, Hannah Sheahan, Michael Tessler, Lucy Campbell-Gillingham, Jan Balaguer, Nat McAleese, Amelia Glaese, John Aslanides, Matt Botvinick, et al. Fine-tuning language models to find agreement among humans with diverse preferences. *Advances in Neural Information Processing Systems*, 35:38176–38189, 2022.
- [29] Eytan Bakshy, Solomon Messing, and Lada A Adamic. Exposure to ideologically diverse news and opinion on facebook. *Science*, 348(6239):1130–1132, 2015. doi: 10.1126/science.aaa1160.
- [30] Dan Bang and Chris D Frith. Making better decisions in groups. *Royal Society open science*, 4(8):170193, 2017.
- [31] Roy Bar-Haim, Indrajit Bhattacharya, Francesco Dinuzzo, Amrita Saha, and Noam Slonim. Stance classification of context-dependent claims. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 251–261, 2017.

- [32] Roy Bar-Haim, Lilach Eden, Roni Friedman, Yoav Kantor, Dan Lahav, and Noam Slonim. 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, Online, July 2020. Association for Computational Linguistics, Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.371.
- [33] Roy Bar-Haim, Yoav Kantor, Lilach Eden, Roni Friedman, Dan Lahav, and Noam Slonim. 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, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.3.
- [34] Roy Bar-Haim, Lilach Eden, Yoav Kantor, Roni Friedman, and Noam Slonim. Every bite is an experience: Key point analysis of business reviews. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *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 3376–3386, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.262.
- [35] Roy Bar-Haim, Yoav Kantor, Elad Venezian, Yoav Katz, and Noam Slonim. Project Debater APIs: Decomposing the AI grand challenge. In Heike Adel and Shuming Shi, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 267–274, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-demo.31.
- [36] Jason Barabas. How deliberation affects policy opinions. *American political science review*, 98(4):687–701, 2004.
- [37] Sugato Basu, Arindam Banerjee, and Raymond J. Mooney. *Active Semi-Supervision for Pairwise Constrained Clustering*, chapter 32, pages 333–344. Society for Industrial and Applied Mathematics, 2004. doi: 10.1137/1.9781611972740.31.
- [38] Eric PS Baumer, Mahmood Jasim, Ali Sarvghad, and Narges Mahyar. Of course it’s political! a critical inquiry into underemphasized dimensions in civic text visualization. In *Computer Graphics Forum*, volume 41, pages 1–14. Wiley Online Library, 2022.
- [39] Connor Baumler, Anna Sotnikova, and Hal Daumé III. Which Examples Should be Multiply Annotated? Active Learning When Annotators May Disagree. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10352–10371. ACL, 2023.
- [40] Connor Baumler, Anna Sotnikova, and Hal Daumé III. Which examples should be multiply annotated? active learning when annotators may disagree. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10352–10371, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.658.

- [41] Jordan Beck, Bikalpa Neupane, and John M. Carroll. Managing conflict in online debate communities. *First Monday*, 24(7), Jun. 2019. doi: 10.5210/fm.v24i7.9585.
- [42] Tilman Beck, Ji-Ung Lee, Christina Viehmann, Marcus Maurer, Oliver Quiring, and Iryna Gurevych. 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, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.1.
- [43] Tilman Beck, Hendrik Schuff, Anne Lauscher, and Iryna Gurevych. How (not) to use sociodemographic information for subjective NLP tasks. *arXiv preprint arXiv:2309.07034*, 2023.
- [44] Tilman Beck, Hendrik Schuff, Anne Lauscher, and Iryna Gurevych. Sensitivity, performance, robustness: Deconstructing the effect of sociodemographic prompting. In Yvette Graham and Matthew Purver, editors, *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2589–2615, St. Julian’s, Malta, March 2024. Association for Computational Linguistics.
- [45] Jacob Beel, Tong Xiang, Sandeep Soni, and Diyi Yang. Linguistic characterization of divisive topics online: Case studies on contentiousness in abortion, climate change, and gun control. *Proceedings of the International AAAI Conference on Web and Social Media*, 16(1):32–42, May 2022. doi: 10.1609/icwsm.v16i1.19270.
- [46] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623, 2021.
- [47] Jamal Bentahar, Bernard Moulin, and Micheline Bélanger. A taxonomy of argumentation models used for knowledge representation. *The Artificial Intelligence Review*, 33(3):211, 2010.
- [48] Surbhi Bhatia. A comparative study of opinion summarization techniques. *IEEE Transactions on Computational Social Systems*, 8(1):110–117, 2020.
- [49] Haji Binali, Vidyasagar Potdar, and Chen Wu. A state of the art opinion mining and its application domains. In *2009 IEEE International Conference on Industrial Technology*, pages 1–6. IEEE, 2009.
- [50] Or Biran and Owen Rambow. Identifying justifications in written dialogs. In *2011 IEEE Fifth International Conference on Semantic Computing*, pages 162–168, 2011. doi: 10.1109/ICSC.2011.41.
- [51] Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. Language (technology) is power: A critical survey of “bias” in nlp. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, 2020.

- [52] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, oct 2008. doi: 10.1088/1742-5468/2008/10/p10008.
- [53] James Bohman. Deliberative democracy and the epistemic benefits of diversity. *Episteme*, 3(3):175–191, 2006.
- [54] Filip Boltužić and Jan Šnajder. Identifying prominent arguments in online debates using semantic textual similarity. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 110–115, 2015.
- [55] Thijs Bouman, Linda Steg, and Henk A. L. Kiers. Measuring values in environmental research: A test of an environmental portrait value questionnaire. *Frontiers in Psychology*, 9, 2018. ISSN 1664-1078. doi: 10.3389/fpsyg.2018.00564.
- [56] Samuel R. Bowman and George Dahl. What will it take to fix benchmarking in natural language understanding? In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4843–4855, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.385.
- [57] Ryan Boyd, Steven Wilson, James Pennebaker, Michal Kosinski, David Stillwell, and Rada Mihalcea. Values in words: Using language to evaluate and understand personal values. *Proceedings of the International AAAI Conference on Web and Social Media*, 9 (1):31–40, Aug. 2021. doi: 10.1609/icwsm.v9i1.14589.
- [58] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
- [59] Moritz Büchi, Eduard Fosch-Villaronga, Christoph Lutz, Aurelia Tamò-Larrieux, Shruthi Velidi, and Salome Viljoen. The chilling effects of algorithmic profiling: Mapping the issues. *Computer law & security review*, 36:105367, 2020.
- [60] Federico Cabitza, Andrea Campagner, and Valerio Basile. Toward a perspectivist turn in ground truthing for predictive computing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(6):6860–6868, Jun. 2023. doi: 10.1609/aaai.v37i6.25840.
- [61] Federico Cabitza, Andrea Campagner, and Valerio Basile. Toward a perspectivist turn in ground truthing for predictive computing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 6860–6868, Jun. 2023. doi: 10.1609/aaai.v37i6.25840.

- [62] Elena Cabrio and Serena Villata. Five years of argument mining: A data-driven analysis. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, IJCAI'18, page 5427–5433. AAAI Press, 7 2018. ISBN 9780999241127. doi: 10.24963/ijcai.2018/766.
- [63] Philippe Caillou, Jonas Renault, Jean-Daniel Fekete, Anne-Catherine Letournel, and Michèle Sebag. Cartolabe: A web-based scalable visualization of large document collections. *IEEE Computer Graphics and Applications*, 41(2):76–88, 2020.
- [64] Paul Cairney. *The politics of evidence-based policy making*. Springer, 2016.
- [65] Joel Castaño, Silverio Martínez-Fernández, Xavier Franch, and Justus Bogner. Analyzing the evolution and maintenance of ml models on hugging face. *arXiv preprint arXiv:2311.13380*, 2023.
- [66] Arie Cattan, Lilach Eden, Yoav Kantor, and Roy Bar-Haim. From key points to key point hierarchy: Structured and expressive opinion summarization. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 912–928, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.52.
- [67] Chengliang Chai, Guoliang Li, Jian Li, Dong Deng, and Jianhua Feng. Cost-effective crowdsourced entity resolution: A partial-order approach. In *Proceedings of the 2016 International Conference on Management of Data*, SIGMOD '16, page 969–984, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450335317. doi: 10.1145/2882903.2915252.
- [68] Tuhin Chakrabarty, Christopher Hidey, Smaranda Muresan, Kathy Mckeown, and Alyssa Hwang. Ampersand: Argument mining for persuasive online discussions. *arXiv preprint arXiv:2004.14677*, pages 2933–2943, November 2020. doi: 10.18653/v1/d19-1291.
- [69] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. A survey on evaluation of large language models. *ACM Trans. Intell. Syst. Technol.*, 15(3), mar 2024. ISSN 2157-6904. doi: 10.1145/3641289.
- [70] Jilin Chen, Gary Hsieh, Jalal U. Mahmud, and Jeffrey Nichols. Understanding individuals' personal values from social media word use. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*, CSCW '14, page 405–414, New York, NY, USA, 2014. Association for Computing Machinery. ISBN 9781450325400. doi: 10.1145/2531602.2531608.
- [71] Sihao Chen, Daniel Khashabi, Wenpeng Yin, Chris Callison-Burch, and Dan Roth. Seeing things from a different angle: Discovering diverse perspectives about claims. In Jill Burstein, Christy Doran, and Tamar Solorio, editors, *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, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1053.
- [72] Sihao Chen, Siyi Liu, Xander Uyttendaele, Yi Zhang, William Bruno, and Dan Roth. Design challenges for a multi-perspective search engine. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 293–303, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-naacl.22.
 - [73] Liying Cheng, Lidong Bing, Qian Yu, Wei Lu, and Luo Si. Ape: argument pair extraction from peer review and rebuttal via multi-task learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7000–7011, 2020.
 - [74] Robert Chew, Caroline Kery, Laura Baum, Thomas Bukowski, Annice Kim, and Mario Navarro. Predicting age groups of reddit users based on posting behavior and metadata: Classification model development and validation. *JMIR Public Health Surveill*, 7(3):e25807, Mar 2021. ISSN 2369-2960. doi: 10.2196/25807.
 - [75] Eric Chu and Peter Liu. Meansum: A neural model for unsupervised multi-document abstractive summarization. In *International Conference on Machine Learning*, pages 1223–1232. PMLR, 2019.
 - [76] Hyung Won Chung, Thibault Fevry, Henry Tsai, Melvin Johnson, and Sebastian Ruder. Rethinking embedding coupling in pre-trained language models. In *International Conference on Learning Representations*, 2020.
 - [77] Matteo Cinelli, Gianmarco De Francisci Morales, Alessandro Galeazzi, Walter Quattrociocchi, and Michele Starnini. The echo chamber effect on social media. *Proceedings of the National Academy of Sciences*, 118(9):e2023301118, 2021. doi: 10.1073/pnas.2023301118.
 - [78] Paul Clough, Robert Gaizauskas, Scott SL Piao, and Yorick Wilks. Measuring text reuse. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 152–159, 2002.
 - [79] Katherine M Collins, Umang Bhatt, and Adrian Weller. Eliciting and learning with soft labels from every annotator. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 10, pages 40–52, 2022.
 - [80] Luke Collins and Brigitte Nerlich. Examining user comments for deliberative democracy: A corpus-driven analysis of the climate change debate online. In *Climate Change Communication and the Internet*, pages 41–59. Routledge, 2019.
 - [81] Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 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, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1070.
- [82] Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*, 2022.
- [83] Michael Crawshaw. Multi-task learning with deep neural networks: A survey. *arXiv preprint arXiv:2009.09796*, 2020.
- [84] Scott Crossley, Luc Paquette, Mihai Dascalu, Danielle S McNamara, and Ryan S Baker. Combining click-stream data with nlp tools to better understand mooc completion. In *Proceedings of the sixth international conference on learning analytics & knowledge*, LAK '16, pages 6–14, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450341905. doi: 10.1145/2883851.2883931.
- [85] Christophe Croux and Catherine Dehon. Influence functions of the spearman and kendall correlation measures. *Statistical Methods & Applications*, 19(4):497–515, Nov 2010. ISSN 1613-981X. doi: 10.1007/s10260-010-0142-z.
- [86] Rowena Cullen. Addressing the digital divide. *Online information review*, 25(5):311–320, 2001. doi: 10.1108/14684520110410517.
- [87] Nicole Curato, John S Dryzek, Selen A Ercan, Carolyn M Hendriks, and Simon Niemeyer. Twelve key findings in deliberative democracy research. *Daedalus*, 146(3):28–38, 2017.
- [88] Lincoln Dahlberg. The internet and democratic discourse: Exploring the prospects of online deliberative forums extending the public sphere. *Information, communication & society*, 4(4):615–633, 2001. doi: 10.1080/13691180110097030.
- [89] Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110, 2022.
- [90] Jo Davies and Martin Graff. Performance in e-learning: online participation and student grades. *British Journal of Educational Technology*, 36(4):657–663, 2005. doi: <https://doi.org/10.1111/j.1467-8535.2005.00542.x>.
- [91] Todd Davies and Reid Chandler. *Democracy in motion: evaluating the practice and impact of deliberative civic engagement*, chapter Online Deliberation Design: Choices, Criteria, and Evidence, pages 103–131. Oxford University Press, 2012.
- [92] Johannes Daxenberger, Steffen Eger, Ivan Habernal, Christian Stab, and Iryna Gurevych. 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, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1218.

- [93] Johannes Daxenberger, Benjamin Schiller, Chris Stahlhut, Erik Kaiser, and Iryna Gurevych. Argumentext: argument classification and clustering in a generalized search scenario. *Datenbank-Spektrum*, 20:115–121, 2020.
- [94] Angel Daza and Anette Frank. 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, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.321.
- [95] Maaïke H de Boer, Jasper van der Waa, Sophie van Gent, Quirine T.S. Smit, Wouter Korteling, Robin M. van Stokkum, and Mark Neerincx. A contextual hybrid intelligent system design for diabetes lifestyle management. In *International Workshop Modelling and Representing Context, ECAI*, volume 23, 2023.
- [96] Christine De Kock and Andreas Vlachos. I beg to differ: A study of constructive disagreement in online conversations. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2017–2027, 2021.
- [97] Davide Dell’Anna, Pradeep K. Murukannaiah, Bernd Dudzik, Davide Grossi, Catholijn M. Jonker, Catharine Oertel, and Pinar Yolum. Toward a quality model for hybrid intelligence teams. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems*, pages 1–10, Auckland, 2024. To appear.
- [98] Dominik Dellermann, Philipp Ebel, Matthias Söllner, and Jan Marco Leimeister. Hybrid intelligence. *Business & Information Systems Engineering*, 61:637–643, 2019. doi: 10.1007/s12599-019-00595-2.
- [99] Emily Denton, Alex Hanna, Razvan Amironesei, Andrew Smart, Hilary Nicole, and Morgan Klaus Scheuerman. Bringing the people back in: Contesting benchmark machine learning datasets, 2020.
- [100] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423.
- [101] Thomas G Dietterich. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural computation*, 10(7):1895–1923, 1998.
- [102] Zijian Ding, Alison Smith-Renner, Wenjuan Zhang, Joel R Tetreault, and Alejandro Jaimes. Harnessing the power of llms: Evaluating human-ai text co-creation through the lens of news headline generation. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *The 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3321–3339, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.217.

- [103] Mark Dingemanse and Andreas Liesenfeld. From text to talk: Harnessing conversational corpora for humane and diversity-aware language technology. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5614–5633, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.385.
- [104] Tim Draws, Oana Inel, Nava Tintarev, Christian Baden, and Benjamin Timmermans. Comprehensive viewpoint representations for a deeper understanding of user interactions with debated topics. In *Proceedings of the 2022 Conference on Human Information Interaction and Retrieval, CHIIR '22*, pages 135–145, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450391863. doi: 10.1145/3498366.3505812.
- [105] Tim Draws, Nirmal Roy, Oana Inel, Alisa Rieger, Rishav Hada, Mehmet Orcun Yalcin, Benjamin Timmermans, and Nava Tintarev. Viewpoint diversity in search results. In *European Conference on Information Retrieval*, pages 279–297. Springer, 2023.
- [106] John S Dryzek, André Bächtiger, Simone Chambers, Joshua Cohen, James N Druckman, Andrea Felicetti, James S Fishkin, David M Farrell, Archon Fung, Amy Gutmann, et al. The crisis of democracy and the science of deliberation. *Science*, 363 (6432):1144–1146, 2019. doi: 10.1126/science.aaw2694.
- [107] Olive Jean Dunn. Multiple comparisons using rank sums. *Technometrics*, 6(3):241–252, 1964.
- [108] Esin Durmus, Karina Nyugen, Thomas I Liao, Nicholas Schiefer, Amanda Askill, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, et al. Towards measuring the representation of subjective global opinions in language models. *arXiv preprint arXiv:2306.16388*, 2023.
- [109] Charlie Egan, Advait Siddharthan, and Adam Wyner. Summarising the points made in online political debates. In *Proceedings of the 3rd Workshop on Argument Mining, The 54th Annual Meeting of the Association for Computational Linguistics*, pages 134–143. Association for Computational Linguistics (ACL), 2016.
- [110] Steffen Eger, Johannes Daxenberger, and Iryna Gurevych. 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, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1002.
- [111] Steffen Eger, Johannes Daxenberger, Christian Stab, and Iryna Gurevych. 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, August 2018. Association for Computational Linguistics.

