Facilitating Opinion Diversity through Hybrid NLP Approaches
Thesis Proposal

Michiel van der Meer
LIACS
Leiden University
m.t.van.der.meer@liacs.leidenuniv.nl

Abstract

Modern democracies face a critical issue of declining citizen participation in decision-making. Online discussion forums are an important avenue for enhancing citizen participation. This thesis proposal (1) identifies the challenges involved in facilitating large-scale online discussions with Natural Language Processing (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. We illustrate how these representations can draw insights into the diversity of perspectives and allow us to investigate interactions in online discussions.

1 Introduction

Addressing societal problems, such as climate change, pandemics, and resource scarcity, requires citizen engagement. One way to enhance citizen participation is by engaging with the public directly in society-wide conversations on online platforms (Smith, 2009; Friess and Eilders, 2015). Online discussions help identify the problem areas and possible solutions that fit the diverse needs of those affected (Surowiecki, 2004; Dryzek et al., 2019).

Online discussions generate vast amounts of content, which is challenging to manage and navigate (Dahlberg, 2001). Further, the content is scattered across time and threads, and it contains frequently repeating arguments and abundant unconnected ideas. This makes it difficult for users to know where to add new contributions, resulting in low-quality content (Klein, 2012). These issues can be addressed by employing moderators or facilitators, e.g., to structure the content of a discussion or to steer user interactions (Trénel, 2009). However, given the amount of data, manually facilitating online discussions is not feasible.

Instead, we turn to NLP for interpreting text-based opinions at scale (Sun et al., 2017), powered by the recent surge of Large Language Models (LLMs) (Min et al., 2023; Argyle et al., 2023). Central to our approach to facilitation is extracting structured *perspectives* from users in a discussion. The perspectives provide high-level insights into the arguments employed by citizens (Vecchi et al., 2021) or the motivations underlying the opinions in a community (Weld et al., 2022). These representations may, in turn, influence the facilitation strategies (Falk et al., 2021) or shape policies following the discussion (Mouter et al., 2021).

Using NLP for analyzing opinions sourced from online platforms comes with its own set of challenges. For instance, online platforms have been centered on managing large volumes of information, e.g., through personalized recommendations (Adomavicius and Tuzhilin, 2005) or argument structuring (Iandoli et al., 2014) but have neglected inclusive design aspects (Shortall et al., 2022). This can cause majority opinions to be heard while suppressing dissent voices (Neubaum and Krämer, 2017). Similarly, we see that LLMs capture majority opinions well, but do not distill all voices equally (e.g., Mustafaraj et al., 2011; van der Meer et al., 2024c). Further, LLMs lack deep social reasoning (Liang et al., 2021), may be biased (Hartmann et al., 2023; Santurkar et al., 2023), and make mistakes in ways humans cannot anticipate (Huang et al., 2023). Finally, straightforward automated discussion analysis runs the danger of ignoring diverse opinions, which undermines the wisdom of the crowd effect (Lorenz et al., 2011). In light of these challenges, we ask our first research question:

Q1 *What fundamental issues arise in using NLP to analyze perspectives in online discussions?*

Next, our goal is to obtain structured information

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

from online societal discussions that provide insights into the opinions involved. However, we see that NLP-based methods for analyzing online deliberation are limited in the degree to which **diverse** perspectives can be obtained. To combat these limitations, we develop an approach that adopts a "hybrid" mindset, i.e., incorporates humans-in-the-loop to address diversity directly. We leverage LLMs and humans jointly, with their different 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, analyzing opinions, in practice, is modeled by different tasks. 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. Higher-order abstractions, such as personal values, deeply motivate choices and the attitudes of individuals but are difficult to estimate automatically. Conversely, surface-level stance recognition tasks are more widely applicable but uncover little information about an individual's opinion. 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 combine different tasks for representing diverse opinions in online discussions?*

Sections 2, 3, and 4 describe our progress on the three questions. Section 5 concludes the paper.

2 Use of NLP in Societal Discussions

> **Q1** What fundamental issues arise in using NLP to analyze perspectives in online discussions?

NLP research regarding the facilitation of online societal discussions has seen recent interest (e.g., Crossley et al., 2016; Jelodar et al., 2020; Xia et al., 2020). Research is focused on (1) using NLP tools, in particular few-shot prompted LLMs, to analyze the discussions (e.g., Xia et al., 2020; Syed et al., 2023), and (2) using discussion data to benchmark the capabilities of NLP tools (e.g., Feng et al., 2023). In the next two sections, we outline related work in these directions, highlighting fundamental issues that cross-cut techniques and applications.

2.1 Discussion Analysis

Online social interaction through text is common, and the use of NLP for analyzing large amounts of such data is mainstream (Liu, 2012). Discussions happen in various specific contexts, e.g., reviews (Jo and Oh, 2011) or e-learning (Davies and Graff, 2005), but also broader contemporary topics such as climate change (Lörcher and Taddicken, 2017). Their scale, combined with their pertinence makes analyzing such discussions interesting.

