

THE OPEN UNIVERSITY

# Computational Argumentation Approaches to Improve Sensemaking and Evidence-based Reasoning in Online Deliberation Systems

Lucas ANASTASIOU

A thesis submitted for the degree of

*Doctor of Philosophy*

Science Technology Engineering and Mathematics Faculty

Knowledge Media Institute (KMi)



# Declaration

I, Lucas Anastasiou, declare that the research in this dissertation is my own work or that where it has been carried out in collaboration with, or supported by others, that this is duly acknowledged below and my contribution indicated. I attest that I have exercised reasonable care to ensure that the work is original, and to the best of my knowledge does not break any UK law, infringe any third party's copyright or other Intellectual Property Right, or contain any confidential material. I confirm that this dissertation has not been previously submitted for the award of a degree by this or any other university.

*to Piero*

# Acknowledgements

Before starting a PhD, I had already heard many times that “a PhD is not a sprint, is a marathon”. To me, it sounded cliché and a bit pretentious. It was not long after I engaged in the whole process of “*PhD-ing*”, that this little piece of wisdom struck me. Moreover, a quote that I heard from a running coach to a guy that just finished a marathon race in the last position: “..the last in a race is the first that did not quit”, gave me the motivation to embrace the process and keep pushing till the end. This work would not have been made possible without the support of many people over the span of several years. First and foremost, I need to offer my deepest and most sincere gratitude to my supervisors, Prof. Anna de Liddo and Dr. Petr Knöth for guiding me and supporting me through this long PhD journey. For Anna, my primary supervisor, a special thanks for the mentorship, for putting up with me and guidance. I am forever indebted.

For Petr (and all of his CORE team that I used to be part of before starting my PhD), an extended thanks for planting in me the seed of a researcher mind.

I am also grateful to Prof. Advait Siddharthan and Dr. Aisling Third - the examiners in my mini-viva- for their insightful comments and feedback, which helped shape my work for the rest of the PhD. My most sincere thanks go to my PhD examiners, Prof. John Domingue and Dr. Elena Musi, for their valuable comments and for making my viva examination an enjoyable experience and an opportunity for discussion.

I would also like to thank Office of Naval Research of the US that funded my PhD programme, through the BCause research project. My gratitude also goes to Alberto Ardito and Riccardo Pala that we have worked together endless hours to develop the BCause application website and services. Thanks also to Aldo de Moor and Barbara Brayshay that we have worked together for disseminating the work of BCause.

My gratitude extends to KMI research institute which has been my home for more than 10 years. As an ex-member of the KMI band, a director of a KMI movie, a proud contributor to Knowledge Makers group, and many other endeavours over the years, I feel KMI offered me more than a great working environment: a truly unique place to *be*.

More than anything else, it gave me the chance to encounter so many interesting and inspiring people; the list would be exhaustively long - you know who you are! However, I cannot omit to thank Prof. Fridolin Wild (an ex-KMIer) who first saw something in me and hired me as Research Assistant to work on his technology enhanced learning projects. I extend my thanks to the PG Forum, held by Daniel Gooch and Prof. Marian Petre, for offering an incredible space of learning and sharing your experiences about the academic practice.

I also need to thank my family, for aspiring to me the high value of education since I was a child, and supporting me to pursue a doctoral degree.

Last, I save the most special spot and wholesome gratitude for the love of my life, my partner in crime Lara, the beautiful mother of my beautiful son.

## Related publications

- Lucas Anastasiou and Anna De Liddo. 2021. *Making Sense of Online Discussions: Can Automated Reports help?*. In CHI Conf. Human Factors in Computing Systems Extended Abstracts (CHI '21 Extended Abstracts), May 8–13, 2021, Yokohama, Japan. ACM, New York, NY, USA, 10 pages. <https://doi.org/10.1145/3411763.3451815>
- Lucas Anastasiou, Aldo de Moor, Barbara Brayshay, and Anna De Liddo. 2023. *A tale of struggles: an evaluation framework for transitioning from individually usable to community-useful online deliberation tools*. In The 11th International Conference on Communities and Technologies (C&T) (C&T'23), May 29-June 2, 2023, Lahti, Finland. ACM, New York, NY, USA, 12 pages. <https://doi.org/10.1145/3593743.3593771>
- Lucas Anastasiou and Anna De Liddo. 2022. *Examining the Impact of Argumentation Structuring on Content Depth, Engagement and Polarisation of Online Discussions: A Comparative Study to linear discussion interfaces*. In Proceedings of the Eighteen Research Workshop of International Conference on Information Systems (ICIS '22), December 11, 2022, Copenhagen, Denmark
- Lucas Anastasiou and Anna De Liddo. 2023. *BCause: Reducing group bias and promoting cohesive discussion in online deliberation processes through a*

*simple and engaging online deliberation tool.* In the 1st Workshop on Social Influence in Conversations (SICon 2023), July 14, 2023, co-located with ACL 2023, Toronto, Canada

- Lucas Anastasiou and Anna De Liddo. 2024. *Integrating and Assessing Computational Argumentation Artifacts in a Real Online Discussion System.* In CHI Conf. Human Factors in Computing Systems Extended Abstracts (CHI '24 Extended Abstracts), ACM, *Under review*

# Abstract

Deliberation is the process through which communities identify potential solutions for a problem and select the solution that most effectively meets their diverse requirements through dialogic communication. Online deliberation is implemented nowadays with means of social media and online discussion platforms; however, these media present significant challenges and issues that can be traced to inadequate support for Sensemaking processes and poor endorsement of the quality characteristics of deliberation.