- [112] Liat Ein-Dor, Eyal Shnarch, Lena Dankin, Alon Halfon, Benjamin Sznajder, Ariel Gera, Carlos Alzate, Martin Gleize, Leshem Choshen, Yufang Hou, Yonatan Bilu, Ranit Aharonov, and Noam Slonim. Corpus wide argument mining—a working solution. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7683–7691, Apr. 2020. doi: 10.1609/aaai.v34i05.6270.
- [113] Wafaa S El-Kassas, Cherif R Salama, Ahmed A Rafea, and Hoda K Mohamed. Automatic text summarization: A comprehensive survey. *Expert systems with applications*, 165:113679, 2021.
- [114] Tina Eliassi-Rad, Henry Farrell, David Garcia, Stephan Lewandowsky, Patricia Palacios, Don Ross, Didier Sornette, Karim Thébault, and Karoline Wiesner. What science can do for democracy: a complexity science approach. *Humanities and Social Sciences Communications*, 7(1):1–4, 2020.
- [115] Stephen Elstub and Peter McLaverty. *Deliberative Democracy: Issues and Cases*. Edinburgh University Press, Edinburgh, 2014. ISBN 9780748643509. doi: 10.1515/9780748643509.
- [116] Alexander Erdmann, David Joseph Wrisley, Benjamin Allen, Christopher Brown, Sophie Cohen-Bodénès, Micha Elsner, Yukun Feng, Brian Joseph, Béatrice Joyeux-Prunel, and Marie Catherine de Marneffe. Practical, Efficient, and Customizable Active Learning for Named Entity Recognition in the Digital Humanities. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics, NAACL’19*, pages 2223–2234, Minneapolis, Minnesota, USA, 2019. ACL. doi: 10.18653/v1/n19-1231.
- [117] K Esau and Dennis Friess. What creates listening online? exploring reciprocity in online political discussions with relational content analysis. *Journal of Deliberative Democracy*, 18(1):1–16, June 2022. doi: 10.16997/jdd.1021.
- [118] Katharina Esau, Dennis Friess, and Christiane Eilders. Design matters! an empirical analysis of online deliberation on different news platforms. *Policy & Internet*, 9(3): 321–342, 2017. doi: 10.1002/poi3.154.
- [119] Kevin M Esterling, Archon Fung, and Taeku Lee. How much disagreement is good for democratic deliberation? *Political Communication*, 32(4):529–551, 2015.
- [120] Alexander R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409, 2021.
- [121] Neele Falk, Iman Jundi, Eva Maria Vecchi, and Gabriella Lapesa. Predicting moderation of deliberative arguments: Is argument quality the key? In Khalid Al-Khatib, Yufang Hou, and Manfred Stede, editors, *Proceedings of the 8th Workshop on Argument Mining*, pages 133–141, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.argmining-1.13.

- [122] Wei Fang, Moin Nadeem, Mitra Mohtarami, and James Glass. Neural multi-task learning for stance prediction. In *Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER)*, pages 13–19, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-6603.
- [123] Andrea Felicetti, Simon Niemeyer, and Nicole Curato. Improving deliberative participation: Connecting mini-publics to deliberative systems. *European Political Science Review*, 8(3):427–448, 2016.
- [124] Shangbin Feng, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair nlp models. *arXiv preprint arXiv:2305.08283*, pages 11737–11762, July 2023. doi: 10.18653/v1/2023.acl-long.656.
- [125] Oliver Ferschke, Iryna Gurevych, and Yevgen Chebotar. Behind the article: Recognizing dialog acts in wikipedia talk pages. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 777–786, 2012.
- [126] Ken Fischer, Justin Reedy, Cameron Piercy, and Rashmi Thapaliya. A typology of reasoning in deliberative processes: A study of the 2010 oregon citizens’ initiative review. *Journal of Deliberative Democracy*, 18(2), 2022.
- [127] Lukas Fluri, Daniel Paleka, and Florian Tramèr. Evaluating superhuman models with consistency checks. *arXiv preprint arXiv:2306.09983*, 2023.
- [128] Chase Foster and Jeffrey Frieden. Crisis of trust: Socio-economic determinants of europeans’ confidence in government. *European Union Politics*, 18(4):511–535, 2017.
- [129] James B Freeman. *Argument structure*. Argumentation Library. Springer, Dordrecht, Netherlands, 2011 edition, March 2011. ISBN 978-94-007-3553-8. doi: 10.1007/978-94-007-0357-5.
- [130] Batya Friedman, Peter H. Kahn, Alan Borning, and Alina Huldtgren. *Value Sensitive Design and Information Systems*, pages 55–95. Springer Netherlands, Dordrecht, 2013. ISBN 978-94-007-7844-3. doi: 10.1007/978-94-007-7844-3_4.
- [131] Roni Friedman-Melamed, Lena Dankin, Yufang Hou, Ranit Aharonov, Yoav Katz, and Noam Slonim. Overview of the 2021 key point analysis shared task. In *Conference on Empirical Methods in Natural Language Processing*, 2021.
- [132] Dennis Friess and Christiane Eilders. A systematic review of online deliberation research. *Policy & Internet*, 7(3):319–339, 2015. doi: 10.1002/poi3.95.
- [133] Adam D Galinsky, William W Maddux, Debra Gilin, and Judith B White. Why it pays to get inside the head of your opponent: The differential effects of perspective taking and empathy in negotiations. *Psychological science*, 19(4):378–384, 2008.
- [134] Tianyu Gao, Xingcheng Yao, and Danqi Chen. SimCSE: Simple contrastive learning of sentence embeddings. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2021.

- [135] Albert Gatt and Emiel Krahmer. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, 61:65–170, 2018.
- [136] Hunter Gehlbach, Maureen E Brinkworth, and Ming-Te Wang. The social perspective taking process: What motivates individuals to take another’s perspective? *Teachers College Record*, 114(1):1–29, 2012.
- [137] Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A Smith. Realexityprompts: Evaluating neural toxic degeneration in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3356–3369, 2020.
- [138] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020. doi: 10.1038/s42256-020-00257-z.
- [139] Debanjan Ghosh, Ritvik Shrivastava, and Smaranda Muresan. “laughing at you or with you”: The role of sarcasm in shaping the disagreement space. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1998–2010, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.171.
- [140] Robert E Goodin and John S Dryzek. Deliberative impacts: The macro-political uptake of mini-publics. *Politics & society*, 34(2):219–244, 2006.
- [141] Jesse Graham, Jonathan Haidt, and Brian A Nosek. Liberals and conservatives rely on different sets of moral foundations. *Journal of personality and social psychology*, 96(5): 1029–1046, 2009. doi: <https://doi.org/10.1037/a0015141>.
- [142] Jesse Graham, Jonathan Haidt, Sena Koleva, Matt Motyl, Ravi Iyer, Sean P. Wojcik, and Peter H. Ditto. Moral Foundations Theory: The Pragmatic Validity of Moral Pluralism. In *Advances in Experimental Social Psychology*, volume 47, pages 55–130. Elsevier, Amsterdam, the Netherlands, 2013. doi: 10.1016/B978-0-12-407236-7.00002-4.
- [143] Jesse Graham, Jonathan Haidt, Sena Koleva, Matt Motyl, Ravi Iyer, Sean P Wojcik, and Peter H Ditto. Moral foundations theory: The pragmatic validity of moral pluralism. In Patricia Devine and Ashby Plant, editors, *Advances in experimental social psychology*, volume 47 of *Advances in Experimental Social Psychology*, pages 55–130. Elsevier, 2013. doi: <https://doi.org/10.1016/B978-0-12-407236-7.00002-4>.
- [144] Shai Gretz, Roni Friedman, Edo Cohen-Karlik, Assaf Toledo, Dan Lahav, Ranit Aharonov, and Noam Slonim. A large-scale dataset for argument quality ranking: Construction and analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7805–7813, Apr. 2020. doi: 10.1609/aaai.v34i05.6285.

- [145] Shai Gretz, Assaf Toledo, Roni Friedman, Dan Lahav, Rose Weeks, Naor Bar-zeev, João Sedoc, Pooja Sangha, Yoav Katz, and Noam Slonim. Benchmark data and evaluation framework for intent discovery around covid-19 vaccine hesitancy. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1328–1340, 2023.
- [146] Leo Groarke. Informal Logic. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Spring 2024 edition, 2024.
- [147] Maarten Grootendorst. Bertopic: Neural topic modeling with a class-based tf-idf procedure, 2022.
- [148] Davide Grossi, Ulrike Hahn, Michael Mäs, Andreas Nitsche, Jan Behrens, Niclas Boehmer, Markus Brill, Ulle Endriss, Umberto Grandi, Adrian Haret, et al. Enabling the digital democratic revival: A research program for digital democracy. *arXiv preprint arXiv:2401.16863*, 2024.
- [149] Cornelia Gruber, Katharina Hechinger, Matthias Assenmacher, Göran Kauermann, and Barbara Plank. More labels or cases? assessing label variation in natural language inference. In *Proceedings of the Third Workshop on Understanding Implicit and Under-specified Language*, pages 22–32, Malta, March 2024. Association for Computational Linguistics.
- [150] Max Grusky. Rogue scores. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1914–1934, 2023.
- [151] Ken Gu and Akshay Budhkar. A package for learning on tabular and text data with transformers. In *Proceedings of the Third Workshop on Multimodal Artificial Intelligence*, pages 69–73, Mexico City, Mexico, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.maiworkshop-1.10.
- [152] Zhen Guo, Zhe Zhang, and Munindar Singh. In opinion holders’s shoes: Modeling cumulative influence for view change in online argumentation. In *Proceedings of The Web Conference 2020*, pages 2388–2399, 2020. doi: 10.1145/3366423.3380302.
- [153] Ali Gürkan, Luca Iandoli, Mark Klein, and Giuseppe Zollo. Mediating debate through on-line large-scale argumentation: Evidence from the field. *Information Sciences*, 180 (19):3686–3702, 2010.
- [154] Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. Annotation artifacts in natural language inference data. In Marilyn Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-2017.
- [155] J. Habermas and T. Burger. *The Structural Transformation of the Public Sphere: An Inquiry into a Category of Bourgeois Society*. Studies in Contemporary German Social Thought. MIT Press, 1991. ISBN 9780262581080.

- [156] Ivan Habernal and Iryna Gurevych. Argumentation mining in user-generated web discourse. *Computational Linguistics*, 43(1):125–179, April 2017. doi: 10.1162/COLI_a_00276.
- [157] Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. The argument reasoning comprehension task: Identification and reconstruction of implicit warrants. In Marilyn Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1930–1940, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1175.
- [158] Rafik Hadfi and Takayuki Ito. Augmented democratic deliberation: Can conversational agents boost deliberation in social media? In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*, pages 1794–1798, 2022.
- [159] Daniel Halpern and Jennifer Gibbs. Social media as a catalyst for online deliberation? exploring the affordances of facebook and youtube for political expression. *Computers in human behavior*, 29(3):1159–1168, 2013.
- [160] Xuehua Han, Juanle Wang, Min Zhang, and Xiaojie Wang. Using social media to mine and analyze public opinion related to covid-19 in china. *International journal of environmental research and public health*, 17(8):2788, 2020.
- [161] Momchil Hardalov, Arnav Arora, Preslav Nakov, and Isabelle Augenstein. Cross-domain label-adaptive stance detection. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9011–9028, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.710.
- [162] Jochen Hartmann, Jasper Schwenzow, and Maximilian Witte. The political ideology of conversational ai: Converging evidence on chatgpt’s pro-environmental, left-libertarian orientation. *arXiv preprint arXiv:2301.01768*, 2023.
- [163] Brian W Head and John Alford. Wicked problems: Implications for public policy and management. *Administration & society*, 47(6):711–739, 2015.
- [164] Jeffrey Heer. Agency plus automation: Designing artificial intelligence into interactive systems. *Proceedings of the National Academy of Sciences*, 116(6):1844–1850, 2019. doi: 10.1073/pnas.1807184115.
- [165] Philipp Heinisch, Anette Frank, Juri Opitz, Moritz Plenz, and Philipp Cimiano. Overview of the 2022 validity and novelty prediction shared task. In Gabriella Lapesa, Jodi Schneider, Yohan Jo, and Sougata Saha, editors, *Proceedings of the 9th Workshop on Argument Mining*, pages 84–94, Online and in Gyeongju, Republic of Korea, October 2022. International Conference on Computational Linguistics.

- [166] Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. Aligning ai with shared human values. In *International Conference on Learning Representations*, 2020.
- [167] Kristin Page Hocevar, Andrew J Flanagin, and Miriam J Metzger. Social media self-efficacy and information evaluation online. *Computers in Human Behavior*, 39:254–262, 2014. ISSN 0747-5632. doi: <https://doi.org/10.1016/j.chb.2014.07.020>.
- [168] Lu Hong and Scott E Page. Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proceedings of the National Academy of Sciences*, 101(46):16385–16389, 2004.
- [169] Joe Hoover, Gwenyth Portillo-Wightman, Leigh Yeh, Shreya Havaldar, Aida Mostafazadeh Davani, Ying Lin, Brendan Kennedy, Mohammad Atari, Zahra Kamel, Madelyn Mendlen, Gabriela Moreno, Christina Park, Tingyee E. Chang, Jenna Chin, Christian Leong, Jun Yen Leung, Arineh Mirinjian, and Morteza Dehghani. Moral foundations twitter corpus: A collection of 35k tweets annotated for moral sentiment. *Social Psychological and Personality Science*, 11(8):1057–1071, 11 2020. ISSN 19485514. doi: [10.1177/1948550619876629](https://doi.org/10.1177/1948550619876629).
- [170] Dirk Hovy and Shrimai Prabhumoye. Five sources of bias in natural language processing. *Language and Linguistics Compass*, 15(8):e12432, 2021.
- [171] Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. Learning whom to trust with mace. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1120–1130, 2013.
- [172] Jeremy Howard and Sebastian Ruder. 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, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: [10.18653/v1/P18-1031](https://doi.org/10.18653/v1/P18-1031).
- [173] Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.
- [174] J. Edward Hu, Abhinav Singh, Nils Holzenberger, Matt Post, and Benjamin Van Durme. Large-scale, diverse, paraphrastic bitexts via sampling and clustering. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 44–54, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: [10.18653/v1/K19-1005](https://doi.org/10.18653/v1/K19-1005).
- [175] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*, 2023.

- [176] Nannan Huang, Lin Tian, Haytham Fayek, and Xiuzhen Zhang. Examining bias in opinion summarisation through the perspective of opinion diversity. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 149–161, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.wassa-1.14.
- [177] Dieuwke Hupkes, Mario Giulianelli, Verna Dankers, Mikel Artetxe, Yanai Elazar, Tiago Pimentel, Christos Christodoulopoulos, Karim Lasri, Naomi Saphra, Arabella Sinclair, et al. A taxonomy and review of generalization research in nlp. *Nature Machine Intelligence*, 5(10):1161–1174, 2023.
- [178] Luca Iandoli, Ivana Quinto, Anna De Liddo, and Simon Buckingham Shum. Socially augmented argumentation tools: Rationale, design and evaluation of a debate dashboard. *International Journal of Human-Computer Studies*, 72(3):298–319, 2014. ISSN 1071-5819. doi: <https://doi.org/10.1016/j.ijhcs.2013.08.006>.
- [179] IBM Research. The Project Debater Service API. <https://developer.ibm.com/apis/catalog/debater--project-debater-service-api/Introduction>, 2023. Accessed: October 2023.
- [180] David Inouye and Jugal K Kalita. Comparing twitter summarization algorithms for multiple post summaries. In *2011 IEEE Third international conference on privacy, security, risk and trust and 2011 IEEE third international conference on social computing*, pages 298–306. IEEE, 2011.
- [181] IPCC. Climate change 2022: Impacts, adaptation, and vulnerability. <https://www.ipcc.ch/report/ar6/syr/>, 2022.
- [182] Karen A. Jehn. Enhancing effectiveness: An investigation of advantages and disadvantages of value-based intragroup conflict. *International Journal of Conflict Management*, 5(3):223–238, Jan 1994. ISSN 1044-4068. doi: 10.1108/eb022744.
- [183] Hamed Jelodar, Yongli Wang, Rita Orji, and Shucheng Huang. Deep sentiment classification and topic discovery on novel coronavirus or covid-19 online discussions: Nlp using lstm recurrent neural network approach. *IEEE Journal of Biomedical and Health Informatics*, 24(10):2733–2742, 2020. doi: 10.1109/JBHI.2020.3001216.
- [184] Shagun Jhaver, Amy Bruckman, and Eric Gilbert. Does transparency in moderation really matter? user behavior after content removal explanations on reddit. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW), nov 2019. doi: 10.1145/3359252.
- [185] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12), mar 2023. ISSN 0360-0300. doi: 10.1145/3571730.
- [186] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

- [187] Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. Tinybert: Distilling bert for natural language understanding. *arXiv preprint arXiv:1909.10351*, 2019.
- [188] Yohan Jo and Alice H Oh. Aspect and sentiment unification model for online review analysis. In *Proceedings of the fourth ACM international conference on Web search and data mining*, WSDM '11, pages 815–824, New York, NY, USA, 2011. Association for Computing Machinery. ISBN 9781450304931. doi: 10.1145/1935826.1935932.
- [189] Carolina Centeio Jorge, Emma M van Zoelen, Ruben Verhagen, Siddharth Mehrotra, Catholijn M Jonker, and Myrthe L Tielman. Appropriate context-dependent artificial trust in human-machine teamwork. In *Putting AI in the Critical Loop*, pages 41–60. Elsevier, 2024.
- [190] Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. The state and fate of linguistic diversity and inclusion in the NLP world. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.560.
- [191] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain, April 2017. Association for Computational Linguistics.
- [192] Kamil Kanclerz, Konrad Karanowski, Julita Bielaniewicz, Marcin Gruza, Piotr Miłkowski, Jan Kocon, and Przemysław Kazienko. Pals: Personalized active learning for subjective tasks in nlp. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13326–13341, 2023.
- [193] Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. Large language models struggle to learn long-tail knowledge. In *International Conference on Machine Learning*, pages 15696–15707. PMLR, 2023.
- [194] Georgi Karadzhov, Andreas Vlachos, and Tom Stafford. The effect of diversity on group decision-making. *arXiv preprint arXiv:2402.01427*, 2024.
- [195] Robert E. Kass and Adrian E. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795, 1995. doi: 10.1080/01621459.1995.10476572.
- [196] Brendan Kennedy, Mohammad Atari, Aida Mostafazadeh Davani, Joe Hoover, Ali Omrani, Jesse Graham, and Morteza Dehghani. Moral concerns are differentially observable in language. *Cognition*, 212:104696, 2021.
- [197] Dana Khartabil, Christopher Collins, Simon Wells, Benjamin Bach, and Jessie Kennedy. Design and evaluation of visualization techniques to facilitate argument exploration. *Computer Graphics Forum*, 40(6):447–465, 2021.