Analyzing how humans express themselves through text is the core task in many NLP areas, e.g., Opinion Summarization (Liu, 2012), Argument Mining (Lawrence and Reed, 2020), Sentiment Analysis (Wankhade et al., 2022), and Value Classification (Hoover et al., 2020). These tasks lie at the heart of creating insights into online (political) discourse and may be used e.g. for estimating the quality of discussions (Steenbergen et al., 2003), extracting the arguments involved (Lapesa et al., 2023), or reasoning over inconsistencies between choices and their justifications (Liscio et al., 2024). In the age of LLMs, these tasks have seen considerable performance improvements (Jiang et al., 2023), though new challenges such as dealing with shortcut learning (Geirhos et al., 2020) or mitigating social biases (Liang et al., 2021) arise.

Extracting diverse views from online discussions is challenging for three reasons. First, data sourced from social media platforms inherits biases that are present on these platforms, including fake news, trolling, and polarization (Cinelli et al., 2021). This impacts how opinions are shaped (Hocevar et al., 2014) and the distribution of opinions (Xiong and Liu, 2014). Second, when analyzing the opinions about societal issues, it is necessary to realize that not all citizens have equal access due to the digital divide (Cullen, 2001) or differences in tech-illiteracy (Knobel and Lankshear, 2008). This makes the users in online discussions biased and less diverse. Third, since users are free to join in discussions of their choosing, there may be undesired echo chambers or self-selection effects among the messages seen by users (Song et al., 2020).

Despite these challenges, we can use NLP to investigate questions about human behavior at scale (Lazer et al., 2009). Analyses about behavior may lead to insights on both individual and group levels. This can be useful for improving democratic processes (Collins and Nerlich, 2019), but also applies in other areas, such as faithfully interpreting product feedback (Bar-Haim et al., 2021), service improvement (Skiera et al., 2022), or course management (Lin et al., 2009).

2.2 Benchmarking

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 (Durmus et al., 2024; Hendrycks et al., 2021; van der Meer et al., 2024c). Whether LLMs can learn to represent them in the future remains unclear (Wei et al., 2022; Schaeffer et al., 2023), but research suggests that they cannot (Feng et al., 2023; Argyle et al., 2023), in part due to the limitations mentioned in Section 2.1. Therefore, we work with the assumption that this is a fundamental limitation of LLMs, and we have to find other approaches to 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 (Laban et al., 2022; Wu et al., 2020) for shaping an individual's perspective (Bakshy et al., 2015). Others strive to make LLMs better represent a diverse group of annotators based on their labeling behavior and demographics (Bakker et al., 2022; Lahoti et al., 2023). In such approaches, models have a large reliance on annotated data. Labels are obtained from a few human annotators per instance, and often aggregated by majority voting, painting an incomplete picture of the true range of interpretations for a potentially controversial text (Plank, 2022). The role of subjectivity in these tasks remains unclear (Aroyo and Welty, 2015; Cabitza et al., 2023). This holds for traditional supervised learning, but also for the latest trends in instruction-tuning (Uma et al., 2021; Wang et al., 2023).

In the rest of this proposal, we argue that the aforementioned challenges can be overcome by using LLMs to **assist humans** in mining opinionated text data rather than replacing them, and we provide an example of how hybrid approaches can uncover perspectives of the opinion holders.

[1]Although linguistic diversity generally refers to diversity of language proficiencies (Joshi et al., 2020; Dingemanse and Liesenfeld, 2022), we are specifically interested in diversity in arguments, communication styles, and values in online discussions.

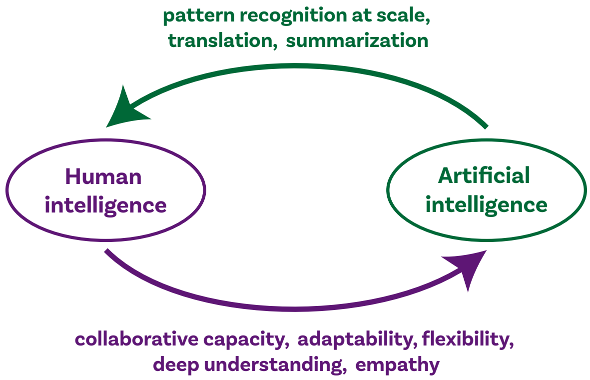

Figure 1: Feedback loops in Hybrid Intelligence.

3 Hybrid Intelligence

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

Central to our proposal on facilitating deliberation is the notion of *hybrid intelligence* (Dellermann et al., 2019; Akata et al., 2020; Dell'Anna et al., 2024). In Hybrid Intelligent Systems (HISs), artificial intelligence is a collaborator that enhances human abilities such as reasoning, decision-making, and problem-solving (Tiddi et al., 2023). 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 LLMs, highlighting bidirectional gains, as shown in Figure 1.

3.1 Related Work

NLP has had a profound impact on how researchers analyze human behavior at scale. To do so responsibly, we must ensure that these methods do so effectively while upholding democratic values. Previous work on hybrid approaches for NLP includes user adaptation (Lynn et al., 2017), human-in-the-loop computing (Wang et al., 2021), human-AI interaction (Heer, 2019) and others (e.g., Ding et al., 2023; Team et al., 2022). Recent interest in explainable AI has focused on human understanding of NLP models (Lertvittayakumjorn and Toni, 2021). Specifically for NLP, much focus is on approaches that mix crowd, expert, and automated decision-making, which have been applied to analyzing discussion content (Kong et al., 2022; Pacheco et al., 2023). However, these approaches have a one-way interaction between the NLP model and humans, as we will describe in the next section.

