Opinion diversity through hybrid intelligence

Meer, M.T. van der

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 underrepresented 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 commonsense 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.

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 HISs 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.