- [198] Urja Khurana, Eric Nalisnick, Antske Fokkens, and Swabha Swayamdipta. Crowd-calibrator: Can annotator disagreement inform calibration in subjective tasks? In *First Conference on Language Modeling*, Philadelphia, PA, 2024.
- [199] Urja Khurana, Eric Nalisnick, Antske Fokkens, and Swabha Swayamdipta. Crowd-calibrator: Can annotator disagreement inform calibration in subjective tasks?, 2024.
- [200] Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. Dynabench: Rethinking benchmarking in NLP. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4110–4124, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.324.
- [201] Johannes Kiesel, Milad Alshomary, Nicolas Handke, Xiaoni Cai, Henning Wachsmuth, and Benno Stein. Identifying the human values behind arguments. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4459–4471, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.306.
- [202] Johannes Kiesel, Milad Alshomary, Nailia Mirzakhmedova, Maximilian Heinrich, Nicolas Handke, Henning Wachsmuth, and Benno Stein. Semeval-2023 task 4: Valueeval: Identification of human values behind arguments. In Atul Kr. Ojha, A. Seza Doğruöz, Giovanni Da San Martino, Harish Tayyar Madabushi, Ritesh Kumar, and Elisa Sartori, editors, *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2287–2303, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.semeval-1.313.
- [203] Johannes Kiesel, Milad Alshomary, Nailia Mirzakhmedova, Maximilian Heinrich, Nicolas Handke, Henning Wachsmuth, and Benno Stein. SemEval-2023 Task 4: ValueEval: Identification of Human Values behind Arguments. In *17th International Workshop on Semantic Evaluation*, SemEval '23, pages 2290–2306, Toronto, Canada, July 2023. Association for Computational Linguistics.
- [204] Mark Klein. Enabling large-scale deliberation using attention-mediation metrics. *Computer Supported Cooperative Work (CSCW)*, 21(4-5):449–473, 2012. doi: 10.1007/s10606-012-9156-4.
- [205] Mark Klein. Crowd-scale deliberation for group decision-making. *Handbook of group decision and negotiation*, pages 355–369, 2021.
- [206] Michele Knobel and Colin Lankshear. Digital literacy and participation in online social networking spaces. *Digital literacies: Concepts, policies and practices*, 11:249–278, 2008.

- [207] Jonathan Kobbe, Ines Rehbein, Ioana Hulpuş, and Heiner Stuckenschmidt. Exploring morality in argumentation. In *Proceedings of the 7th Workshop on Argument Mining*, pages 30–40, Online, December 2020. Association for Computational Linguistics.
- [208] Quyu Kong, Emily Booth, Francesco Bailo, Amelia Johns, and Marian-Andrei Rizoiu. Slipping to the extreme: A mixed method to explain how extreme opinions infiltrate online discussions. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, pages 524–535, May 2022. doi: 10.1609/icwsm.v16i1.19312.
- [209] Erik Körner, Ahmad Dawar Hakimi, Gerhard Heyer, and Martin Potthast. Casting the same sentiment classification problem. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 584–590, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.53.
- [210] Erik Körner, Gregor Wiedemann, Ahmad Dawar Hakimi, Gerhard Heyer, and Martin Potthast. On classifying whether two texts are on the same side of an argument. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10130–10138, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.795.
- [211] John K. Kruschke. Rejecting or accepting parameter values in bayesian estimation. *Advances in Methods and Practices in Psychological Science*, 1(2):270–280, 2018. doi: 10.1177/2515245918771304.
- [212] William H Kruskal and W Allen Wallis. Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association*, 47(260):583–621, 1952.
- [213] Gillian Ku, Cynthia S Wang, and Adam D Galinsky. The promise and perversity of perspective-taking in organizations. *Research in Organizational Behavior*, 35:79–102, 2015.
- [214] Dilek Küçük and Fazli Can. Stance detection: A survey. *ACM Computing Surveys (CSUR)*, 53(1):1–37, feb 2020. ISSN 0360-0300. doi: 10.1145/3369026.
- [215] Andrew P. Kythreotis, Chrystal Mantyka-Pringle, Theresa G. Mercer, Lorraine E. Whitmarsh, Adam Corner, Jouni Paavola, Chris Chambers, Byron A. Miller, and Noel Castree. Citizen social science for more integrative and effective climate action: A science-policy perspective. *Frontiers in Environmental Science*, 7, 2019. ISSN 2296-665X. doi: 10.3389/fenvs.2019.00010.
- [216] Philippe Laban, Chien-Sheng Wu, Lidiya Murakhovs’ka, Xiang Chen, and Caiming Xiong. Discord questions: A computational approach to diversity analysis in news coverage. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5180–5194, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.380.

- [217] Preethi Lahoti, Nicholas Blumm, Xiao Ma, Raghavendra Kotikalapudi, Sahitya Potluri, Qijun Tan, Hansa Srinivasan, Ben Packer, Ahmad Beirami, Alex Beutel, and Jilin Chen. Improving diversity of demographic representation in large language models via collective-critiques and self-voting. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10383–10405, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.643.
- [218] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.
- [219] Hélène Landemore. *Democratic reason: Politics, collective intelligence, and the rule of the many*. Princeton University Press, 2012.
- [220] Gabriella Lapesa, Eva Maria Vecchi, Serena Villata, and Henning Wachsmuth. Mining, assessing, and improving arguments in NLP and the social sciences. In Fabio Massimo Zanzotto and Sameer Pradhan, editors, *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts*, pages 1–6, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.eacl-tutorials.1.
- [221] Dora C Lau and J Keith Murnighan. Demographic diversity and faultlines: The compositional dynamics of organizational groups. *Academy of management review*, 23(2): 325–340, 1998.
- [222] Anne Lauscher, Lily Ng, Courtney Napoles, and Joel Tetreault. Rhetoric, logic, and dialectic: Advancing theory-based argument quality assessment in natural language processing. In Donia Scott, Nuria Bel, and Chengqing Zong, editors, *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4563–4574, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.402.
- [223] Anne Lauscher, Henning Wachsmuth, Iryna Gurevych, and Goran Glavaš. Scientia potentia est—on the role of knowledge in computational argumentation. *Transactions of the Association for Computational Linguistics*, 10:1392–1422, 2022.
- [224] John Lawrence and Chris Reed. Argument Mining: A Survey. *Computational Linguistics*, 45(4):765–818, 01 2020. ISSN 0891-2017. doi: 10.1162/coli_a_00364.
- [225] David Lazer, Alex Pentland, Lada Adamic, Sinan Aral, Albert-László Barabási, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, et al. Computational social science. *Science*, 323(5915):721–723, 2009. doi: 10.1126/science.1167742.
- [226] Sabinne Lee, Changho Hwang, and M. Jae Moon. Policy learning and crisis policy-making: quadruple-loop learning and covid-19 responses in south korea. *Policy*

- and Society*, 39(3):363–381, 2020. doi: 10.1080/14494035.2020.1785195. PMID: 35039726.
- [227] Jan Marco Leimeister. Collective intelligence. *Business & Information Systems Engineering*, 2:245–248, 2010.
- [228] Elisa Leonardelli, Stefano Menini, Alessio Palmero Aprosio, Marco Guerini, and Sara Tonelli. Agreeing to disagree: Annotating offensive language datasets with annotators’ disagreement. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10528–10539, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.822.
- [229] Roger X. Lera-Leri, Enrico Liscio, Filippo Bistaffa, Catholijn M. Jonker, Maite Lopez-Sanchez, Pradeep K. Murukannaiah, Juan A. Rodriguez-Aguilar, and Francisco Salas-Molina. Aggregating value systems for decision support. *Knowledge-Based Systems*, 287:111453, 2024.
- [230] Piyawat Lertvittayakumjorn and Francesca Toni. Explanation-based human debugging of nlp models: A survey. *Transactions of the Association for Computational Linguistics*, 9:1508–1528, 12 2021. ISSN 2307-387X. doi: 10.1162/tacl_a_00440.
- [231] Hao Li, Viktor Schlegel, Riza Batista-Navarro, and Goran Nenadic. Do you hear the people sing? key point analysis via iterative clustering and abstractive summarisation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14064–14080, Toronto, Canada, July 2023. Association for Computational Linguistics.
- [232] Paul Pu Liang, Chiyu Wu, Louis-Philippe Morency, and Ruslan Salakhutdinov. Towards understanding and mitigating social biases in language models. In Marina Meila and Tong Zhang, editors, *International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 6565–6576. PMLR, PMLR, 18–24 Jul 2021.
- [233] Fu-Ren Lin, Lu-Shih Hsieh, and Fu-Tai Chuang. Discovering genres of online discussion threads via text mining. *Computers & Education*, 52(2):481–495, 2009. ISSN 0360-1315. doi: <https://doi.org/10.1016/j.compedu.2008.10.005>.
- [234] Han Lin and Yonghwan Kim. Learning from disagreement on social media: The mediating role of like-minded and cross-cutting discussion and the moderating role of fact-checking. *Computers in Human Behavior*, 139:107558, 2023. ISSN 0747-5632. doi: <https://doi.org/10.1016/j.chb.2022.107558>.
- [235] Enrico Liscio, Michiel van der Meer, Luciano C. Siebert, Catholijn M. Jonker, Niek Mouter, and Pradeep K. Murukannaiah. Axes: Identifying and evaluating context-specific values. In *Proceedings of the 20th Conference on Autonomous Agents and MultiAgent Systems*, AAMAS ’21, pages 799–808, London, 2021.

- [236] Enrico Liscio, Michiel van der Meer, Luciano C. Siebert, Catholijn M. Jonker, Niek Mouter, and Pradeep K. Murukannaiah. Axies: Identifying and evaluating context-specific values. In *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '21, page 799–808, Richland, SC, 2021. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 9781450383073.
- [237] Enrico Liscio, Alin E Dondera, Andrei Geadau, Catholijn M Jonker, and Pradeep K Murukannaiah. Cross-domain classification of moral values. In *Findings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics. NAACL*, volume 22, pages 1–13, 2022.
- [238] Enrico Liscio, Michiel van der Meer, Luciano C. Siebert, Catholijn M. Jonker, and Pradeep K. Murukannaiah. 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, apr 2022. ISSN 1387-2532. doi: 10.1007/s10458-022-09550-0.
- [239] Enrico Liscio, Oscar Araque, Lorenzo Gatti, Ionut Constantinescu, Catholijn Jonker, Kyriaki Kalimeri, and Pradeep Kumar Murukannaiah. What does a Text Classifier Learn about Morality? An Explainable Method for Cross-Domain Comparison of Moral Rhetoric. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ACL '23, pages 14113–14132, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.789.
- [240] Enrico Liscio, Roger Lera-Leri, Filippo Bistaffa, Roel IJ Dobbe, Catholijn M Jonker, Maite Lopez-Sanchez, Juan A Rodriguez-Aguilar, and Pradeep K Murukannaiah. Inferring values via hybrid intelligence. In *HHAI 2023: Augmenting Human Intellect*, pages 373–378. IOS Press, 2023.
- [241] Enrico Liscio, Roger Lera-Leri, Filippo Bistaffa, Roel IJ Dobbe, Catholijn M Jonker, Maite Lopez-Sanchez, Juan A Rodriguez-Aguilar, and Pradeep K Murukannaiah. Value inference in sociotechnical systems. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*, volume 23 of AAMAS '23, pages 1774–1780, Richland, SC, 2023. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 9781450394321.
- [242] Enrico Liscio, Roger Lera-Leri, Filippo Bistaffa, Roel I.J. Dobbe, Catholijn M. Jonker, Maite Lopez-Sanchez, Juan A. Rodriguez-Aguilar, and Pradeep K. Murukannaiah. Value inference in sociotechnical systems. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '23, page 1774–1780, Richland, SC, 2023. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 9781450394321.
- [243] Enrico Liscio, Luciano C. Siebert, Catholijn M. Jonker, and Pradeep K. Murukannaiah. Value preferences estimation and disambiguation in hybrid participatory systems, 2024.

- [244] Bing Liu. *Sentiment Analysis and Opinion Mining*. Springer International Publishing, 2012. ISBN 9783031021459. doi: 10.1007/978-3-031-02145-9.
- [245] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35, 2023.
- [246] Rui Liu, Zheng Lin, Yutong Tan, and Weiping Wang. Enhancing zero-shot and few-shot stance detection with commonsense knowledge graph. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3152–3157, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.278.
- [247] Yang Liu and Mirella Lapata. Text summarization with pretrained encoders. *arXiv preprint arXiv:1908.08345*, 2019. doi: 10.18653/v1/d19-1387.
- [248] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 1(1), 2019.
- [249] Ines Lörcher and Monika Taddicken. Discussing climate change online. topics and perceptions in online climate change communication in different online public arenas. *Journal of Science Communication*, 16(2):A03, 2017.
- [250] Jan Lorenz, Heiko Rauhut, Frank Schweitzer, and Dirk Helbing. How social influence can undermine the wisdom of crowd effect. *Proceedings of the national academy of sciences*, 108(22):9020–9025, 2011. doi: 10.1073/pnas.1008636108.
- [251] Philipp Lorenz-Spreen, Lisa Oswald, Stephan Lewandowsky, and Ralph Hertwig. A systematic review of worldwide causal and correlational evidence on digital media and democracy. *Nature human behaviour*, 7(1):74–101, 2023.
- [252] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2018.
- [253] Nicholas Lourie, Ronan Le Bras, and Yejin Choi. Scruples: A corpus of community ethical judgments on 32,000 real-life anecdotes. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15):13470–13479, May 2021. doi: 10.1609/aaai.v35i15.17589.
- [254] Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8086–8098, 2022.
- [255] Yun Luo, Zihan Liu, Stan Z. Li, and Yue Zhang. Improving (dis)agreement detection with inductive social relation information from comment-reply interactions. In *Proceedings of the ACM Web Conference 2023, WWW ’23*, page 1584–1593, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450394161. doi: 10.1145/3543507.3583314.

- [256] Veronica Lynn, Youngseo Son, Vivek Kulkarni, Niranjan Balasubramanian, and H Andrew Schwartz. Human centered nlp with user-factor adaptation. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel, editors, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1146–1155, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1119.
- [257] Qingsong Ma, Johnny Wei, Ondřej Bojar, and Yvette Graham. Results of the WMT19 metrics shared task: Segment-level and strong MT systems pose big challenges. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 62–90, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-5302.
- [258] Noni E. MacDonald. Vaccine hesitancy: Definition, scope and determinants. *Vaccine*, 33(34):4161–4164, 2015. ISSN 0264-410X. doi: <https://doi.org/10.1016/j.vaccine.2015.04.036>. WHO Recommendations Regarding Vaccine Hesitancy.
- [259] David Madras, Toni Pitassi, and Richard Zemel. Predict responsibly: improving fairness and accuracy by learning to defer. *Advances in Neural Information Processing Systems*, 31, 2018.
- [260] Jane Mansbridge, James Bohman, Simone Chambers, Thomas Christiano, Archon Fung, John Parkinson, Dennis F Thompson, and Mark E Warren. A systemic approach to deliberative democracy. *Deliberative systems: Deliberative democracy at the large scale*, pages 1–26, 2012. doi: 10.1017/cbo9781139178914.002.
- [261] Katerina Margatina and Nikolaos Aletras. On the limitations of simulating active learning. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4402–4419, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.269.
- [262] Tom McCoy, Ellie Pavlick, and Tal Linzen. 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, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1334.
- [263] Spencer McKay and Chris Tenove. Disinformation as a threat to deliberative democracy. *Political research quarterly*, 74(3):703–717, 2021.
- [264] Duncan McLaren, Karen A Parkhill, Adam Corner, Naomi E Vaughan, and Nicholas F Pidgeon. Public conceptions of justice in climate engineering: Evidence from secondary analysis of public deliberation. *Global Environmental Change*, 41:64–73, 2016.
- [265] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys*, 54(6), jul 2021. ISSN 0360-0300. doi: 10.1145/3457607.

- [266] Bonan Min, Hayley Ross, Elinor Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz, and Dan Roth. Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Computing Surveys*, 56(2):1–40, sep 2023. ISSN 0360-0300. doi: 10.1145/3605943.
- [267] Amita Misra, Pranav Anand, Jean E Fox Tree, and Marilyn Walker. Using summarization to discover argument facets in online ideological dialog. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 430–440, 2015.
- [268] Mirjam Moerbeek. Bayesian updating: increasing sample size during the course of a study. *BMC Medical Research Methodology*, 21(1):137, Jul 2021. ISSN 1471-2288. doi: 10.1186/s12874-021-01334-6.
- [269] Negar Mokherian, Myrl Marmarelis, Frederic Hopp, Valerio Basile, Fred Morstatter, and Kristina Lerman. Capturing perspectives of crowdsourced annotators in subjective learning tasks. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7337–7349, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.407.
- [270] Ines Montani and Matthew Honnibal. Prodigy: A modern and scriptable annotation tool for creating training data for machine learning models, 2022.
- [271] Marlon Mooijman, Joe Hoover, Ying Lin, Heng Ji, and Morteza Dehghani. Moralization in social networks and the emergence of violence during protests. *Nature human behaviour*, 2(6):389–396, 2018. doi: 10.1038/s41562-018-0353-0.
- [272] Roser Morante, Chantal Van Son, Isa Maks, and Piek Vossen. Annotating perspectives on vaccination. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4964–4973, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4.
- [273] Eduardo Mosqueira-Rey, Elena Hernández-Pereira, David Alonso-Ríos, José Bobes-Bascarán, and Ángel Fernández-Leal. Human-in-the-loop machine learning: a state of the art. *Artificial Intelligence Review*, 56(4):3005–3054, 2023.
- [274] Niek Mouder, Jose Ignacio Hernandez, and Anatol Valerian Itten. 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, 05 2021. doi: 10.1371/journal.pone.0250614.
- [275] Niek Mouder, Paul Koster, and Thijs Dekker. Contrasting the recommendations of participatory value evaluation and cost-benefit analysis in the context of urban mobility investments. *Transportation research part A: policy and practice*, 144:54–73, 2021.