3.2 Approach

We observe that LLMs still have many challenges to overcome in representing diverse perspectives (Section 2). 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 power users, moderators, or even site admins (Saxena and Reddy, 2022). They provide text and behavioral data, such as post-voting, 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. Much room for making these procedures more informative exist, such as expanding the label set (van de Ven et al., 2022), including free-form text response (Ouyang et al., 2023), asking a crowd of annotators rather than individuals (Nie et al., 2020), and more (e.g., Plank, 2022; Santy et al., 2023).

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 (Tomašev et al., 2020). First, NLP may enable, rather than restrict, access to certain services, for example by 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 consider each sample identically. 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. 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 Fig. 1, iteratively refining the analysis procedure while performing discussion analyses. By building on the bidirectional contributions, we allow for continual improvement.

Our work involves discussion analysis approaches that involve (1) selecting samples for human inspection that are interesting to annotate, (2) accounting for diversity (e.g., leveraging contextualized embeddings (Reimers and Gurevych, 2019)), (3) seeking labels from multiple annotators. We find that a hybrid approach can capture more diverse interpretations of the arguments in a discussion than a purely manual or purely automatic approach (van der Meer et al., 2022, 2024b). When extracting arguments from online comments, human annotators are more precise than NLP methods. At the same time, we use sampling based on the maximum embedding distance to ensure diverse content is observed (Basu et al., 2004) and automatically merge similar arguments (Chai et al., 2016). In this setup, we obtain labels from a crowd over diverse samples that promote perspective-taking. 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. 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 (van der Meer et al., 2024a).

Developing hybrid approaches requires a new evaluation paradigm. We need to compare our method's effectiveness with human-only and machine-only baselines. In NLP, test sets are usually collected manually. This may make the upper bound on performance unfair, though performance gaps between hybrid and manual approaches can be addressed (Xu et al., 2023; Fluri et al., 2023).

4 Perspective Hierarchy

Q3 How to combine different tasks for representing diverse opinions in online discussions?

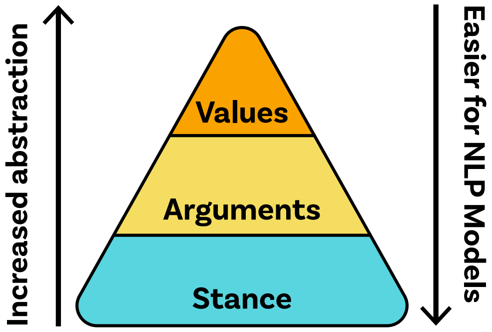

Figure 2: The perspective hierarchy. The higher the level of abstraction, the more human intelligence is required for interpreting the component.

Given that NLP can process large amounts of discussion data, but is limited in its capabilities (Section 2), and that we may construct hybrid procedures to account for these limits (Section 3), we address the challenge on how to capture perspectives. Uncovering them from online societal discussions requires a representation for identifying how people feel about potential decisions, how this is communicated in the discussions, and what their underlying motivations are.

4.1 Related Work

Few attempts to represent perspectives holistically exist (Chen et al., 2019; van Son et al., 2016). These works focus on annotating utterances for low-level claim information (Morante et al., 2020), or investigating some of the reasoning behind the views held in discussions (Draws et al., 2022). Stances and arguments are inherently linked in argumentation models (Toulmin, 2003; Van Eemeren et al., 2015), and form the basis of frameworks for representing perspectives (Wiebe et al., 2005; Chen et al., 2022).

However, neither stance nor arguments aim to represent opinions on a deeper personal level. A fundamental concept for explaining the motivations underlying opinions and actions is personal values (Schwartz, 2012). There are various theories of personal values (e.g., Rokeach, 1967; Schwartz, 2012; Graham et al., 2013). Preferences among values describe the attitude of individuals and groups (Ponizovskiy et al., 2020), and can be extracted from behavioral cues to investigate political affiliation (Roy et al., 2021), perform moral reasoning (Mooijman et al., 2018), or positively influence lifestyle (de Boer et al., 2023). Values are abstract and need to be interpreted inside their context, making it difficult for both humans and NLP methods to reliably measure them (Liscio et al., 2023). One way to contextualize them is to connect values to argumentation, focusing on how choices are justified and reasoned over (Kiesel et al., 2022). Using this insight, we incorporate personal values into our perspective representation and aim to obtain them using a hybrid approach.

4.2 Approach

We propose a perspective hierarchy to represent a person's perspective at different levels of abstraction, shown in Figure 2. Our perspective hierarchy is composed of stances, arguments, and values.

Stance Whether, or how much, support or opposition is expressed to a claim. Stance detection has been studied extensively and remains a popular task for investigating opinions on claims (Küçük and Can, 2020).
Arguments The reasons given for adopting a stance towards a claim. In real-world contexts, argumentation manifests in many forms and is predominantly informal (Groarke, 2024). Mining arguments from text works well within known contexts (Ein-Dor et al., 2020), but suffers from implicit reasoning (Habernal et al., 2018). Hence, we require more human guidance to correct for possible mistakes in automated methods.
Values The motivations underlying opinions and actions (Schwartz, 2012). Values are communicated implicitly through actions or written motivations. Estimating values automatically remains difficult even within their context (Kiesel et al., 2023). Only through iterative hybrid procedures can we accurately reason about preferences among human values.