- [276] Niek Mouter, Ruth M Shortall, Shannon L Spruit, and Anatol V Itten. Including young people, cutting time and producing useful outcomes: Participatory value evaluation as a new practice of public participation in the dutch energy transition. *Energy Research & Social Science*, 75:101965, 2021.
- [277] Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. When does label smoothing help? *Advances in neural information processing systems*, 32, 2019.
- [278] Eni Mustafaraj, Samantha Finn, Carolyn Whitlock, and Panagiotis T Metaxas. Vocal minority versus silent majority: Discovering the opinions of the long tail. In *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*, pages 103–110. IEEE, 2011. doi: 10.1109/PASSAT/SocialCom.2011.188.
- [279] Yoon-Eui Nahm. A novel approach to prioritize customer requirements in qfd based on customer satisfaction function for customer-oriented product design. *Journal of Mechanical Science and Technology*, 27:3765–3777, 2013.
- [280] United Nations. Transforming our world: the 2030 agenda for sustainable development. <https://sdgs.un.org/2030agenda>, 2015.
- [281] Ani Nenkova, Kathleen McKeown, et al. Automatic summarization. *Foundations and Trends® in Information Retrieval*, 5(2–3):103–233, 2011.
- [282] German Neubaum and Nicole C Krämer. Monitoring the opinion of the crowd: Psychological mechanisms underlying public opinion perceptions on social media. *Media psychology*, 20(3):502–531, 2017. doi: 10.1080/15213269.2016.1211539.
- [283] M. E. J. Newman, D. J. Watts, and S. H. Strogatz. Random graph models of social networks. *Proceedings of the National Academy of Sciences*, 99(suppl_1):2566–2572, 2002. doi: 10.1073/pnas.012582999.
- [284] Vlad Niculae and Cristian Danescu-Niculescu-Mizil. Conversational markers of constructive discussions. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 568–578, 2016.
- [285] Vlad Niculae, Joonsuk Park, and Claire Cardie. 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, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1091.
- [286] Yixin Nie, Xiang Zhou, and Mohit Bansal. What can we learn from collective human opinions on natural language inference data? In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9131–9143, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.734.

- [287] Sem Nouws, Íñigo Martínez De Rituerto De Troya, Roel Dobbe, and Marijn Janssen. Diagnosing and addressing emergent harms in the design process of public ai and algorithmic systems. In *Proceedings of the 24th Annual International Conference on Digital Government Research*, pages 679–681, 2023.
- [288] Humphrey O. Obie, Waqar Hussain, Xin Xia, John Grundy, Li Li, Burak Turhan, Jon Whittle, and Mojtaba Shahin. A first look at human values-violation in app reviews. In *2021 IEEE/ACM 43rd International Conference on Software Engineering: Software Engineering in Society (ICSE-SEIS)*, pages 29–38, 2021. doi: 10.1109/ICSE-SEIS52602.2021.00012.
- [289] OECD. *Innovative Citizen Participation and New Democratic Institutions*. OECD, 2020. doi: <https://doi.org/https://doi.org/10.1787/339306da-en>.
- [290] OpenAI. The OpenAI Python library. <https://github.com/openai/openai-python>, 2023. Accessed: October 2023.
- [291] Matthias Orlikowski, Paul Röttger, Philipp Cimiano, and Dirk Hovy. The Ecological Fallacy in Annotation: Modeling Human Label Variation goes beyond Sociodemographics. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics Volume 2: Short Papers*, pages 1017–1029. ACL, 2023. ISBN 9781959429715. doi: 10.18653/v1/2023.acl-short.88.
- [292] Elinor Ostrom. *Governing the commons: The evolution of institutions for collective action*. Cambridge University Press, Cambridge, UK, 1990.
- [293] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.
- [294] Siru Ouyang, Shuohang Wang, Yang Liu, Ming Zhong, Yizhu Jiao, Dan Iter, Reid Pryzant, Chenguang Zhu, Heng Ji, and Jiawei Han. The shifted and the overlooked: A task-oriented investigation of user-gpt interactions. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2375–2393, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.146.
- [295] Maria Leonor Pacheco, Tunazzina Islam, Lyle Ungar, Ming Yin, and Dan Goldwasser. Interactive concept learning for uncovering latent themes in large text collections. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5059–5080, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.313.
- [296] Raquel Mochales Palau and Marie-Francine Moens. Argumentation mining: the detection, classification and structure of arguments in text. In *Proceedings of the 12th international conference on artificial intelligence and law*, pages 98–107, 2009.

- [297] Zizi Papacharissi. The virtual sphere: The internet as a public sphere. *New media & society*, 4(1):9–27, 2002.
- [298] John Parkinson. Legitimacy problems in deliberative democracy. *Political studies*, 51(1):180–196, 2003.
- [299] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12(85):2825–2830, 2011.
- [300] Joshua C Peterson, Ruairidh M Battleday, Thomas L Griffiths, and Olga Russakovsky. Human uncertainty makes classification more robust. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9617–9626, 2019.
- [301] Hoang Phan, Long Nguyen, and Khanh Doan. Matching the statements: A simple and accurate model for key point analysis. In *Proceedings of the 8th Workshop on Argument Mining*, pages 165–174, 2021.
- [302] Barbara Plank. The “problem” of human label variation: On ground truth in data, modeling and evaluation. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.731.
- [303] Francesca Polletta and Beth Gardner. The Forms of Deliberative Communication. In *The Oxford Handbook of Deliberative Democracy*. Oxford University Press, 09 2018. ISBN 9780198747369. doi: 10.1093/oxfordhb/9780198747369.013.45.
- [304] Vladimir Ponizovskiy, Murat Ardag, Lusine Grigoryan, Ryan Boyd, Henrik Döbwall, and Peter Holtz. Development and validation of the personal values dictionary: A theory-driven tool for investigating references to basic human values in text. *European Journal of Personality*, 34(5):885–902, 2020. doi: 10.1002/per.2294.
- [305] John Pougué-Biyong, Valentina Semenova, Alexandre Matton, Rachel Han, Aerin Kim, Renaud Lambiotte, and Doyne Farmer. Disagreement: A comment-reply dataset for (dis) agreement detection in online debates. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.
- [306] Katrin Preckel, Philipp Kanske, and Tania Singer. On the interaction of social affect and cognition: empathy, compassion and theory of mind. *Current Opinion in Behavioral Sciences*, 19:1–6, 2018.
- [307] Vincent Price. Social Identification and Public Opinion: Effects of Communicating Group Conflict. *Public Opinion Quarterly*, 53(2):197–224, 01 1989. ISSN 0033-362X. doi: 10.1086/269503.

- [308] Nicholas Proferes, Naiyan Jones, Sarah Gilbert, Casey Fiesler, and Michael Zimmer. Studying Reddit: A systematic overview of disciplines, approaches, methods, and ethics. *Social Media + Society*, 7(2):20563051211019004, 2021. doi: 10.1177/20563051211019004.
- [309] Yada Pruksachatkun, Jason Phang, Haokun Liu, Phu Mon Htut, Xiaoyi Zhang, Richard Yuanzhe Pang, Clara Vania, Katharina Kann, and Samuel R. Bowman. Intermediate-task transfer learning with pretrained language models: When and why does it work? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5231–5247, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.467.
- [310] James Pustejovsky, José M Castano, Robert Ingria, Roser Sauri, Robert J Gaizauskas, Andrea Setzer, Graham Katz, and Dragomir R Radev. Timeml: Robust specification of event and temporal expressions in text. *New directions in question answering*, 3: 28–34, 2003.
- [311] Liang Qiu, Yizhou Zhao, Jinchao Li, Pan Lu, Baolin Peng, Jianfeng Gao, and Song-Chun Zhu. Valuenet: A new dataset for human value driven dialogue system. *arXiv preprint arXiv:2112.06346*, 2021.
- [312] Alexander J Quinn and Benjamin B Bederson. Human computation: a survey and taxonomy of a growing field. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 1403–1412, 2011.
- [313] John Rawls. *A Theory of Justice*. Oxford University Press, Oxford, 1973.
- [314] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1410.
- [315] Nils Reimers, Benjamin Schiller, Tilman Beck, Johannes Daxenberger, Christian Stab, and Iryna Gurevych. 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, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1054.
- [316] Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po Yao Huang, Zhihui Li, Brij B. Gupta, Xiaojiang Chen, and Xin Wang. A Survey of Deep Active Learning. *ACM Computing Surveys*, 54(9):1–40, 2021. doi: 10.1145/3472291.
- [317] Bram M Renting, Holger H Hoos, and Catholijn M Jonker. Automated configuration and usage of strategy portfolios for mixed-motive bargaining. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems, AAMAS ’22*, pages 1101–1109, Richland, SC, 2022. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 9781450392136.

- [318] Myrthe Reuver, Nicolas Mattis, Marijn Sax, Suzan Verberne, Nava Tintarev, Natali Helberger, Judith Moeller, Sanne Vrijenhoek, Antske Fokkens, and Wouter van Atteveldt. Are we human, or are we users? the role of natural language processing in human-centric news recommenders that nudge users to diverse content. In *Proceedings of the 1st Workshop on NLP for Positive Impact*, pages 47–59, 2021.
- [319] Myrthe Reuver, Suzan Verberne, Roser Morante, and Antske Fokkens. Is stance detection topic-independent and cross-topic generalizable?-a reproduction study. In *Proceedings of the 8th Workshop on Argument Mining*, pages 46–56, 2021.
- [320] Horst WJ Rittel and Melvin M Webber. Dilemmas in a general theory of planning. *Policy sciences*, 4(2):155–169, 1973.
- [321] Milton Rokeach. Rokeach value survey. *The nature of human values.*, 1967. doi: 10.1037/t01381-000.
- [322] Julia Romberg. Is your perspective also my perspective? enriching prediction with subjectivity. In *Proceedings of the 9th Workshop on Argument Mining*, pages 115–125, 2022.
- [323] Andrew Rosenberg and Julia Hirschberg. V-measure: A conditional entropy-based external cluster evaluation measure. In Jason Eisner, editor, *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 410–420, Prague, Czech Republic, June 2007. Association for Computational Linguistics.
- [324] Jeffrey N. Rouder, Paul L. Speckman, Dongchu Sun, Richard D. Morey, and Geoffrey Iverson. Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, 16(2):225–237, Apr 2009. ISSN 1531-5320. doi: 10.3758/PBR.16.2.225.
- [325] Gene Rowe and Lynn J Frewer. Public participation methods: a framework for evaluation. *Science, technology, & human values*, 25(1):3–29, 2000.
- [326] Shamik Roy, Maria Leonor Pacheco, and Dan Goldwasser. Identifying morality frames in political tweets using relational learning. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9939–9958, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.783.
- [327] Ramon Ruiz-Dolz, Jose Alemany, Stella M Heras Barberá, and Ana García-Fornes. Transformer-based models for automatic identification of argument relations: A cross-domain evaluation. *IEEE Intelligent Systems*, 36(6):62–70, 2021.
- [328] Pratik Sachdeva, Renata Barreto, Geoff Bacon, Alexander Sahn, Claudia von Vacano, and Chris Kennedy. The measuring hate speech corpus: Leveraging rasch measurement theory for data perspectivism. In Gavin Abercrombie, Valerio Basile, Sara Tonelli, Verena Rieser, and Alexandra Uma, editors, *Proceedings of the 1st Workshop*

- on Perspectivist Approaches to NLP @LREC2022*, pages 83–94, Marseille, France, June 2022. European Language Resources Association.
- [329] Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. “everyone wants to do the model work, not the data work”: Data cascades in high-stakes ai. In *proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–15, 2021.
 - [330] Ahnaf Mozib Samin, Behrooz Nikandish, and Jingyan Chen. Arguments to key points mapping with prompt-based learning. *ICNLSP 2022*, page 303, 2022.
 - [331] Selene Baez Santamaría, Piek Vossen, and Thomas Baier. Evaluating agent interactions through episodic knowledge graphs. In Heuseok Lim, Seungryong Kim, Yeonsoo Lee, Steve Lin, Paul Hongsuck Seo, Yumin Suh, Yoonna Jang, Jungwoo Lim, Yuna Hur, and Suhyune Son, editors, *Proceedings of the 1st Workshop on Customized Chat Grounding Persona and Knowledge@ COLING2022*, pages 15–28, Gyeongju, Republic of Korea, October 2022. Association for Computational Linguistics (ACL), Association for Computational Linguistics.
 - [332] Filippo Santoni de Sio and Jeroen Van den Hoven. Meaningful human control over autonomous systems: A philosophical account. *Frontiers in Robotics and AI*, 5:15, 2018.
 - [333] Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. Whose opinions do language models reflect? *arXiv preprint arXiv:2303.17548*, 202:29971–30004, 23–29 Jul 2023.
 - [334] Sebastin Santy, Jenny Liang, Ronan Le Bras, Katharina Reinecke, and Maarten Sap. NLPositionality: Characterizing design biases of datasets and models. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9080–9102, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.505.
 - [335] Martin Saveski, Nabeel Gillani, Ann Yuan, Prashanth Vijayaraghavan, and Deb Roy. Perspective-taking to reduce affective polarization on social media. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, pages 885–895, 2022.
 - [336] Akрати Saxena and Harita Reddy. Users roles identification on online crowdsourced q&a platforms and encyclopedias: a survey. *Journal of Computational Social Science*, 5(1):285–317, 2022. doi: 10.1007/s42001-021-00125-9.
 - [337] Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. Are emergent abilities of large language models a mirage? *Advances in Neural Information Processing Systems*, 36: 55565–55581, 2024.
 - [338] Mike Schaekermann, Joslin Goh, Kate Larson, and Edith Law. Resolvable vs. irresolvable disagreement: A study on worker deliberation in crowd work. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–19, 2018.

- [339] Patrick Schober, Christa Boer, and Lothar A Schwarte. Correlation coefficients: appropriate use and interpretation. *Anesthesia & analgesia*, 126(5):1763–1768, 2018.
- [340] Mirco Schönfeld, Steffen Eckhard, Ronny Patz, and Hilde van Meegdenburg. The un security council debates 1995-2017. *arXiv preprint arXiv:1906.10969*, 2019.
- [341] Claudia Schulz, Christian M. Meyer, Jan Kiesewetter, Michael Sailer, Elisabeth Bauer, Martin R. Fischer, Frank Fischer, and Iryna Gurevych. 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, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1265.
- [342] Stefan Schulz-Hardt, Dieter Frey, Carsten Lüthgens, and Serge Moscovici. Biased information search in group decision making. *Journal of Personality and Social Psychology*, 78(4):655–669, 2000. doi: 10.1037/0022-3514.78.4.655.
- [343] Shalom Schwartz. A repository of schwartz value scales with instructions and an introduction. *Online Readings in Psychology and Culture*, 2, 09 2021. doi: 10.9707/2307-0919.1173.
- [344] Shalom H Schwartz. An overview of the schwartz theory of basic values. *Online readings in Psychology and Culture*, 2(1):11, 2012. doi: 10.9707/2307-0919.1116.
- [345] Shalom H Schwartz, Jan Cieciuch, Michele Vecchione, Eldad Davidov, Ronald Fischer, Constanze Beierlein, Alice Ramos, Markku Verkasalo, Jan-Erik Lönnqvist, Kursad Demirutku, et al. Refining the theory of basic individual values. *Journal of personality and social psychology*, 103(4):663, 2012. doi: <https://doi.org/10.1037/a0029393>.
- [346] Abigail See, Peter J. Liu, and Christopher D. Manning. Get to the point: Summarization with pointer-generator networks. In Regina Barzilay and Min-Yen Kan, editors, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1099.
- [347] Thibault Sellam, Dipanjan Das, and Ankur Parikh. Bleurt: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, 2020.
- [348] Bryan C Semaan, Scott P Robertson, Sara Douglas, and Misa Maruyama. Social media supporting political deliberation across multiple public spheres: towards depolarization. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, pages 1409–1421, 2014.
- [349] Rico Sennrich, Barry Haddow, and Alexandra Birch. 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, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1009.

- [350] Burr Settles. *Active Learning*. Morgan & Claypool, 2012.
- [351] Eyal Shnarch, Ariel Gera, Alon Halfon, Lena Dankin, Leshem Choshen, Ranit Aharonov, and Noam Slonim. Cluster & tune: Boost cold start performance in text classification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7639–7653, 2022.
- [352] Ruth Shortall, Anatol Itten, Michiel van der Meer, Pradeep K. Murukannaiah, and Catholijn M. Jonker. Reason against the machine? future directions for mass online deliberation. *Frontiers in Political Science*, 4:01–17, 2022. ISSN 2673-3145. doi: 10.3389/fpos.2022.946589.
- [353] Patrick E Shrout and Joseph L Fleiss. Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin*, 86(2):420, 1979.
- [354] Luciano C Siebert, Enrico Liscio, Pradeep K Murukannaiah, Lionel Kaptein, Shannon Spruit, Jeroen van den Hoven, and Catholijn Jonker. Estimating value preferences in a hybrid participatory system. In *Proceedings of the first International Conference on Hybrid Human-Artificial Intelligence (HHAI 2022)*, pages 1–14, Amsterdam, the Netherlands, 2022. IOS Press.
- [355] Tejpsalsingh Siledar, Jigar Makwana, and Pushpak Bhattacharyya. Aspect-sentiment-based opinion summarization using multiple information sources. In *Proceedings of the 6th Joint International Conference on Data Science & Management of Data (10th ACM IKDD CODS and 28th COMAD)*, pages 55–61, 2023.
- [356] Amila Silva, Pei-Chi Lo, and Ee Peng Lim. On predicting personal values of social media users using community-specific language features and personal value correlation. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, pages 680–690, 2021.
- [357] Julius Sim and Chris C Wright. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. *Physical therapy*, 85(3):257–268, 2005.
- [358] Bernd Skiera, Shunyao Yan, Johannes Daxenberger, Marcus Dombois, and Iryna Gurevych. Using information-seeking argument mining to improve service. *Journal of Service Research*, 25(4):537–548, 2022. doi: 10.1177/10946705221110845.
- [359] Gabriella Skitalinskaya, Jonas Klaff, and Henning Wachsmuth. Learning from revisions: Quality assessment of claims in argumentation at scale. In Paola Merlo, Jorg Tiedemann, and Reut Tsarfay, editors, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1718–1729, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.147.
- [360] Graham Smith, Robert C Richards Jr, and John Gastil. The potential of participedia as a crowdsourcing tool for comparative analysis of democratic innovations. *Policy & Internet*, 7(2):243–262, 2015.

- [361] Anders Søgaard, Sebastian Ebert, Jasmijn Bastings, and Katja Filippova. We need to talk about random splits. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1823–1832, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.156.
- [362] Abdurashid Solijonov. International idea: Voter turnout trends around the world. <https://www.idea.int/publications/catalogue/voter-turnout-trends-around-world>, 2016. Accessed: 2024-02-19.
- [363] Hyunjin Song, Jaeho Cho, and Grace A Benefield. The dynamics of message selection in online political discussion forums: Self-segregation or diverse exposure? *Communication Research*, 47(1):125–152, 2020. doi: 10.1177/0093650218790144.
- [364] Taylor Sorensen, Joshua Robinson, Christopher Michael Rytting, Alexander Glenn Shaw, Kyle Jeffrey Rogers, Alexia Pauline Delorey, Mahmoud Khalil, Nancy Fulda, and David Wingate. An information-theoretic approach to prompt engineering without ground truth labels, 2022.
- [365] Christian Stab and Iryna Gurevych. 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, Dublin, Ireland, August 2014. Dublin City University and Association for Computational Linguistics.
- [366] Christian Stab, Johannes Daxenberger, Chris Stahlhut, Tristan Miller, Benjamin Schiller, Christopher Tauchmann, Steffen Eger, and Iryna Gurevych. ArgumenText: Searching for arguments in heterogeneous sources. In Yang Liu, Tim Paek, and Manasi Patwardhan, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 21–25, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-5005.
- [367] Christian Stab, Tristan Miller, Benjamin Schiller, Pranav Rai, and Iryna Gurevych. Cross-topic argument mining from heterogeneous sources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3664–3674, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1402.
- [368] Marco R Steenbergen, André Bächtiger, Markus Spörndli, and Jürg Steiner. Measuring political deliberation: A discourse quality index. *Comparative European Politics*, 1(1): 21–48, 2003. doi: 10.1057/palgrave.cep.6110002.
- [369] Cor Steging, Silja Renooij, and Bart Verheij. Discovering the rationale of decisions: towards a method for aligning learning and reasoning. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*, pages 235–239, 2021.
- [370] Kim Strandberg and Kimmo Grönlund. Online deliberation. *The Oxford handbook of deliberative democracy*, pages 365–377, 2018.

- [371] Jennifer Stromer-Galley and Peter Muhlberger. Agreement and disagreement in group deliberation: Effects on deliberation satisfaction, future engagement, and decision legitimacy. *Political Communication*, 26(2):173–192, 2009. doi: 10.1080/10584600902850775.
- [372] Jennifer Stromer-Galley, Lauren Bryant, and Bruce Bimber. Context and medium matter: Expressing disagreements online and face-to-face in political deliberations. *Journal of Deliberative Democracy*, 11(1), 2020. doi: <https://doi.org/10.16997/jdd.218>.
- [373] Yoshihiko Suhara, Xiaolan Wang, Stefanos Angelidis, and Wang-Chiew Tan. Opiniondigest: A simple framework for opinion summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5789–5798, 2020.
- [374] Shiliang Sun, Chen Luo, and Junyu Chen. A review of natural language processing techniques for opinion mining systems. *Information fusion*, 36:10–25, 2017. ISSN 1566-2535. doi: <https://doi.org/10.1016/j.inffus.2016.10.004>.
- [375] James Surowiecki. *The wisdom of crowds*. Anchor, 2005.
- [376] Shahbaz Syed, Khalid Al Khatib, Milad Alshomary, Henning Wachsmuth, and Martin Potthast. Generating informative conclusions for argumentative texts. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3482–3493, 2021.
- [377] Shahbaz Syed, Dominik Schwabe, Khalid Al Khatib, and Martin Potthast. Indicative summarization of long discussions. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2752–2788, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.166.
- [378] Zeerak Talat, Hagen Blix, Josef Valvoda, Maya Indira Ganesh, Ryan Cotterell, and Adina Williams. 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*, pages 769–779, Seattle, USA, 2022. doi: 10.18653/v1/2022.naacl-main.56.
- [379] Alex Tamkin, Miles Brundage, Jack Clark, and Deep Ganguli. Understanding the capabilities, limitations, and societal impact of large language models. *arXiv preprint arXiv:2102.02503*, 2021.
- [380] Dapeng Tao, Jun Cheng, Zhengtao Yu, Kun Yue, and Lizhen Wang. Domain-weighted majority voting for crowdsourcing. *IEEE transactions on neural networks and learning systems*, 30(1):163–174, 2018.
- [381] Milagro Teruel, Cristian Cardellino, Fernando Cardellino, Laura Alonso Alemany, and Serena Villata. 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)*, pages 4061–4064, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA).

- [382] Terne Sasha Thorn Jakobsen, Maria Barrett, and Anders Søgaard. 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, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.starsem-1.25.
- [383] Ilaria Tiddi, Victor de Boer, Stefan Schlobach, and André Meyer-Vitali. Knowledge engineering for hybrid intelligence. In *Proceedings of the 12th Knowledge Capture Conference 2023, K-CAP '23*, page 75–82, Pensacola, FL, USA., 2023. Association for Computing Machinery. ISBN 9798400701412. doi: 10.1145/3587259.3627541.
- [384] Nenad Tomašev, Julien Cornebise, Frank Hutter, Shakir Mohamed, Angela Picciariello, Bec Connelly, Danielle CM Belgrave, Daphne Ezer, Fanny Cachat van der Haert, Frank Mugisha, et al. Ai for social good: unlocking the opportunity for positive impact. *Nature Communications*, 11(1):2468, 2020. doi: 10.1038/s41467-020-15871-z.
- [385] Petter Törnberg. Chatgpt-4 outperforms experts and crowd workers in annotating political twitter messages with zero-shot learning. *arXiv preprint arXiv:2304.06588*, 2023.
- [386] Stephen E Toulmin. *The uses of argument*. Cambridge university press, 2003. ISBN 9780521534833.
- [387] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 0(0), 2023.
- [388] Jackson Trager, Alireza S Ziabari, Aida Mostafazadeh Davani, Prenti Golazanian, Farzan Karimi-Malekabadi, Ali Omrani, Zhihe Li, Brendan Kennedy, Nils Karl Reimer, Melissa Reyes, et al. The moral foundations reddit corpus. *arXiv preprint arXiv:2208.05545*, 2022.
- [389] Nhat Tran and Diane Litman. Multi-task learning in argument mining for persuasive online discussions. In *Proceedings of the 8th Workshop on Argument Mining*, pages 148–153, 2021.
- [390] Matthias Trénel. Facilitation and inclusive deliberation. *Online deliberation: Design, research, and practice*, pages 253–257, 2009.
- [391] Roman Trötschel, Joachim Hüffmeier, David D Loschelder, Katja Schwartz, and Peter M Gollwitzer. Perspective taking as a means to overcome motivational barriers in negotiations: When putting oneself into the opponent's shoes helps to walk toward agreements. *Journal of personality and social psychology*, 101(4):771, 2011.
- [392] Petter Törnberg. How digital media drive affective polarization through partisan sorting. *Proceedings of the National Academy of Sciences*, 119(42):e2207159119, 2022. doi: 10.1073/pnas.2207159119.

- [393] Alexandra N Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. Learning from disagreement: A survey. *Journal of Artificial Intelligence Research*, 72:1385–1470, jan 2021. ISSN 1076-9757. doi: 10.1613/jair.1.12752.
- [394] Steven Umbrello and Ibo Van de Poel. Mapping value sensitive design onto ai for social good principles. *AI and Ethics*, 1(3):283–296, 2021.
- [395] Raphael Vallat. Pingouin: statistics in python. *Journal of Open Source Software*, 3(31): 1026, 2018. doi: 10.21105/joss.01026.
- [396] Guido M van de Ven, Tinne Tuytelaars, and Andreas S Tolias. Three types of incremental learning. *Nature Machine Intelligence*, 4(12):1185–1197, 2022. doi: 10.1038/s42256-022-00568-3.
- [397] Michiel van der Meer, Enrico Liscio, Catholijn M. Jonker, Aske Plaat, Piek Vossen, and Pradeep K. Murukannaiah. HyEnA: A hybrid method for extracting arguments from opinions. In *HHA12022: Augmenting Human Intellect*, pages 17–31. IOS Press, 2022.
- [398] Michiel van der Meer, Enrico Liscio, Catholijn M. Jonker, Aske Plaat, Piek Vossen, and Pradeep K. Murukannaiah. Hyena: A hybrid method for extracting arguments from opinions. In *Proceedings of the first International Conference on Hybrid Human-Artificial Intelligence (HHA1 2022)*, pages 17–31, Amsterdam, the Netherlands, 2022. IOS Press.
- [399] Michiel van der Meer, Myrthe Reuver, Urja Khurana, Lea Krause, and Selene Baez Santamaria. Will it blend? Mixing training paradigms & prompting for argument quality prediction. In *Proceedings of the 9th Workshop on Argument Mining*, pages 95–103, Online and in Gyeongju, Republic of Korea, 10 2022. International Conference on Computational Linguistics.
- [400] Michiel van der Meer, Piek Vossen, Catholijn M. Jonker, and Pradeep K. Murukannaiah. Do differences in values influence disagreements in online discussions? In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15986–16008, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.992.
- [401] Michiel van der Meer, Piek Vossen, Catholijn M. Jonker, and Pradeep K. Murukannaiah. Do differences in values influence disagreements in online discussions? Supplementary material, Dec 2023.
- [402] Michiel Van Der Meer, Neele Falk, Pradeep K. Murukannaiah, and Enrico Liscio. Annotator-centric active learning for subjective NLP tasks. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18537–18555, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

- [403] Michiel van der Meer, Enrico Liscio, Catholijn M. Jonker, Aske Plaat, Piek Vossen, and Pradeep K. Murukannaiah. A hybrid intelligence method for argument mining. *Journal of Artificial Intelligence Research*, 80:1187–1222, 2024. doi: <https://doi.org/10.1613/jair.1.15135>.
- [404] Michiel van der Meer, Enrico Liscio, Catholijn M. Jonker, Aske Plaat, Piek Vossen, and Pradeep K. Murukannaiah. A hybrid intelligence method for argument mining: Supplementary material, 2024.
- [405] Michiel van der Meer, Piek Vossen, Catholijn M. Jonker, and Pradeep K. Murukannaiah. An empirical analysis of diversity in argument summarization. In Yvette Graham and Matthew Purver, editors, *(To appear) Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2028–2045, St. Julian's, Malta, March 2024. Association for Computational Linguistics.
- [406] Max van Duijn, Bram van Dijk, Tom Kouwenhoven, Werner de Valk, Marco Spruit, and Peter van der Putten. Theory of mind in large language models: Examining performance of 11 state-of-the-art models vs. children aged 7-10 on advanced tests. In Jing Jiang, David Reitter, and Shumin Deng, editors, *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 389–402, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.conll-1.25.
- [407] Frans H. van Eemeren, Rob Grootendorst, and Tjark Krugier. *Handbook of Argumentation Theory*. De Gruyter Mouton, Berlin, Boston, 1987. ISBN 9783110846096. doi: 10.1515/9783110846096.
- [408] Frans H Van Eemeren, Frans H van Eemeren, Sally Jackson, and Scott Jacobs. Argumentation. *Reasonableness and effectiveness in argumentative discourse: Fifty contributions to the development of Pragma-dialectics*, pages 3–25, 2015.
- [409] Daan Van Knippenberg, Carsten KW De Dreu, and Astrid C Homan. Work group diversity and group performance: an integrative model and research agenda. *Journal of applied psychology*, 89(6):1008, 2004.
- [410] Don van Ravenzwaaij and Eric-Jan Wagenmakers. Advantages masquerading as “issues” in bayesian hypothesis testing: A commentary on tendeiro and kiers (2019). *Psychological Methods*, 27(3):451–465, June 2022. doi: 10.1037/met0000415.
- [411] Chantal van Son, Tommaso Caselli, Antske Fokkens, Isa Maks, Roser Morante, Lora Aroyo, and Piek Vossen. GRASP: A multilayered annotation scheme for perspectives. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1177–1184, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA).
- [412] Chantal van Son, Tommaso Caselli, Antske Fokkens, Isa Maks, Roser Morante, Lora Aroyo, and Piek Vossen. GRASP: A multilayered annotation scheme for perspectives. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk,

- and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1177–1184, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA).
- [413] Emma Van Zoelen, Tina Mioch, Mani Tajaddini, Christian Fleiner, Stefani Tsaneva, Pietro Camin, Thiago S. Gouvêa, Kim Baraka, Maaïke H. T. De Boer, and Mark A. Neerincx. Developing team design patterns for hybrid intelligence systems. In *Proceedings of the second International Conference on Hybrid Human-Artificial Intelligence (HHAI 2023)*, pages 3–16. IOS Press, 2023. doi: 10.3233/faia230071.
- [414] Eva Maria Vecchi, Neele Falk, Iman Jundi, and Gabriella Lapesa. Towards argument mining for social good: A survey. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1338–1352, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.107.
- [415] Jacky Visser, Barbara Konat, Rory Duthie, Marcin Koszowy, Katarzyna Budzyska, and Chris Reed. Argumentation in the 2016 us presidential elections: annotated corpora of television debates and social media reaction. *Language Resources and Evaluation*, 54(1):123–154, 2020.
- [416] Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein. Computational argumentation quality assessment in natural language. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 176–187, Valencia, Spain, April 2017. Association for Computational Linguistics.
- [417] Henning Wachsmuth, Shahbaz Syed, and Benno Stein. Retrieval of the best counter-argument without prior topic knowledge. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 241–251, 2018.
- [418] Douglas Walton, Christopher Reed, and Fabrizio Macagno. *Argumentation schemes*. Cambridge University Press, 2008.
- [419] Jindong Wang, Xixu Hu, Wenxin Hou, Hao Chen, Runkai Zheng, Yidong Wang, Linyi Yang, Haojun Huang, Wei Ye, Xiubo Geng, et al. On the robustness of chatgpt: An adversarial and out-of-distribution perspective. *arXiv preprint arXiv:2302.12095*, 2023: 1–19, 2 2023.
- [420] Lu Wang and Wang Ling. Neural network-based abstract generation for opinions and arguments. In Kevin Knight, Ani Nenkova, and Owen Rambow, editors, *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 47–57, San Diego, California, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-1007.

- [421] Xinpeng Wang and Barbara Plank. Actor: Active learning with annotator-specific classification heads to embrace human label variation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2046–2052, 2023.
- [422] Yizhong Wang, Hamish Ivison, Pradeep Dasigi, Jack Hessel, Tushar Khot, Khyathi Chandu, David Wadden, Kelsey MacMillan, Noah A Smith, Iz Beltagy, et al. How far can camels go? exploring the state of instruction tuning on open resources. *Advances in Neural Information Processing Systems*, 36:74764–74786, 2024.
- [423] Zijie J. Wang, Dongjin Choi, Shenyu Xu, and Diyi Yang. Putting humans in the natural language processing loop: A survey. In Su Lin Blodgett, Michael Madaio, Brendan O'Connor, Hanna Wallach, and Qian Yang, editors, *Proceedings of the First Workshop on Bridging Human–Computer Interaction and Natural Language Processing*, pages 47–52, Online, April 2021. Association for Computational Linguistics.
- [424] Mayur Wankhade, Annavarapu Chandra Sekhara Rao, and Chaitanya Kulkarni. A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review*, 55(7):5731–5780, 2022. doi: 10.1007/s10462-022-10144-1.
- [425] Cedric Waterschoot, Ernst van den Hemel, and Antal van den Bosch. Detecting minority arguments for mutual understanding: A moderation tool for the online climate change debate. In Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico Santus, Francis Bond, and Seung-Hoon Na, editors, *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6715–6725, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics.
- [426] Leon Weber-Genzel, Siyao Peng, Marie-Catherine de Marneffe, and Barbara Plank. Varierr nli: Separating annotation error from human label variation, 2024.
- [427] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856. Survey Certification.
- [428] Maxwell A Weinzierl and Sanda M Harabagiu. From hesitancy framings to vaccine hesitancy profiles: A journey of stance, ontological commitments and moral foundations. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, pages 1087–1097, 2022.
- [429] Galen Weld, Amy X. Zhang, and Tim Althoff. What makes online communities ‘better’? measuring values, consensus, and conflict across thousands of subreddits. *Proceedings of the International AAAI Conference on Web and Social Media*, 16(1):1121–1132, May 2022. doi: 10.1609/icwsm.v16i1.19363.
- [430] Orion Weller, Kevin Seppi, and Matt Gardner. When to use multi-task learning vs intermediate fine-tuning for pre-trained encoder transfer learning. In *Proceedings of*

- the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 272–282, 2022.
- [431] Jevin D West and Carl T Bergstrom. Misinformation in and about science. *Proceedings of the National Academy of Sciences*, 118(15):e1912444117, 2021.
 - [432] Janyce Wiebe, Theresa Wilson, and Claire Cardie. Annotating expressions of opinions and emotions in language. *Language resources and evaluation*, 39(2-3):165–210, 2005. doi: 10.1007/s10579-005-7880-9.
 - [433] Adina Williams, Nikita Nangia, and Samuel Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics, 2018.
 - [434] Michael Wojatzki, Torsten Zesch, Saif Mohammad, and Svetlana Kiritchenko. Agree or disagree: Predicting judgments on nuanced assertions. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 214–224, 2018.
 - [435] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Huggingface’s transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019.
 - [436] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics.
 - [437] Scott Wright and John Street. Democracy, deliberation and design: the case of online discussion forums. *New media & society*, 9(5):849–869, 2007. doi: 10.1177/1461444807081230.
 - [438] Chuhan Wu, Fangzhao Wu, Tao Qi, and Yongfeng Huang. SentiRec: Sentiment diversity-aware neural news recommendation. In Kam-Fai Wong, Kevin Knight, and Hua Wu, editors, *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 44–53, Suzhou, China, December 2020. Association for Computational Linguistics.
 - [439] Adam Wyner, Jodi Schneider, Katie Atkinson, and Trevor Bench-Capon. Semi-automated argumentative analysis of online product reviews. In *Computational Models of Argument*, pages 43–50. IOS Press, 2012. doi: 10.3233/978-1-61499-111-3-43.
 - [440] Yan Xia, Haiyi Zhu, Tun Lu, Peng Zhang, and Ning Gu. Exploring antecedents and consequences of toxicity in online discussions: A case study on reddit. *Proceedings of the ACM on Human-computer Interaction*, 4(CSCW2):1–23, oct 2020. doi: 10.1145/3415179.