Mining Perspective Hierarchies We illustrate how we used data from large online social media platforms to investigate perspective hierarchies for individuals (van der Meer et al., 2023). 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 (Pougué-Biyong et al., 2021), 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 to answer questions about human behavior.

We probed the relationship between disagreements in stance and deeper conflicts in values. Our experiments show that when values are diverse, conflicts in values can correlate to stance disagreement. Based on purely automated estimations, this evidence is weak. When we incorporate human-provided self-reports, the evidence becomes stronger, showing that the hybrid approach is crucial to performing a meaningful analysis. On the other hand, when strong value diversity is absent, we cannot correlate disagreement and value conflict directly. Thus, we require a more complete picture, and should therefore incorporate the arguments to complete the perspective hierarchy.

5 Conclusions

We identified the strengths and weaknesses of using NLP to represent diverse perspectives in online societal discussions. NLP techniques, in particular few-shot prompting with LLMs, allow us to analyze discussion data for perspectives at a large scale. However, open challenges include (1) a difficulty in acquiring opinions from diverse opinion holders, and (2) limitations of LLMs to represent minority opinions. Our approach combines the complementary abilities of humans and LLMs into hybrid intelligence methods to obtain better analyses than automated or manual analysis alone. We propose a perspective hierarchy to guide the investigation of human behavior in online societal discussions at scale. We find that this hierarchy is useful for uncovering perspectives, for instance, in observing that diversity in opinions can be signaled by differences among value preferences.

Future Directions

First, integrating human and artificial work requires careful task balancing. In some cases, obtaining an automated judgment from an LLM is sufficient, but in others, we need to query a pool of diverse human annotators. We can use frameworks like learning to defer (Madras et al., 2018) or other active learning approaches (Baumler et al., 2023) to directly obtain diverse opinions (Waterschoot et al., 2022).

Second, evaluation of hybrid intelligence systems 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 (Chang et al., 2024). Alternative approaches can instead incorporate interactive crowd-sourced benchmarks that develop over time (Kiela et al., 2021), or turn to use-case-specific evaluation, leveraging objective behavioral cues to assess our methods, e.g., in measuring interaction structure to reveal the quality of a conversation (Santamaría et al., 2022).

Lastly, our proposed hybrid human-AI approach engages with citizens to learn their perspectives. We represent the cares, incentives, and preferences of those involved in societal discussions. In the long run, we may be able to adopt components in the perspective hierarchy for not only facilitating discussions but supporting negotiations (Renting et al., 2022) among societal stakeholders, e.g., on which portfolio of choices to make to combat a pandemic (Mouter et al., 2021).

References

G. Adomavicius and A. Tuzhilin. 2005. 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.

Zeynep Akata, Dan Balliet, Maarten de Rijke, Frank Dignum, Virginia Dignum, Guszti Eiben, Antske Fokkens, Davide Grossi, Koen Hindriks, Holger Hoos, Hayley Hung, Catholijn Jonker, Christof Monz, Mark Neerincx, Frans Oliehoek, Henry Prakken, Stefan Schlobach, Linda van der Gaag, Frank van Harmelen, Herke van Hoof, Birna van Riemsdijk, Aimee van Wynsberghe, Rineke Verbrugge, Bart Verheij, Piek Vossen, and Max Welling. 2020. A research agenda for hybrid intelligence: Augmenting human intellect with collaborative, adaptive, responsible, and explainable artificial intelligence. *Computer*, 53(8):18–28.

Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. 2023. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351.

Lora Aroyo and Chris Welty. 2015. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1):15–24.

Michiel Bakker, Martin Chadwick, Hannah Sheahan, Michael Tessler, Lucy Campbell-Gillingham, Jan

Balaguer, Nat McAleese, Amelia Glaese, John Aslanides, Matt Botvinick, and Christopher Summerfield. 2022. Fine-tuning language models to find agreement among humans with diverse preferences. In *Advances in Neural Information Processing Systems*, volume 35, pages 38176–38189. Curran Associates, Inc.

Eytan Bakshy, Solomon Messing, and Lada A. Adamic. 2015. Exposure to ideologically diverse news and opinion on facebook. *Science*, 348(6239):1130–1132.

Roy Bar-Haim, Lilach Eden, Yoav Kantor, Roni Friedman, and Noam Slonim. 2021. Every bite is an experience: Key Point Analysis of business reviews. 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 3376–3386, Online. Association for Computational Linguistics.

Sugato Basu, Arindam Banerjee, and Raymond J. Mooney. 2004. Active semi-supervision for pairwise constrained clustering. In *Proceedings of the 2004 SIAM International Conference on Data Mining (SDM)*, pages 333–344.

Connor Baumler, Anna Sotnikova, and Hal Daumé III. 2023. 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, Toronto, Canada. Association for Computational Linguistics.