- [441] Fei Xiong and Yun Liu. Opinion formation on social media: an empirical approach. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 24(1):013130, 03 2014. ISSN 1054-1500. doi: 10.1063/1.4866011.
- [442] Jun Xu, Xiangnan He, and Hang Li. Deep learning for matching in search and recommendation. *Foundations and Trends® in Information Retrieval*, 14(2–3):102–288, 2020. ISSN 1554-0669. doi: 10.1561/15000000076.
- [443] Qiongkai Xu, Christian Walder, and Chenchen Xu. Humanly certifying superhuman classifiers. *arXiv preprint arXiv:2109.07867*, 2021.
- [444] Naomi Yamashita and Toru Ishida. Effects of machine translation on collaborative work. In *Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work*, pages 515–524, 2006.
- [445] Weizhe Yuan, Graham Neubig, and Pengfei Liu. Bartscore: Evaluating generated text as text generation. *Advances in Neural Information Processing Systems*, 34:27263–27277, 2021.
- [446] Lihi Zelnik-Manor and Pietro Perona. Self-tuning spectral clustering. In *Proceedings of the 17th International Conference on Neural Information Processing Systems*, volume 17 of *NIPS’04*, page 1601–1608, Cambridge, MA, USA, 2004. MIT Press.
- [447] Leihan Zhang, Jichang Zhao, and Ke Xu. Who creates trends in online social media: The crowd or opinion leaders? *Journal of Computer-Mediated Communication*, 21(1): 1–16, 2016.
- [448] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*, 2020.
- [449] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*, 2020.
- [450] Ye Zhang, Matthew Lease, and Byron C. Wallace. Active Discriminative Text Representation Learning. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence*, pages 3386–3392, San Francisco, California, USA, 2017.
- [451] Zhisong Zhang, Emma Strubell, and Eduard Hovy. A Survey of Active Learning for Natural Language Processing. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, EMNLP ’22, pages 6166–6190. ACL, 2022.
- [452] Yuekai Zhao, Haoran Zhang, Shuchang Zhou, and Zhihua Zhang. Active Learning Approaches to Enhancing Neural Machine Translation. In *Findings of the Association for Computational Linguistics*, EMNLP 2020, pages 1796–1806, Online, 2020. ACL. doi: 10.18653/v1/2020.findings-emnlp.162.

- [453] Ming Zhong, Pengfei Liu, Yiran Chen, Danqing Wang, Xipeng Qiu, and Xuanjing Huang. Extractive summarization as text matching. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6197–6208, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.552.
- [454] Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. Can large language models transform computational social science? *Computational Linguistics*, pages 1–55, 2024.
- [455] Elena Ziliotti, Patricia Reyes Benavides, Arthur Gwagwa, and Matthew Dennis. *Social Media and Democracy*, chapter 2, pages 33–52. Open Book Publishers, 2023. doi: 10.11647/obp.0366.02.

Acknowledgments

Some of the people listed in this acknowledgments section will reside in more than one subgroup, in which case please know that I might have grouped you arbitrarily. Figuring out interconnections is left as an exercise for the reader.

Pradeep, you were my closest supervisor, the one leading me through it all. Thank you for your guidance during the years in my PhD and beyond. You taught me how to do so many things the scientific way: writing, collaborating, discussing, and organizing, to name a few. I consider you a great example of how to be a great person, professionally and personally.

Catholijn, thank you for always having my back and being kind in cases where I could use some steering. You inspire me to never let go of my enthusiasm and to keep talking and honing my research. I will always remember your advice to *frappez toujours*.

Piek, in you I found a strong sparring partner. Your office was always open, and inviting for a healthy discussion. You went along with my endless fidgeting about what to do next and let me find my path. I am very grateful you facilitated my second home at the VU.

Aske, thank you for welcoming me to Leiden University. You were (physically!) there when I started during COVID and provided advice when I needed it. You showed me the value of pursuing what I want to do over all other things.

My closest colleagues had a significant hand in my PhD. Enrico, if there is such a thing as an academic brother I am sure it is you. You were there from the first interview to my final paper, and I am grateful you are the best scholar-in-crime I could have imagined. I'd gladly share a five-star hotel room with you again. Bram, you were not only my closest colleague in Leiden but also a friend outside of it. Your enjoyment of the good things in life, as well as your endless capacity for banter, makes me enjoy every situation we find ourselves in, and you lifted the quality of the many events we attended together. To complete the group of culinarians, Mani, thank you for starting our tradition of high-quality dinner evenings after going to Berlin together. I hope we find our way to the same table again soon.

To my colleagues from CLTL, I will dearly miss our (occasionally lengthy) coffee breaks. Selene, thank you for your exceptional kindness, for being up for exploring new foods on our adventures, and for never failing to lift my spirit. Lea, your warm heart and critical mind make you one of the kindest and coolest people I know. Urja, we go way back to the BSc KI at UvA, I'm glad our paths have crossed for so long. Myrthe, you are my opinion-mining buddy and your drive for academic diligence motivates me to be a better scientist. Throughout my PhD, I managed to travel to different places, be it for conferences or vacation afterward. Jaap, Stella, Jonathan, and Baran, the times together were among my favorites. Tom, we explored Singapore together before ending up as office mates in Leiden, the times in the new building were short but sweet. To the HI IGLU team: Tae, Putra, and Kata, thank you for besting the challenge together and finding each other again in different parts of the world. To my other coauthors, thank you for teaching me much about (interdisciplinary) research. Anatol and Ruth, thanks for calmly explaining the basics of political science, Luciano, for kickstarting my PhD alongside Enrico, and Neele for pushing the last project in my PhD to successful completion.

The past years would have been very different if it were not for the groups I found myself in. At LIACS, my thanks extend to Thomas, Matthias, Andreas (I hope your plants live long lives), Zhao, Mike, Andrius, Alan, Álvaro, Felix, Koen, Annie, Joost, and all other present and past members of the RLG group, as well as Suzan, Amin, Arian, Gijs, Zhaochun, Juan, and the other members of the TMR group, and finally Bernard, Ramira, Max, Tessa, Bram, Yuchen, and the other members of the CIL group. At the VU Amsterdam, my thanks go to the CLTL group for hosting me: Leon, Annika, Wende, Stefan, Ellie, Pia, Antske, Levi, Angel, Ilia, Lisa, and all other present and past members of the CLTL group at the VU. Lastly, at the TU Delft: Ruben, Pei-Yu, Masha, Morita, Emma, Nele, Amir, Carolina, Zuzanna, Deborah, Myrthe, Catha, Frans, Siddharth, and a special mention for the organizational miracle that is Anita, as well as any other present or past member in the II group.

None of my work would have come to fruition without those in the Hybrid Intelligence consortium. Thank you for powering through the difficult COVID times with great success. Bart, and Cor, thank you for instilling the HI values and academic lifeblood in me. Tiffany, I greatly enjoyed co-organizing the HI PhD reading group with you. Thanks for the wonderful times to the early HI PhDs: Johanna (I hope you still have the light you stole), Annet, Ludi, Giacomo, Wijnand, Mark, Íñigo, Niklas, Maria, Sharvaree, Nicole, Aishwarya, Merle, Loan, Emre, J.D., Delaram, Anna, Jonne, Feline, as well as the later cohorts of HI PhDs. I would also like to express my gratitude to the senior members of the HI consortium, including Rineke, Ilaria, Victor, Davide, Davide, Jasper, Henry, Erman, Annette, Dan, Kim, Herke, Silja, Pinar, Birna, Shenghui, Roel, Bernd, and all other present and past HI members. Finally, I'd like to thank the management team, including Frank, Wendy, and Stefan for steering the HI meetings in the right direction and succeeding in creating a tight-knit group that felt like a home.

Ik had mijn PhD niet kunnen afmaken zonder mijn vrienden en familie. Jullie zorgen er voor dat ik mijn werk even kan vergeten en ruimte krijg voor andere dingen. Aan de Kumdo instructor group: Bas, Dominic, en Kelly (en Devil), jullie zijn meer dan vrienden en ik hoop dat wij samen trainen tot we oud zijn (en dat zijn we nog niet). Kevin (honk!), Sterre, Maya, Ray, Lotte, Oumaima, Max, Nele, Quintin en Quinten, Evgeniia, Zena, Tibor, Yuri, Meg, en alle anderen bij Dojo Den Drijver en Gungsul Academy, 해동! Jonathan, bedankt voor het altijd beschikbaar zijn om te praten over alle dingen in het leven, jouw enthousiasme over werkelijk elk onderwerp is ontzettend aanstekelijk. Arianne, Bart, Thomas, Roan, en de rest van mijn goeie vrienden, ik ben dankbaar voor jullie warmte en gesprekken over de jaren. Sabina, jij hebt mij van het begin tot het eind altijd van aanmoediging voorzien, zowel in tijden van deadlinestress, als in de vakanties samen erna. Door jouw rol als een onuitputbaar klankbord is dit proefschrift ook een beetje van jou.

Als laatste gaat mijn dank naar degenen die mij al mijn hele leven samen met mij zijn. Arwen, Sanne, Theo en Petra, jullie staan aan de basis van mijn nieuwsgierigheid om de wereld om me heen beter te begrijpen. Mama, en papa en Jolanda, bedankt voor jullie onvoorwaardelijke steun. Jullie hebben me gevormd tot de persoon die ik ben. Annemein, Felix en Aster, wij zijn met z'n vieren een eenheid, en dat was niet anders tijdens mijn PhD. Ik kan me niet voorstellen hoe ik dit allemaal zou hebben gedaan zonder zo'n geweldig liefdevol zootje zusjes en broertje om me heen.

Curriculum Vitæ

Michiel Theo van der Meer was born on October 29, 1995, in Groningen, the Netherlands. He graduated with a gymnasium degree from the Jan van Egmond Lyceum in Purmerend, Noord-Holland, in 2014. Michiel then pursued both his Bachelor of Science and Master of Science in Artificial Intelligence at the University of Amsterdam, completing these degrees from 2014 to 2017 and from 2017 to 2020, respectively.

During his academic career, Michiel was actively involved in various roles. He served as a teaching assistant, was part of the education committee in his study association, and was a dedicated team member of the Dutch Nao Team from 2015 to 2019. In this team, he held several positions, including conducting weekly meetings, facilitating group decisions, and developing soccer-playing robots. Michiel traveled with his team to notable events such as Techfest 2015 in Mumbai, India, and RoboCup competitions from 2016 to 2019 in Leipzig, Germany; Tehran, Iran; Nagoya, Japan; and Montréal, Canada. He completed his master's degree cum laude in 2020, with a thesis focused on explainability in reinforcement learning.

Alongside his academic pursuits, Michiel has garnered diverse professional experience in the AI field. From 2018 to 2020, he worked as an AI developer and trainer at Millennials.ai, where he developed machine learning approaches for healthcare services and delivered lectures on AI and NLP basics to beginner audiences at various companies, governmental departments, and public institutions. In 2020, he served as a Data Engineer at InBiome in Amsterdam before embarking on his PhD studies at Leiden University, which he completed from 2020 to 2024. During his PhD studies, he took courses in transferable skills such as research methods for computer science, project and data management, and scientific conduct, among others. Starting from August 2024, Michiel is a postdoctoral researcher at the Idiap Research Institute in Martigny, Switzerland.

List of Publications

2024

1. **Michiel van der Meer**, Neele Falk, Pradeep K. Murukannaiah, Enrico Liscio. 2024. Annotator-Centric Active Learning for Subjective NLP Tasks. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18537–18555, Miami, Florida, USA. Association for Computational Linguistics.
2. **Michiel van der Meer**, Enrico Liscio, Catholijn M. Jonker, Aske Plaat, Piek Vossen, Pradeep K. Murukannaiah. 2024. A Hybrid Intelligence Method for Argument Mining. In *Journal of Artificial Intelligence Research* 80, pages 1187–1222.
3. **Michiel van der Meer**. Facilitating Online Opinion Diversity through Hybrid NLP Approaches (Thesis proposal). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop)*, pages 272–284, Mexico City, Mexico. Association for Computational Linguistics.
4. **Michiel van der Meer**, Piek Vossen, Catholijn M. Jonker, and Pradeep K. Murukannaiah. Value-Sensitive Disagreement Analysis for Online Deliberation. In *HHA1 2024: Hybrid Human AI Systems for the Social Good (Extended Abstracts)*, pages 481–484, Malmö, Sweden. IOS Press.
5. **Michiel van der Meer**, Piek Vossen, Catholijn M. Jonker, and Pradeep K. Murukannaiah. An Empirical Analysis of Diversity in Argument Summarization. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2028–2045, St. Julian's, Malta. Association for Computational Linguistics.

2023

1. **Michiel van der Meer**, Piek Vossen, Catholijn M. Jonker, and Pradeep K. Murukannaiah. 2023. Do Differences in Values Influence Disagreements in Online Discussions? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15986–16008, Singapore. Association for Computational Linguistics.
2. Lea Krause, Selene Báez Santamaría, **Michiel van der Meer**, and Urja Khurana. 2023. Leveraging Few-Shot Data Augmentation and Waterfall Prompting for Response Generation. In *Proceedings of The Eleventh Dialog System Technology Challenge*, pages 193–205, Prague, Czech Republic. Association for Computational Linguistics.
3. Thomas M. Moerland, Matthias Müller-Brockhausen, Zhao Yang, Andrius Bernatavicius, Koen Ponse, Tom Kouwenhoven, Andreas Sauter, **Michiel van der Meer**, Bram Renting, and Aske Plaat. EduGym: An Environment Suite for Reinforcement Learning Education. *arXiv preprint*.

2022

1. Ruth Shortall, Anatol Ippen, **Michiel van der Meer**, Pradeep K. Murukannaiah, Catholijn M. Jonker. 2022. Reason Against the Machine: Future Directions for Mass Online Deliberation. In *Frontiers in Political Science*.
-  2. **Michiel van der Meer**, Myrthe Reuver, Urja Khurana, Lea Krause, and Selene Báez Santamaría. 2022. Will It Blend? Mixing Training Paradigms & Prompting for Argument Quality Prediction. In *Proceedings of the 9th Workshop on Argument Mining*, pages 95–103, Online and in Gyeongju, Republic of Korea. International Conference on Computational Linguistics.
3. Enrico Liscio, **Michiel van der Meer**, Luciano C. Siebert, Catholijn M. Jonker, Pradeep K. Murukannaiah. 2022. What Values Should an Agent Align with? An Empirical Comparison of General and Context-Specific Values. In *Autonomous Agents and Multi-Agent Systems* 36, 23.
-  4. **Michiel van der Meer**, Enrico Liscio, Catholijn M. Jonker, Aske Plaat, Piek Vossen, Pradeep K. Murukannaiah. 2022. HyEnA: A Hybrid Method for Extracting Arguments from Opinions. In *HHAI2022: Augmenting Human Intellect*, pages 17–31, Amsterdam, the Netherlands. IOS Press. **[Best paper award]**

2021

1. Enrico Liscio, **Michiel van der Meer**, Luciano C. Siebert, Catholijn M. Jonker, Niek Mouter, Pradeep K. Murukannaiah. 2021. Axes: Identifying and Evaluating Context-Specific Values. In *Proceedings of the 20th International Conference on Autonomous Agents and Multiagent Systems*, pages 799–808. Online. IFAAMAS.
2. Enrico Liscio, **Michiel van der Meer**, Catholijn M. Jonker, Pradeep K. Murukannaiah. 2021. A Collaborative Platform for Identifying Context-Specific Values: Demo Track. In *Proceedings of the 20th International Conference on Autonomous Agents and Multiagent Systems*, pages 1773–1775, Online. IFAAMAS.

 Included in this thesis.