Federico Cabitza, Andrea Campagner, and Valerio Basile. 2023. Toward a perspectivist turn in ground truthing for predictive computing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(6):6860–6868.

Chengliang Chai, Guoliang Li, Jian Li, Dong Deng, and Jianhua Feng. 2016. Cost-effective crowdsourced entity resolution: A partial-order approach. In *Proceedings of the 2016 International Conference on Management of Data*, SIGMOD '16, page 969–984, New York, NY, USA. Association for Computing Machinery.

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. 2024. A survey on evaluation of large language models. *ACM Trans. Intell. Syst. Technol.*, 15(3).

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

Sihao Chen, Siyi Liu, Xander Uyttendaele, Yi Zhang, William Bruno, and Dan Roth. 2022. Design challenges for a multi-perspective search engine. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 293–303, Seattle, United States. Association for Computational Linguistics.

Matteo Cinelli, Gianmarco De Francisci Morales, Alessandro Galeazzi, Walter Quattrociocchi, and Michele Starnini. 2021. The echo chamber effect on social media. *Proceedings of the National Academy of Sciences*, 118(9):e2023301118.

Luke Collins and Brigitte Nerlich. 2019. 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.

Scott Crossley, Luc Paquette, Mihai Dascalu, Danielle S. McNamara, and Ryan S. Baker. 2016. 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, page 6–14, New York, NY, USA. Association for Computing Machinery.

Rowena Cullen. 2001. Addressing the digital divide. *Online information review*, 25(5):311–320.

Lincoln Dahlberg. 2001. The internet and democratic discourse: Exploring the prospects of online deliberative forums extending the public sphere. *Information, communication & society*, 4(4):615–633.

Jo Davies and Martin Graff. 2005. Performance in e-learning: online participation and student grades. *British Journal of Educational Technology*, 36(4):657–663.

Maaike H de Boer, Jasper van der Waa, Sophie van Gent, Quirine T.S. Smit, Wouter Korteling, Robin M. van Stokkum, and Mark Neerincx. 2023. A contextual hybrid intelligent system design for diabetes lifestyle management. In *International Workshop Modelling and Representing Context, ECAI*, volume 23.

Davide Dell'Anna, Pradeep K. Murukannaiah, Bernd Dudzik, Davide Grossi, Catholijn M. Jonker, Catharine Oertel, and Pınar Yolum. 2024. 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. To appear.

Dominik Dellermann, Philipp Ebel, Matthias Söllner, and Jan Marco Leimeister. 2019. Hybrid intelligence. *Business & Information Systems Engineering*, 61:637–643.

Zijian Ding, Alison Smith-Renner, Wenjuan Zhang, Joel Tetreault, and Alejandro Jaimes. 2023. Harnessing the power of LLMs: Evaluating human-AI text co-creation through the lens of news headline generation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3321–3339, Singapore. Association for Computational Linguistics.

Mark Dingemanse and Andreas Liesenfeld. 2022. From text to talk: Harnessing conversational corpora for humane and diversity-aware language technology. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5614–5633, Dublin, Ireland. Association for Computational Linguistics.

Tim Draws, Oana Inel, Nava Tintarev, Christian Baden, and Benjamin Timmermans. 2022. 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, page 135–145, New York, NY, USA. Association for Computing Machinery.

John S. Dryzek, André Bächtiger, Simone Chambers, Joshua Cohen, James N. Druckman, Andrea Felicetti, James S. Fishkin, David M. Farrell, Archon Fung, Amy Gutmann, Hélène Landemore, Jane Mansbridge, Sofie Marien, Michael A. Neblo, Simon Niemeyer, Maija Setälä, Rune Slothuus, Jane Suiter, Dennis Thompson, and Mark E. Warren. 2019. The crisis of democracy and the science of deliberation. *Science*, 363(6432):1144–1146.

Esin Durmus, Karina Nguyen, Thomas I. Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, Liane Lovitt, Sam McCandlish, Orowa Sikder, Alex Tamkin, Janel Thamkul, Jared Kaplan, Jack Clark, and Deep Ganguli. 2024. Towards measuring the representation of subjective global opinions in language models.

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. 2020. Corpus wide argument mining—a working solution. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7683–7691.

Neele Falk, Iman Jundi, Eva Maria Vecchi, and Gabriella Lapesa. 2021. Predicting moderation of deliberative arguments: Is argument quality the key? In *Proceedings of the 8th Workshop on Argument Mining*, pages 133–141, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Shangbin Feng, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. 2023. From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair NLP models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11737–11762, Toronto, Canada. Association for Computational Linguistics.

Lukas Fluri, Daniel Paleka, and Florian Tramèr. 2023. Evaluating superhuman models with consistency checks. In *Socially Responsible Language Modelling Research*.

Dennis Friess and Christiane Eilders. 2015. A systematic review of online deliberation research. *Policy & Internet*, 7(3):319–339.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673.

Jesse Graham, Jonathan Haidt, Sena Koleva, Matt Motyl, Ravi Iyer, Sean P. Wojcik, and Peter H. Ditto. 2013. Chapter two - 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. Academic Press.

Leo Groarke. 2024. Informal Logic. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*, Spring 2024 edition. Metaphysics Research Lab, Stanford University.

Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. 2018. The argument reasoning comprehension task: Identification and reconstruction of implicit warrants. 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 1930–1940, New Orleans, Louisiana. Association for Computational Linguistics.

Jochen Hartmann, Jasper Schwenzow, and Maximilian Witte. 2023. The political ideology of conversational ai: Converging evidence on chatgpt's proenvironmental, left-libertarian orientation.

Jeffrey Heer. 2019. Agency plus automation: Designing artificial intelligence into interactive systems. *Proceedings of the National Academy of Sciences*, 116(6):1844–1850.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2021. Aligning AI with shared human values. In *International Conference on Learning Representations*.

Kristin Page Hocevar, Andrew J. Flanagin, and Miriam J. Metzger. 2014. Social media self-efficacy and information evaluation online. *Computers in Human Behavior*, 39:254–262.

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. 2020. Moral foundations twitter corpus: A collection of 35k tweets annotated for moral sentiment. *Social Psychological and Personality Science*, 11(8):1057–1071.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions.

Luca Iandoli, Ivana Quinto, Anna De Liddo, and Simon Buckingham Shum. 2014. Socially augmented argumentation tools: Rationale, design and evaluation of a debate dashboard. *International Journal of Human-Computer Studies*, 72(3):298–319.

Hamed Jelodar, Yongli Wang, Rita Orji, and Shucheng Huang. 2020. 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.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Yohan Jo and Alice H. Oh. 2011. 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, page 815–824, New York, NY, USA. Association for Computing Machinery.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

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. 2021. Dynabench: Rethinking benchmarking in NLP. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4110–4124, Online. Association for Computational Linguistics.

Johannes Kiesel, Milad Alshomary, Nicolas Handke, Xiaoni Cai, Henning Wachsmuth, and Benno Stein. 2022. Identifying the human values behind arguments. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4459–4471, Dublin, Ireland. Association for Computational Linguistics.

Johannes Kiesel, Milad Alshomary, Nailia Mirzakhmedova, Maximilian Heinrich, Nicolas Handke, Henning Wachsmuth, and Benno Stein. 2023. SemEval-2023 task 4: ValueEval: Identification of human values behind arguments. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2287–2303, Toronto, Canada. Association for Computational Linguistics.

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

Michele Knobel and Colin Lankshear. 2008. Digital literacy and participation in online social networking spaces. *Digital literacies: Concepts, policies and practices*, 11:249–278.

Quyu Kong, Emily Booth, Francesco Bailo, Amelia Johns, and Marian-Andrei Rizoiu. 2022. Slipping to the extreme: A mixed method to explain how extreme opinions infiltrate online discussions. *Proceedings of the International AAAI Conference on Web and Social Media*, 16(1):524–535.

Dilek Küçük and Fazli Can. 2020. Stance detection: A survey. *ACM Comput. Surv.*, 53(1).

Philippe Laban, Chien-Sheng Wu, Lidiya Murakhovs'ka, Xiang Chen, and Caiming Xiong. 2022. Discord questions: A computational approach to diversity analysis in news coverage. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5180–5194, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Preethi Lahoti, Nicholas Blumm, Xiao Ma, Raghavendra Kotikalapudi, Sahitya Potluri, Qijun Tan, Hansa Srinivasan, Ben Packer, Ahmad Beirami, Alex Beutel, and Jilin Chen. 2023. Improving diversity of demographic representation in large language models via collective-critiques and self-voting. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10383–10405, Singapore. Association for Computational Linguistics.

Gabriella Lapesa, Eva Maria Vecchi, Serena Villata, and Henning Wachsmuth. 2023. Mining, assessing, and improving arguments in NLP and the social sciences. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts*, pages 1–6, Dubrovnik, Croatia. Association for Computational Linguistics.

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

David Lazer, Alex Pentland, Lada Adamic, Sinan Aral, Albert-László Barabási, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, Tony Jebara, Gary King, Michael Macy, Deb Roy, and Marshall Van Alstyne. 2009. Computational social science. *Science*, 323(5915):721–723.

Piyawat Lertvittayakumjorn and Francesca Toni. 2021. Explanation-Based Human Debugging of NLP Models: A Survey. *Transactions of the Association for Computational Linguistics*, 9:1508–1528.

Paul Pu Liang, Chiyu Wu, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2021. Towards understanding and mitigating social biases in language models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 6565–6576. PMLR.

Fu-Ren Lin, Lu-Shih Hsieh, and Fu-Tai Chuang. 2009. Discovering genres of online discussion threads via text mining. *Computers & Education*, 52(2):481–495.

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. 2023. 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. International Foundation for Autonomous Agents and Multiagent Systems.

Enrico Liscio, Luciano C. Siebert, Catholijn M. Jonker, and Pradeep K. Murukannaiah. 2024. Value preferences estimation and disambiguation in hybrid participatory systems.

Bing Liu. 2012. *Sentiment Analysis and Opinion Mining*. Springer International Publishing.

Ines Lörcher and Monika Taddicken. 2017. Discussing climate change online. topics and perceptions in online climate change communication in different online public arenas. *Journal of Science Communication*, 16(2):A03.