SIKS Dissertations

-
- 2016 01 Syed Saiden Abbas (RUN), Recognition of Shapes by Humans and Machines
 - 02 Michiel Christiaan Meulendijk (UU), Optimizing medication reviews through decision support: prescribing a better pill to swallow
 - 03 Maya Sappelli (RUN), Knowledge Work in Context: User Centered Knowledge Worker Support
 - 04 Laurens Rietveld (VUA), Publishing and Consuming Linked Data
 - 05 Evgeny Sherkhonov (UvA), Expanded Acyclic Queries: Containment and an Application in Explaining Missing Answers
 - 06 Michel Wilson (TUD), Robust scheduling in an uncertain environment
 - 07 Jeroen de Man (VUA), Measuring and modeling negative emotions for virtual training
 - 08 Matje van de Camp (TiU), A Link to the Past: Constructing Historical Social Networks from Unstructured Data
 - 09 Archana Nottamkandath (VUA), Trusting Crowdsourced Information on Cultural Artefacts
 - 10 George Karafotias (VUA), Parameter Control for Evolutionary Algorithms
 - 11 Anne Schuth (UvA), Search Engines that Learn from Their Users
 - 12 Max Knobbout (UU), Logics for Modelling and Verifying Normative Multi-Agent Systems
 - 13 Nana Baah Gyan (VUA), The Web, Speech Technologies and Rural Development in West Africa - An ICT4D Approach
 - 14 Ravi Khadka (UU), Revisiting Legacy Software System Modernization
 - 15 Steffen Michels (RUN), Hybrid Probabilistic Logics - Theoretical Aspects, Algorithms and Experiments
 - 16 Guangliang Li (UvA), Socially Intelligent Autonomous Agents that Learn from Human Reward
 - 17 Berend Weel (VUA), Towards Embodied Evolution of Robot Organisms
 - 18 Albert Meroño Peñuela (VUA), Refining Statistical Data on the Web
 - 19 Julia Efremova (TU/e), Mining Social Structures from Genealogical Data
 - 20 Daan Odijk (UvA), Context & Semantics in News & Web Search
 - 21 Alejandro Moreno Céleri (UT), From Traditional to Interactive Playspaces: Automatic Analysis of Player Behavior in the Interactive Tag Playground
 - 22 Grace Lewis (VUA), Software Architecture Strategies for Cyber-Foraging Systems
 - 23 Fei Cai (UvA), Query Auto Completion in Information Retrieval
 - 24 Brend Wanders (UT), Repurposing and Probabilistic Integration of Data; An Iterative and data model independent approach
 - 25 Julia Kiseleva (TU/e), Using Contextual Information to Understand Searching and Browsing Behavior

- 26 Dilhan Thilakarathne (VUA), In or Out of Control: Exploring Computational Models to Study the Role of Human Awareness and Control in Behavioural Choices, with Applications in Aviation and Energy Management Domains
 - 27 Wen Li (TUD), Understanding Geo-spatial Information on Social Media
 - 28 Mingxin Zhang (TUD), Large-scale Agent-based Social Simulation - A study on epidemic prediction and control
 - 29 Nicolas Höning (TUD), Peak reduction in decentralised electricity systems - Markets and prices for flexible planning
 - 30 Ruud Mattheij (TiU), The Eyes Have It
 - 31 Mohammad Khelghati (UT), Deep web content monitoring
 - 32 Eelco Vriezekolk (UT), Assessing Telecommunication Service Availability Risks for Crisis Organisations
 - 33 Peter Bloem (UvA), Single Sample Statistics, exercises in learning from just one example
 - 34 Dennis Schunselaar (TU/e), Configurable Process Trees: Elicitation, Analysis, and Enactment
 - 35 Zhaochun Ren (UvA), Monitoring Social Media: Summarization, Classification and Recommendation
 - 36 Daphne Karreman (UT), Beyond R2D2: The design of nonverbal interaction behavior optimized for robot-specific morphologies
 - 37 Giovanni Sileno (UvA), Aligning Law and Action - a conceptual and computational inquiry
 - 38 Andrea Minuto (UT), Materials that Matter - Smart Materials meet Art & Interaction Design
 - 39 Merijn Bruijnes (UT), Believable Suspect Agents; Response and Interpersonal Style Selection for an Artificial Suspect
 - 40 Christian Detweiler (TUD), Accounting for Values in Design
 - 41 Thomas King (TUD), Governing Governance: A Formal Framework for Analysing Institutional Design and Enactment Governance
 - 42 Spyros Martzoukos (UvA), Combinatorial and Compositional Aspects of Bilingual Aligned Corpora
 - 43 Saskia Koldijk (RUN), Context-Aware Support for Stress Self-Management: From Theory to Practice
 - 44 Thibault Sellam (UvA), Automatic Assistants for Database Exploration
 - 45 Bram van de Laar (UT), Experiencing Brain-Computer Interface Control
 - 46 Jorge Gallego Perez (UT), Robots to Make you Happy
 - 47 Christina Weber (UL), Real-time foresight - Preparedness for dynamic innovation networks
 - 48 Tanja Buttler (TUD), Collecting Lessons Learned
 - 49 Gleb Polevoy (TUD), Participation and Interaction in Projects. A Game-Theoretic Analysis
 - 50 Yan Wang (TiU), The Bridge of Dreams: Towards a Method for Operational Performance Alignment in IT-enabled Service Supply Chains
-
- 2017 01 Jan-Jaap Oerlemans (UL), Investigating Cybercrime

- 02 Sjoerd Timmer (UU), Designing and Understanding Forensic Bayesian Networks using Argumentation
- 03 Daniël Harold Telgen (UU), Grid Manufacturing: A Cyber-Physical Approach with Autonomous Products and Reconfigurable Manufacturing Machines
- 04 Mrunal Gawade (CWI), Multi-core Parallelism in a Column-store
- 05 Mahdiah Shadi (UvA), Collaboration Behavior
- 06 Damir Vandic (EUR), Intelligent Information Systems for Web Product Search
- 07 Roel Bertens (UU), Insight in Information: from Abstract to Anomaly
- 08 Rob Konijn (VUA), Detecting Interesting Differences: Data Mining in Health Insurance Data using Outlier Detection and Subgroup Discovery
- 09 Dong Nguyen (UT), Text as Social and Cultural Data: A Computational Perspective on Variation in Text
- 10 Robby van Delden (UT), (Steering) Interactive Play Behavior
- 11 Florian Kunneman (RUN), Modelling patterns of time and emotion in Twitter #anticipointment
- 12 Sander Leemans (TU/e), Robust Process Mining with Guarantees
- 13 Gijs Huisman (UT), Social Touch Technology - Extending the reach of social touch through haptic technology
- 14 Shoshannah Tekofsky (TiU), You Are Who You Play You Are: Modelling Player Traits from Video Game Behavior
- 15 Peter Berck (RUN), Memory-Based Text Correction
- 16 Aleksandr Chuklin (UvA), Understanding and Modeling Users of Modern Search Engines
- 17 Daniel Dimov (UL), Crowdsourced Online Dispute Resolution
- 18 Ridho Reinanda (UvA), Entity Associations for Search
- 19 Jeroen Vuurens (UT), Proximity of Terms, Texts and Semantic Vectors in Information Retrieval
- 20 Mohammadbashir Sedighi (TUD), Fostering Engagement in Knowledge Sharing: The Role of Perceived Benefits, Costs and Visibility
- 21 Jeroen Linssen (UT), Meta Matters in Interactive Storytelling and Serious Gaming (A Play on Worlds)
- 22 Sara Magliacane (VUA), Logics for causal inference under uncertainty
- 23 David Graus (UvA), Entities of Interest — Discovery in Digital Traces
- 24 Chang Wang (TUD), Use of Affordances for Efficient Robot Learning
- 25 Veruska Zamborlini (VUA), Knowledge Representation for Clinical Guidelines, with applications to Multimorbidity Analysis and Literature Search
- 26 Merel Jung (UT), Socially intelligent robots that understand and respond to human touch
- 27 Michiel Joosse (UT), Investigating Positioning and Gaze Behaviors of Social Robots: People's Preferences, Perceptions and Behaviors
- 28 John Klein (VUA), Architecture Practices for Complex Contexts
- 29 Adel Alhuraibi (TiU), From IT-Business Strategic Alignment to Performance: A Moderated Mediation Model of Social Innovation, and Enterprise Governance of IT"
- 30 Wilma Latuny (TiU), The Power of Facial Expressions

-
- 31 Ben Ruijl (UL), Advances in computational methods for QFT calculations
 - 32 Thaer Samar (RUN), Access to and Retrieval of Content in Web Archives
 - 33 Brigit van Loggem (OU), Towards a Design Rationale for Software Documentation: A Model of Computer-Mediated Activity
 - 34 Maren Scheffel (OU), The Evaluation Framework for Learning Analytics
 - 35 Martine de Vos (VUA), Interpreting natural science spreadsheets
 - 36 Yuanhao Guo (UL), Shape Analysis for Phenotype Characterisation from High-throughput Imaging
 - 37 Alejandro Montes Garcia (TU/e), WiBAF: A Within Browser Adaptation Framework that Enables Control over Privacy
 - 38 Alex Kayal (TUD), Normative Social Applications
 - 39 Sara Ahmadi (RUN), Exploiting properties of the human auditory system and compressive sensing methods to increase noise robustness in ASR
 - 40 Altaf Hussain Abro (VUA), Steer your Mind: Computational Exploration of Human Control in Relation to Emotions, Desires and Social Support For applications in human-aware support systems
 - 41 Adnan Manzoor (VUA), Minding a Healthy Lifestyle: An Exploration of Mental Processes and a Smart Environment to Provide Support for a Healthy Lifestyle
 - 42 Elena Sokolova (RUN), Causal discovery from mixed and missing data with applications on ADHD datasets
 - 43 Maaïke de Boer (RUN), Semantic Mapping in Video Retrieval
 - 44 Garm Lucassen (UU), Understanding User Stories - Computational Linguistics in Agile Requirements Engineering
 - 45 Bas Testerink (UU), Decentralized Runtime Norm Enforcement
 - 46 Jan Schneider (OU), Sensor-based Learning Support
 - 47 Jie Yang (TUD), Crowd Knowledge Creation Acceleration
 - 48 Angel Suarez (OU), Collaborative inquiry-based learning
-
- 2018 01 Han van der Aa (VUA), Comparing and Aligning Process Representations
 - 02 Felix Mannhardt (TU/e), Multi-perspective Process Mining
 - 03 Steven Bosems (UT), Causal Models For Well-Being: Knowledge Modeling, Model-Driven Development of Context-Aware Applications, and Behavior Prediction
 - 04 Jordan Janeiro (TUD), Flexible Coordination Support for Diagnosis Teams in Data-Centric Engineering Tasks
 - 05 Hugo Huurdeman (UvA), Supporting the Complex Dynamics of the Information Seeking Process
 - 06 Dan Ionita (UT), Model-Driven Information Security Risk Assessment of Socio-Technical Systems
 - 07 Jieting Luo (UU), A formal account of opportunism in multi-agent systems
 - 08 Rick Smetsers (RUN), Advances in Model Learning for Software Systems
 - 09 Xu Xie (TUD), Data Assimilation in Discrete Event Simulations
 - 10 Julienka Mollee (VUA), Moving forward: supporting physical activity behavior change through intelligent technology
 - 11 Mahdi Sargolzaei (UvA), Enabling Framework for Service-oriented Collaborative Networks

- 12 Xixi Lu (TU/e), Using behavioral context in process mining
 - 13 Seyed Amin Tabatabaei (VUA), Computing a Sustainable Future
 - 14 Bart Joosten (TiU), Detecting Social Signals with Spatiotemporal Gabor Filters
 - 15 Naser Davarzani (UM), Biomarker discovery in heart failure
 - 16 Jaebok Kim (UT), Automatic recognition of engagement and emotion in a group of children
 - 17 Jianpeng Zhang (TU/e), On Graph Sample Clustering
 - 18 Henriette Nakad (UL), De Notaris en Private Rechtspraak
 - 19 Minh Duc Pham (VUA), Emergent relational schemas for RDF
 - 20 Manxia Liu (RUN), Time and Bayesian Networks
 - 21 Aad Slotmaker (OU), EMERGO: a generic platform for authoring and playing scenario-based serious games
 - 22 Eric Fernandes de Mello Araújo (VUA), Contagious: Modeling the Spread of Behaviours, Perceptions and Emotions in Social Networks
 - 23 Kim Schouten (EUR), Semantics-driven Aspect-Based Sentiment Analysis
 - 24 Jered Vroon (UT), Responsive Social Positioning Behaviour for Semi-Autonomous Telepresence Robots
 - 25 Riste Gligorov (VUA), Serious Games in Audio-Visual Collections
 - 26 Roelof Anne Jelle de Vries (UT), Theory-Based and Tailor-Made: Motivational Messages for Behavior Change Technology
 - 27 Maikel Leemans (TU/e), Hierarchical Process Mining for Scalable Software Analysis
 - 28 Christian Willemse (UT), Social Touch Technologies: How they feel and how they make you feel
 - 29 Yu Gu (TiU), Emotion Recognition from Mandarin Speech
 - 30 Wouter Beek (VUA), The "K" in "semantic web" stands for "knowledge": scaling semantics to the web
-
- 2019 01 Rob van Eijk (UL), Web privacy measurement in real-time bidding systems. A graph-based approach to RTB system classification
 - 02 Emmanuelle Beauxis Aussalet (CWI, UU), Statistics and Visualizations for Assessing Class Size Uncertainty
 - 03 Eduardo Gonzalez Lopez de Murillas (TU/e), Process Mining on Databases: Extracting Event Data from Real Life Data Sources
 - 04 Ridho Rahmadi (RUN), Finding stable causal structures from clinical data
 - 05 Sebastiaan van Zelst (TU/e), Process Mining with Streaming Data
 - 06 Chris Dijkshoorn (VUA), Nichesourcing for Improving Access to Linked Cultural Heritage Datasets
 - 07 Soude Fazeli (TUD), Recommender Systems in Social Learning Platforms
 - 08 Frits de Nijs (TUD), Resource-constrained Multi-agent Markov Decision Processes
 - 09 Fahimeh Alizadeh Moghaddam (UvA), Self-adaptation for energy efficiency in software systems
 - 10 Qing Chuan Ye (EUR), Multi-objective Optimization Methods for Allocation and Prediction

- 11 Yue Zhao (TUD), Learning Analytics Technology to Understand Learner Behavioral Engagement in MOOCs
 - 12 Jacqueline Heinerman (VUA), Better Together
 - 13 Guanliang Chen (TUD), MOOC Analytics: Learner Modeling and Content Generation
 - 14 Daniel Davis (TUD), Large-Scale Learning Analytics: Modeling Learner Behavior & Improving Learning Outcomes in Massive Open Online Courses
 - 15 Erwin Walraven (TUD), Planning under Uncertainty in Constrained and Partially Observable Environments
 - 16 Guangming Li (TU/e), Process Mining based on Object-Centric Behavioral Constraint (OCBC) Models
 - 17 Ali Hurriyetoglu (RUN), Extracting actionable information from microtexts
 - 18 Gerard Wagenaar (UU), Artefacts in Agile Team Communication
 - 19 Vincent Koeman (TUD), Tools for Developing Cognitive Agents
 - 20 Chide Groenouwe (UU), Fostering technically augmented human collective intelligence
 - 21 Cong Liu (TU/e), Software Data Analytics: Architectural Model Discovery and Design Pattern Detection
 - 22 Martin van den Berg (VUA), Improving IT Decisions with Enterprise Architecture
 - 23 Qin Liu (TUD), Intelligent Control Systems: Learning, Interpreting, Verification
 - 24 Anca Dumitrache (VUA), Truth in Disagreement - Crowdsourcing Labeled Data for Natural Language Processing
 - 25 Emiel van Miltenburg (VUA), Pragmatic factors in (automatic) image description
 - 26 Prince Singh (UT), An Integration Platform for Synchromodal Transport
 - 27 Alessandra Antonaci (OU), The Gamification Design Process applied to (Massive) Open Online Courses
 - 28 Esther Kuindersma (UL), Cleared for take-off: Game-based learning to prepare airline pilots for critical situations
 - 29 Daniel Formolo (VUA), Using virtual agents for simulation and training of social skills in safety-critical circumstances
 - 30 Vahid Yazdanpanah (UT), Multiagent Industrial Symbiosis Systems
 - 31 Milan Jelisavcic (VUA), Alive and Kicking: Baby Steps in Robotics
 - 32 Chiara Sironi (UM), Monte-Carlo Tree Search for Artificial General Intelligence in Games
 - 33 Anil Yaman (TU/e), Evolution of Biologically Inspired Learning in Artificial Neural Networks
 - 34 Negar Ahmadi (TU/e), EEG Microstate and Functional Brain Network Features for Classification of Epilepsy and PNES
 - 35 Lisa Facey-Shaw (OU), Gamification with digital badges in learning programming
 - 36 Kevin Ackermans (OU), Designing Video-Enhanced Rubrics to Master Complex Skills
 - 37 Jian Fang (TUD), Database Acceleration on FPGAs
 - 38 Akos Kadar (OU), Learning visually grounded and multilingual representations
-

- 2020 01 Armon Toubman (UL), Calculated Moves: Generating Air Combat Behaviour
02 Marcos de Paula Bueno (UL), Unraveling Temporal Processes using Probabilistic Graphical Models
03 Mostafa Deghani (UvA), Learning with Imperfect Supervision for Language Understanding
04 Maarten van Gompel (RUN), Context as Linguistic Bridges
05 Yulong Pei (TU/e), On local and global structure mining
06 Preethu Rose Anish (UT), Stimulation Architectural Thinking during Requirements Elicitation - An Approach and Tool Support
07 Wim van der Vegt (OU), Towards a software architecture for reusable game components
08 Ali Mirsoleimani (UL), Structured Parallel Programming for Monte Carlo Tree Search
09 Myriam Traub (UU), Measuring Tool Bias and Improving Data Quality for Digital Humanities Research
10 Alifah Syamsiyah (TU/e), In-database Preprocessing for Process Mining
11 Sepideh Mesbah (TUD), Semantic-Enhanced Training Data Augmentation Methods for Long-Tail Entity Recognition Models
12 Ward van Breda (VUA), Predictive Modeling in E-Mental Health: Exploring Applicability in Personalised Depression Treatment
13 Marco Virgolin (CWI), Design and Application of Gene-pool Optimal Mixing Evolutionary Algorithms for Genetic Programming
14 Mark Raasveldt (CWI/UL), Integrating Analytics with Relational Databases
15 Konstantinos Georgiadis (OU), Smart CAT: Machine Learning for Configurable Assessments in Serious Games
16 Ilona Wilmont (RUN), Cognitive Aspects of Conceptual Modelling
17 Daniele Di Mitri (OU), The Multimodal Tutor: Adaptive Feedback from Multimodal Experiences
18 Georgios Methenitis (TUD), Agent Interactions & Mechanisms in Markets with Uncertainties: Electricity Markets in Renewable Energy Systems
19 Guido van Capelleveen (UT), Industrial Symbiosis Recommender Systems
20 Albert Hankel (VUA), Embedding Green ICT Maturity in Organisations
21 Karine da Silva Miras de Araujo (VUA), Where is the robot?: Life as it could be
22 Maryam Masoud Khamis (RUN), Understanding complex systems implementation through a modeling approach: the case of e-government in Zanzibar
23 Rianne Conijn (UT), The Keys to Writing: A writing analytics approach to studying writing processes using keystroke logging
24 Lenin da Nóbrega Medeiros (VUA/RUN), How are you feeling, human? Towards emotionally supportive chatbots
25 Xin Du (TU/e), The Uncertainty in Exceptional Model Mining
26 Krzysztof Leszek Sadowski (UU), GAMBIT: Genetic Algorithm for Model-Based mixed-Integer opTimization
27 Ekaterina Muravyeva (TUD), Personal data and informed consent in an educational context