Jan Lorenz, Heiko Rauhut, Frank Schweitzer, and Dirk Helbing. 2011. How social influence can undermine the wisdom of crowd effect. *Proceedings of the National Academy of Sciences*, 108(22):9020–9025.

Veronica Lynn, Youngseo Son, Vivek Kulkarni, Niranjan Balasubramanian, and H. Andrew Schwartz. 2017. Human centered NLP with user-factor adaptation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1146–1155, Copenhagen, Denmark. Association for Computational Linguistics.

David Madras, Toni Pitassi, and Richard Zemel. 2018. Predict responsibly: Improving fairness and accuracy by learning to defer. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz, and Dan Roth. 2023. Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Comput. Surv.*, 56(2).

Marlon Mooijman, Joe Hoover, Ying Lin, Heng Ji, and Morteza Dehghani. 2018. Moralization in social networks and the emergence of violence during protests. *Nature human behaviour*, 2(6):389–396.

Roser Morante, Chantal van Son, Isa Maks, and Piek Vossen. 2020. Annotating perspectives on vaccination. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4964–4973, Marseille, France. European Language Resources Association.

Niek Mouter, Jose Ignacio Hernandez, and Anatol Valerian Itten. 2021. Public participation in crisis policymaking. how 30,000 dutch citizens advised their government on relaxing covid-19 lockdown measures. *PLOS ONE*, 16(5):1–42.

Eni Mustafaraj, Samantha Finn, Carolyn Whitlock, and Panagiotis T. Metaxas. 2011. Vocal minority versus silent majority: Discovering the opionions 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.

German Neubaum and Nicole C. Krämer. 2017. Monitoring the opinion of the crowd: Psychological mechanisms underlying public opinion perceptions on social media. *Media Psychology*, 20(3):502–531.

Yixin Nie, Xiang Zhou, and Mohit Bansal. 2020. What can we learn from collective human opinions on natural language inference data? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9131–9143, Online. Association for Computational Linguistics.

Siru Ouyang, Shuohang Wang, Yang Liu, Ming Zhong, Yizhu Jiao, Dan Iter, Reid Pryzant, Chenguang Zhu, Heng Ji, and Jiawei Han. 2023. The shifted and the overlooked: A task-oriented investigation of user-GPT interactions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2375–2393, Singapore. Association for Computational Linguistics.

Maria Leonor Pacheco, Tunazzina Islam, Lyle Ungar, Ming Yin, and Dan Goldwasser. 2023. Interactive concept learning for uncovering latent themes in large text collections. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5059–5080, Toronto, Canada. Association for Computational Linguistics.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Vladimir Ponizovskiy, Murat Ardag, Lusine Grigoryan, Ryan Boyd, Henrik Dobewall, and Peter Holtz. 2020. 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.

John Pougué-Biyong, Valentina Semenova, Alexandre Matton, Rachel Han, Aerin Kim, Renaud Lambiotte, and Doyne Farmer. 2021. DEBAGREEMENT: 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)*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Bram M. Renting, Holger H. Hoos, and Catholijn M. Jonker. 2022. 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, page 1101–1109, Richland, SC. International Foundation for Autonomous Agents and Multiagent Systems.

Milton Rokeach. 1967. Rokeach value survey. *The nature of human values.*

Shamik Roy, Maria Leonor Pacheco, and Dan Goldwasser. 2021. Identifying morality frames in political tweets using relational learning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9939–9958, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Selene Báez Santamaría, Piek Vossen, and Thomas Baier. 2022. Evaluating agent interactions through episodic knowledge graphs. In *Proceedings of the 1st Workshop on Customized Chat Grounding Persona and Knowledge*, pages 15–28, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 29971–30004. PMLR.

Sebastin Santy, Jenny Liang, Ronan Le Bras, Katharina Reinecke, and Maarten Sap. 2023. NLPositionality: Characterizing design biases of datasets and models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9080–9102, Toronto, Canada. Association for Computational Linguistics.

Akrati Saxena and Harita Reddy. 2022. Users roles identification on online crowdsourced Q& platforms and encyclopedias: a survey. *Journal of Computational Social Science*, 5(1):285–317.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. 2023. Are emergent abilities of large language models a mirage? In *Advances in Neural Information Processing Systems*, volume 36, pages 55565–55581. Curran Associates, Inc.

Shalom H Schwartz. 2012. An overview of the schwartz theory of basic values. *Online readings in Psychology and Culture*, 2(1):11.

Ruth Shortall, Anatol Itten, Michiel van der Meer, Pradeep Murukannaiah, and Catholijn Jonker. 2022. Reason against the machine? Future directions for mass online deliberation. *Frontiers in Political Science*.

Bernd Skiera, Shunyao Yan, Johannes Daxenberger, Marcus Dombois, and Iryna Gurevych. 2022. Using information-seeking argument mining to improve service. *Journal of Service Research*, 25(4):537–548.

G. Smith. 2009. *Democratic Innovations: Designing Institutions for Citizen Participation*. Theories of Institutional Design. Cambridge University Press.

Hyunjin Song, Jaeho Cho, and Grace A. Benefield. 2020. The dynamics of message selection in online political discussion forums: Self-segregation or diverse exposure? *Communication Research*, 47(1):125–152.