-
- 28 Bibeg Limbu (TUD), Multimodal interaction for deliberate practice: Training complex skills with augmented reality
 - 29 Ioan Gabriel Bucur (RUN), Being Bayesian about Causal Inference
 - 30 Bob Zadok Blok (UL), Creatief, Creatiever, Creatiefst
 - 31 Gongjin Lan (VUA), Learning better – From Baby to Better
 - 32 Jason Rhuggenaath (TU/e), Revenue management in online markets: pricing and online advertising
 - 33 Rick Gilsing (TU/e), Supporting service-dominant business model evaluation in the context of business model innovation
 - 34 Anna Bon (UM), Intervention or Collaboration? Redesigning Information and Communication Technologies for Development
 - 35 Siamak Farshidi (UU), Multi-Criteria Decision-Making in Software Production
-
- 2021 01 Francisco Xavier Dos Santos Fonseca (TUD), Location-based Games for Social Interaction in Public Space
 - 02 Rijk Mercur (TUD), Simulating Human Routines: Integrating Social Practice Theory in Agent-Based Models
 - 03 Seyyed Hadi Hashemi (UvA), Modeling Users Interacting with Smart Devices
 - 04 Ioana Jivet (OU), The Dashboard That Loved Me: Designing adaptive learning analytics for self-regulated learning
 - 05 Davide Dell'Anna (UU), Data-Driven Supervision of Autonomous Systems
 - 06 Daniel Davison (UT), "Hey robot, what do you think?" How children learn with a social robot
 - 07 Armel Lefebvre (UU), Research data management for open science
 - 08 Nardie Fanchamps (OU), The Influence of Sense-Reason-Act Programming on Computational Thinking
 - 09 Cristina Zaga (UT), The Design of Robothings. Non-Anthropomorphic and Non-Verbal Robots to Promote Children's Collaboration Through Play
 - 10 Quinten Meertens (UvA), Misclassification Bias in Statistical Learning
 - 11 Anne van Rossum (UL), Nonparametric Bayesian Methods in Robotic Vision
 - 12 Lei Pi (UL), External Knowledge Absorption in Chinese SMEs
 - 13 Bob R. Schadenberg (UT), Robots for Autistic Children: Understanding and Facilitating Predictability for Engagement in Learning
 - 14 Negin Samaeemofrad (UL), Business Incubators: The Impact of Their Support
 - 15 Onat Ege Adali (TU/e), Transformation of Value Propositions into Resource Re-Configurations through the Business Services Paradigm
 - 16 Esam A. H. Ghaleb (UM), Bimodal emotion recognition from audio-visual cues
 - 17 Dario Dotti (UM), Human Behavior Understanding from motion and bodily cues using deep neural networks
 - 18 Remi Wieten (UU), Bridging the Gap Between Informal Sense-Making Tools and Formal Systems - Facilitating the Construction of Bayesian Networks and Argumentation Frameworks
 - 19 Roberto Verdecchia (VUA), Architectural Technical Debt: Identification and Management
 - 20 Masoud Mansoury (TU/e), Understanding and Mitigating Multi-Sided Exposure Bias in Recommender Systems

-
- 21 Pedro Thiago Timbó Holanda (CWI), Progressive Indexes
 - 22 Sihang Qiu (TUD), Conversational Crowdsourcing
 - 23 Hugo Manuel Proença (UL), Robust rules for prediction and description
 - 24 Kaijie Zhu (TU/e), On Efficient Temporal Subgraph Query Processing
 - 25 Eoin Martino Grua (VUA), The Future of E-Health is Mobile: Combining AI and Self-Adaptation to Create Adaptive E-Health Mobile Applications
 - 26 Benno Kruit (CWI/VUA), Reading the Grid: Extending Knowledge Bases from Human-readable Tables
 - 27 Jelte van Waterschoot (UT), Personalized and Personal Conversations: Designing Agents Who Want to Connect With You
 - 28 Christoph Selig (UL), Understanding the Heterogeneity of Corporate Entrepreneurship Programs
-
- 2022 01 Judith van Stegeren (UT), Flavor text generation for role-playing video games
 - 02 Paulo da Costa (TU/e), Data-driven Prognostics and Logistics Optimisation: A Deep Learning Journey
 - 03 Ali el Hassouni (VUA), A Model A Day Keeps The Doctor Away: Reinforcement Learning For Personalized Healthcare
 - 04 Ünal Aksu (UU), A Cross-Organizational Process Mining Framework
 - 05 Shiwei Liu (TU/e), Sparse Neural Network Training with In-Time Over-Parameterization
 - 06 Reza Refaei Afshar (TU/e), Machine Learning for Ad Publishers in Real Time Bidding
 - 07 Sambit Praharaj (OU), Measuring the Unmeasurable? Towards Automatic Co-located Collaboration Analytics
 - 08 Maikel L. van Eck (TU/e), Process Mining for Smart Product Design
 - 09 Oana Andreea Inel (VUA), Understanding Events: A Diversity-driven Human-Machine Approach
 - 10 Felipe Moraes Gomes (TUD), Examining the Effectiveness of Collaborative Search Engines
 - 11 Mirjam de Haas (UT), Staying engaged in child-robot interaction, a quantitative approach to studying preschoolers' engagement with robots and tasks during second-language tutoring
 - 12 Guanyi Chen (UU), Computational Generation of Chinese Noun Phrases
 - 13 Xander Wilcke (VUA), Machine Learning on Multimodal Knowledge Graphs: Opportunities, Challenges, and Methods for Learning on Real-World Heterogeneous and Spatially-Oriented Knowledge
 - 14 Michiel Overeem (UU), Evolution of Low-Code Platforms
 - 15 Jelmer Jan Koorn (UU), Work in Process: Unearthing Meaning using Process Mining
 - 16 Pieter Gijbbers (TU/e), Systems for AutoML Research
 - 17 Laura van der Lubbe (VUA), Empowering vulnerable people with serious games and gamification
 - 18 Paris Mavromoustakos Blom (TiU), Player Affect Modelling and Video Game Personalisation
 - 19 Bilge Yigit Ozkan (UU), Cybersecurity Maturity Assessment and Standardisation

- 20 Fakhra Jabeen (VUA), Dark Side of the Digital Media - Computational Analysis of Negative Human Behaviors on Social Media
 - 21 Seethu Mariyam Christopher (UM), Intelligent Toys for Physical and Cognitive Assessments
 - 22 Alexandra Sierra Rativa (TiU), Virtual Character Design and its potential to foster Empathy, Immersion, and Collaboration Skills in Video Games and Virtual Reality Simulations
 - 23 Ilir Kola (TUD), Enabling Social Situation Awareness in Support Agents
 - 24 Samaneh Heidari (UU), Agents with Social Norms and Values - A framework for agent based social simulations with social norms and personal values
 - 25 Anna L.D. Latour (UL), Optimal decision-making under constraints and uncertainty
 - 26 Anne Dirkson (UL), Knowledge Discovery from Patient Forums: Gaining novel medical insights from patient experiences
 - 27 Christos Athanasiadis (UM), Emotion-aware cross-modal domain adaptation in video sequences
 - 28 Onuralp Ulusoy (UU), Privacy in Collaborative Systems
 - 29 Jan Kolkmeier (UT), From Head Transform to Mind Transplant: Social Interactions in Mixed Reality
 - 30 Dean De Leo (CWI), Analysis of Dynamic Graphs on Sparse Arrays
 - 31 Konstantinos Traganos (TU/e), Tackling Complexity in Smart Manufacturing with Advanced Manufacturing Process Management
 - 32 Cezara Pastrav (UU), Social simulation for socio-ecological systems
 - 33 Brinn Hekkelman (CWI/TUD), Fair Mechanisms for Smart Grid Congestion Management
 - 34 Nimat Ullah (VUA), Mind Your Behaviour: Computational Modelling of Emotion & Desire Regulation for Behaviour Change
 - 35 Mike E.U. Ligthart (VUA), Shaping the Child-Robot Relationship: Interaction Design Patterns for a Sustainable Interaction
-
- 2023 01 Bojan Simoski (VUA), Untangling the Puzzle of Digital Health Interventions
 - 02 Mariana Rachel Dias da Silva (TiU), Grounded or in flight? What our bodies can tell us about the whereabouts of our thoughts
 - 03 Shabnam Najafian (TUD), User Modeling for Privacy-preserving Explanations in Group Recommendations
 - 04 Gineke Wiggers (UL), The Relevance of Impact: bibliometric-enhanced legal information retrieval
 - 05 Anton Bouter (CWI), Optimal Mixing Evolutionary Algorithms for Large-Scale Real-Valued Optimization, Including Real-World Medical Applications
 - 06 António Pereira Barata (UL), Reliable and Fair Machine Learning for Risk Assessment
 - 07 Tianjin Huang (TU/e), The Roles of Adversarial Examples on Trustworthiness of Deep Learning
 - 08 Lu Yin (TU/e), Knowledge Elicitation using Psychometric Learning
 - 09 Xu Wang (VUA), Scientific Dataset Recommendation with Semantic Techniques

- 10 Dennis J.N.J. Soemers (UM), Learning State-Action Features for General Game Playing
 - 11 Fawad Taj (VUA), Towards Motivating Machines: Computational Modeling of the Mechanism of Actions for Effective Digital Health Behavior Change Applications
 - 12 Tessel Bogaard (VUA), Using Metadata to Understand Search Behavior in Digital Libraries
 - 13 Injy Sarhan (UU), Open Information Extraction for Knowledge Representation
 - 14 Selma Čaušević (TUD), Energy resilience through self-organization
 - 15 Alvaro Henrique Chaim Correia (TU/e), Insights on Learning Tractable Probabilistic Graphical Models
 - 16 Peter Blomsma (TiU), Building Embodied Conversational Agents: Observations on human nonverbal behaviour as a resource for the development of artificial characters
 - 17 Meike Nauta (UT), Explainable AI and Interpretable Computer Vision –From Oversight to Insight
 - 18 Gustavo Penha (TUD), Designing and Diagnosing Models for Conversational Search and Recommendation
 - 19 George Aalbers (TiU), Digital Traces of the Mind: Using Smartphones to Capture Signals of Well-Being in Individuals
 - 20 Arkadiy Dushatskiy (TUD), Expensive Optimization with Model-Based Evolutionary Algorithms applied to Medical Image Segmentation using Deep Learning
 - 21 Gerrit Jan de Bruin (UL), Network Analysis Methods for Smart Inspection in the Transport Domain
 - 22 Alireza Shojaifar (UU), Volitional Cybersecurity
 - 23 Theo Theunissen (UU), Documentation in Continuous Software Development
 - 24 Agathe Balayn (TUD), Practices Towards Hazardous Failure Diagnosis in Machine Learning
 - 25 Jurian Baas (UU), Entity Resolution on Historical Knowledge Graphs
 - 26 Loek Tonnaer (TU/e), Linearly Symmetry-Based Disentangled Representations and their Out-of-Distribution Behaviour
 - 27 Ghada Sokar (TU/e), Learning Continually Under Changing Data Distributions
 - 28 Floris den Hengst (VUA), Learning to Behave: Reinforcement Learning in Human Contexts
 - 29 Tim Draws (TUD), Understanding Viewpoint Biases in Web Search Results
-
- 2024 01 Daphne Miedema (TU/e), On Learning SQL: Disentangling concepts in data systems education
 - 02 Emile van Krieken (VUA), Optimisation in Neurosymbolic Learning Systems
 - 03 Feri Wijayanto (RUN), Automated Model Selection for Rasch and Mediation Analysis
 - 04 Mike Huisman (UL), Understanding Deep Meta-Learning
 - 05 Yiyong Gou (UM), Aerial Robotic Operations: Multi-environment Cooperative Inspection & Construction Crack Autonomous Repair
 - 06 Azqa Nadeem (TUD), Understanding Adversary Behavior via XAI: Leveraging Sequence Clustering to Extract Threat Intelligence

- 07 Parisa Shayan (TiU), Modeling User Behavior in Learning Management Systems
- 08 Xin Zhou (UvA), From Empowering to Motivating: Enhancing Policy Enforcement through Process Design and Incentive Implementation
- 09 Giso Dal (UT), Probabilistic Inference Using Partitioned Bayesian Networks
- 10 Cristina-Iulia Bucur (VUA), Linkflows: Towards Genuine Semantic Publishing in Science
- 11 withdrawn
- 12 Peide Zhu (TUD), Towards Robust Automatic Question Generation For Learning
- 13 Enrico Liscio (TUD), Context-Specific Value Inference via Hybrid Intelligence
- 14 Larissa Capobianco Shimomura (TU/e), On Graph Generating Dependencies and their Applications in Data Profiling
- 15 Ting Liu (VUA), A Gut Feeling: Biomedical Knowledge Graphs for Interrelating the Gut Microbiome and Mental Health
- 16 Arthur Barbosa Câmara (TUD), Designing Search-as-Learning Systems
- 17 Razieh Alidoosti (VUA), Ethics-aware Software Architecture Design
- 18 Laurens Stoop (UU), Data Driven Understanding of Energy-Meteorological Variability and its Impact on Energy System Operations
- 19 Azadeh Mozafari Mehr (TU/e), Multi-perspective Conformance Checking: Identifying and Understanding Patterns of Anomalous Behavior
- 20 Ritsart Anne Plantenga (UL), Omgang met Regels
- 21 Federica Vinella (UU), Crowdsourcing User-Centered Teams
- 22 Zeynep Ozturk Yurt (TU/e), Beyond Routine: Extending BPM for Knowledge-Intensive Processes with Controllable Dynamic Contexts
- 23 Jie Luo (VUA), Lamarck's Revenge: Inheritance of Learned Traits Improves Robot Evolution
- 24 Nirmal Roy (TUD), Exploring the effects of interactive interfaces on user search behaviour
- 25 Alisa Rieger (TUD), Striving for Responsible Opinion Formation in Web Search on Debated Topics
- 26 Tim Gubner (CWI), Adaptively Generating Heterogeneous Execution Strategies using the VOILA Framework
- 27 Lincen Yang (UL), Information-theoretic Partition-based Models for Interpretable Machine Learning
- 28 Leon Helwerda (UL), Grip on Software: Understanding development progress of Scrum sprints and backlogs
- 29 David Wilson Romero Guzman (VUA), The Good, the Efficient and the Inductive Biases: Exploring Efficiency in Deep Learning Through the Use of Inductive Biases
- 30 Vijanti Ramautar (UU), Model-Driven Sustainability Accounting
- 31 Ziyu Li (TUD), On the Utility of Metadata to Optimize Machine Learning Workflows
- 32 Vinicius Stein Dani (UU), The Alpha and Omega of Process Mining
- 33 Siddharth Mehrotra (TUD), Designing for Appropriate Trust in Human-AI interaction

-
- 34 Robert Deckers (VUA), From Smallest Software Particle to System Specification - MuDForM: Multi-Domain Formalization Method
 - 35 Sicui Zhang (TU/e), Methods of Detecting Clinical Deviations with Process Mining: a fuzzy set approach
 - 36 Thomas Mulder (TU/e), Optimization of Recursive Queries on Graphs
 - 37 James Graham Nevin (UvA), The Ramifications of Data Handling for Computational Models
 - 38 Christos Koutras (TUD), Tabular Schema Matching for Modern Settings
 - 39 Paola Lara Machado (TU/e), The Nexus between Business Models and Operating Models: From Conceptual Understanding to Actionable Guidance
 - 40 Montijn van de Ven (TU/e), Guiding the Definition of Key Performance Indicators for Business Models
 - 41 Georgios Siachamis (TUD), Adaptivity for Streaming Dataflow Engines
 - 42 Emmeke Veltmeijer (VUA), Small Groups, Big Insights: Understanding the Crowd through Expressive Subgroup Analysis
 - 43 Cedric Waterschoot (KNAW Meertens Instituut), The Constructive Conundrum: Computational Approaches to Facilitate Constructive Commenting on Online News Platforms
 - 44 Marcel Schmitz (OU), Towards learning analytics-supported learning design
 - 45 Sara Salimzadeh (TUD), Living in the Age of AI: Understanding Contextual Factors that Shape Human-AI Decision-Making
 - 46 Georgios Stathis (Leiden University), Preventing Disputes: Preventive Logic, Law & Technology
 - 47 Daniel Daza (VUA), Exploiting Subgraphs and Attributes for Representation Learning on Knowledge Graphs
 - 48 Ioannis Petros Samiotis (TUD), Crowd-Assisted Annotation of Classical Music Compositions
-
- 2025 01 Max van Haastrecht (UL), Transdisciplinary Perspectives on Validity: Bridging the Gap Between Design and Implementation for Technology-Enhanced Learning Systems
 - 02 Jurgen van den Hoogen (JADS), Time Series Analysis Using Convolutional Neural Networks
 - 03 Andra-Denis Ionescu (TUD), Feature Discovery for Data-Centric AI
 - 04 Rianne Schouten (TU/e), Exceptional Model Mining for Hierarchical Data
 - 05 Nele Albers (TUD), Psychology-Informed Reinforcement Learning for Situated Virtual Coaching in Smoking Cessation
 - 06 Daniël Vos (TUD), Decision Tree Learning: Algorithms for Robust Prediction and Policy Optimization
 - 07 Ricky Maulana Fajri (TU/e), Towards Safer Active Learning: Dealing with Unwanted Biases, Graph-Structured Data, Adversary, and Data Imbalance
 - 08 Stefan Bloemheuvel (TiU), Spatio-Temporal Analysis Through Graphs: Predictive Modeling and Graph Construction
 - 09 Fadime Kaya (VUA), Decentralized Governance Design - A Model-Based Approach

- 10 Zhao Yang (UL), Enhancing Autonomy and Efficiency in Goal-Conditioned Reinforcement Learning
- 11 Shahin Sharifi Noorian (TUD), From Recognition to Understanding: Enriching Visual Models Through Multi-Modal Semantic Integration
- 12 Lijun Lyu (TUD), Interpretability in Neural Information Retrieval
- 13 Fuda van Diggelen (VUA), Robots Need Some Education: on the complexity of learning in evolutionary robotics
- 14 Gennaro Gala (TU/e), Probabilistic Generative Modeling with Latent Variable Hierarchies
- 15 Michiel van der Meer (UL), Opinion Diversity through Hybrid Intelligence