Marco R Steenbergen, André Bächtiger, Markus Spörndli, and Jürg Steiner. 2003. Measuring political deliberation: A discourse quality index. *Comparative European Politics*, 1:21–48.

Shiliang Sun, Chen Luo, and Junyu Chen. 2017. A review of natural language processing techniques for opinion mining systems. *Information Fusion*, 36:10–25.

James Surowiecki. 2004. *The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business, economies, societies, and nations.* Doubleday.

Shahbaz Syed, Dominik Schwabe, Khalid Al-Khatib, and Martin Potthast. 2023. Indicative summarization of long discussions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2752–2788, Singapore. Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Ilaria Tiddi, Victor de Boer, Stefan Schlobach, and André Meyer-Vitali. 2023. Knowledge engineering for hybrid intelligence. In *Proceedings of the 12th Knowledge Capture Conference 2023*, K-CAP '23, page 75–82, Pensacola, FL, USA,. Association for Computing Machinery.

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. 2020. AI for social good: unlocking the opportunity for positive impact. *Nature Communications*, 11(1):2468.

S.E. Toulmin. 2003. *The Uses of Argument*. Cambridge University Press.

Matthias Trénel. 2009. Facilitation and inclusive deliberation. *Online deliberation: Design, research, and practice*, pages 253–257.

Alexandra N Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2021. Learning from disagreement: A survey. *Journal of Artificial Intelligence Research*, 72:1385–1470.

Gido M van de Ven, Tinne Tuytelaars, and Andreas S Tolias. 2022. Three types of incremental learning. *Nature Machine Intelligence*, 4(12):1185–1197.

Michiel van der Meer, Neele Falk, Pradeep K. Murukannaiah, and Enrico Liscio. 2024a. Annotator-centric active learning for subjective NLP tasks.

Michiel van der Meer, Enrico Liscio, Catholijn M. Jonker, Aske Plaat, Piek Vossen, and Pradeep K. Murukannaiah. 2022. HyEnA: A Hybrid Method for Extracting Arguments from Opinions. In *Proceedings of the first International Conference on Hybrid Human-Artificial Intelligence (HHAI 2022)*, pages 1–15, Amsterdam, the Netherlands. IOS Press.

Michiel van der Meer, Enrico Liscio, Catholijn M Jonker, Aske Plaat, Piek Vossen, and Pradeep K Murukannaiah. 2024b. A hybrid intelligence method for argument mining. *Journal of AI Research (JAIR, to appear)*.

Michiel van der Meer, Piek Vossen, Catholijn Jonker, and Pradeep 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.

Michiel van der Meer, Piek Vossen, Catholijn Jonker, and Pradeep Murukannaiah. 2024c. 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.

Frans H Van Eemeren, Frans H van Eemeren, Sally Jackson, and Scott Jacobs. 2015. Argumentation. *Reasonableness and effectiveness in argumentative discourse: Fifty contributions to the development of Pragma-dialectics*, pages 3–25.

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

Eva Maria Vecchi, Neele Falk, Iman Jundi, and Gabriella Lapesa. 2021. Towards argument mining for social good: A survey. 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 1338–1352, Online. Association for Computational Linguistics.

Yizhong Wang, Hamish Ivison, Pradeep Dasigi, Jack Hessel, Tushar Khot, Khyathi Chandu, David Wadden, Kelsey MacMillan, Noah A Smith, Iz Beltagy, and Hannaneh Hajishirzi. 2023. How far can camels go? exploring the state of instruction tuning on open resources. In *Advances in Neural Information Processing Systems*, volume 36, pages 74764–74786. Curran Associates, Inc.

Zijie J. Wang, Dongjin Choi, Shenyu Xu, and Diyi Yang. 2021. Putting humans in the natural language processing loop: A survey. In *Proceedings of the First Workshop on Bridging Human–Computer Interaction and Natural Language Processing*, pages 47–52, Online. Association for Computational Linguistics.

Mayur Wankhade, Annavarapu Chandra Sekhara Rao, and Chaitanya Kulkarni. 2022. A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review*, 55(7):5731–5780.

Cedric Waterschoot, Ernst van den Hemel, and Antal van den Bosch. 2022. Detecting minority arguments for mutual understanding: A moderation tool for the online climate change debate. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6715–6725, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models. *Transactions on Machine Learning Research*. Survey Certification.

Galen Weld, Amy X. Zhang, and Tim Althoff. 2022. 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.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language resources and evaluation*, 39:165–210.

Chuhan Wu, Fangzhao Wu, Tao Qi, and Yongfeng Huang. 2020. SentiRec: Sentiment diversity-aware neural news recommendation. In *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. Association for Computational Linguistics.

Yan Xia, Haiyi Zhu, Tun Lu, Peng Zhang, and Ning Gu. 2020. Exploring antecedents and consequences of toxicity in online discussions: A case study on reddit. *Proc. ACM Hum.-Comput. Interact.*, 4(CSCW2).

Fei Xiong and Yun Liu. 2014. Opinion formation on social media: An empirical approach. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 24(1):013130.

Qiongkai Xu, Christian Walder, and Chenchen Xu. 2023. Humanly certifying superhuman classifiers. In *The Eleventh International Conference on Learning Representations*.