This thesis investigates integrating computational argumentation methods in online deliberation platforms as an effective way to improve participants' perception of the quality of the deliberation process, their way of making sense of the overall process and producing healthier social dynamics.

For that, two computational artefacts are proposed: (i) a Synoptical summariser of long discussions and (ii) a Scientific Argument Recommender System (SciArgRecSys).

The two artefacts are designed and developed with state-of-the-art methods (with the



use of Large Language Models - LLMs) and evaluated intrinsically and extrinsically when deployed in a real live platform (BCause).

Through extensive evaluation, the positive effect of both artefacts is illustrated in human Sensemaking and essential quality characteristics of deliberation such as reciprocal Engagement, Mutual Understanding, and Social dynamics. In addition, it has been demonstrated that these interventions effectively reduce polarisation, the formation of sub-communities while significantly enhancing the quality of the discussion by making it more coherent and diverse.

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Problem Scope . . . . .	1
1.2	Motivation . . . . .	5
1.3	Research Inquiry Rationale . . . . .	7
1.3.1	Main Research Hypothesis . . . . .	7
1.3.2	Research Questions . . . . .	8
1.4	Methodology . . . . .	12
1.4.1	Approach and Logic of Enquiry . . . . .	12
1.4.2	Exploring the Problem . . . . .	14
1.5	Thesis Outline . . . . .	15
<b>2</b>	<b>Online Deliberation and Argumentation Support</b>	<b>17</b>
2.1	Social Media for Collective Decision Making Scenarios . . . . .	18
2.2	User Study I - Online Deliberation within Social Media Platforms, Fears and Aspiration . . . . .	20
2.2.1	Method . . . . .	20
2.2.2	Semi-Structured Expert Interviews: Preparation and Recruit- ment . . . . .	22
2.2.3	Interviews' Questions . . . . .	23
2.2.4	Analysis . . . . .	25

2.2.5	Aspirations and Concerns . . . . .	27
2.2.6	Design Guidelines . . . . .	39
2.3	How can Argument Computation Help? . . . . .	41
2.4	Research Design . . . . .	44
2.5	Summary . . . . .	49
<b>3</b>	<b>Literature Review</b>	<b>51</b>
3.1	Online Deliberation as a Mean of Deliberative Democracy . . . . .	51
3.2	Public Deliberation for Wicked Problems . . . . .	52
3.2.1	Sensemaking in Online Deliberation Systems . . . . .	54
3.2.2	Deliberation Quality Metrics . . . . .	58
3.3	Online Deliberation Current State and Issues . . . . .	61
3.3.1	Overview of Online Deliberation Systems . . . . .	66
3.3.2	Computer Supported Public Deliberation Platforms Review . . . . .	67
3.3.3	Technological Gap Identification Analysis . . . . .	71
3.3.4	Defining the Research Gap . . . . .	73
3.4	Argumentation Theory, Models and Mining Approaches . . . . .	75
3.5	Automated Reporting . . . . .	83
3.6	Natural Language Processing in the Era of Large Language Models . . . . .	86
3.6.1	Evolution of LLMs . . . . .	86
3.6.2	Text Generation Strategies with LLMs . . . . .	88
3.7	Automatic Summarisation Systems . . . . .	91
3.7.1	Summary as a Text-to-text Task . . . . .	92
3.7.2	Summarisation Evaluation . . . . .	93
<b>4</b>	<b>Automated Methods of Reporting for Online Deliberation</b>	<b>97</b>
4.1	User Study II - Argumentation-based Automated Reports . . . . .	98
4.2	Study Research Question and Experimentation Design . . . . .	99

4.2.1	Experiment Design . . . . .	101
4.3	Quantitative Comparative Study . . . . .	102
4.3.1	Task . . . . .	104
4.3.2	Analysis . . . . .	106
4.4	Qualitative study . . . . .	108
4.5	Discussion . . . . .	113
4.6	Summary . . . . .	115
<b>5</b>	<b>Summarising Online Discussions - an Argumentation Approach</b>	<b>117</b>
5.1	Study III - Effect of Prompting Input in Quality Characteristics of Generating Summary . . . . .	118
5.2	Methodology . . . . .	121
5.2.1	Generation Strategies . . . . .	126
5.2.2	Experimental design . . . . .	129
5.2.3	Task . . . . .	130
5.3	Results . . . . .	133
5.3.1	Computational Automatic Evaluation . . . . .	133
5.3.2	Human evaluation . . . . .	134
5.4	Discussion . . . . .	135
5.5	Summary . . . . .	138
<b>6</b>	<b>Scientific Argument Recommender System</b>	<b>140</b>
6.1	Extracting Arguments from Scientific Literature in Scale . . . . .	141
6.1.1	Argument Mining for Scientific Discourse . . . . .	141
6.2	Recommending Scientific Evidence . . . . .	148
6.2.1	Scientific Argument Recommendation Task . . . . .	149
6.2.2	Content-based Filtering with Embeddings . . . . .	149
6.2.3	Recommending an Argumentative Summary . . . . .	151

6.3	Study IV - Human-centric Evaluation of SciArgRecSys . . . . .	156
6.3.1	Recommender System Human Evaluation Factors . . . . .	156
6.3.2	Methodology . . . . .	160
6.3.3	Results . . . . .	166
6.3.4	Discussion . . . . .	173
6.4	Summary . . . . .	175
<b>7</b>	<b>Integrating and Assessing Computational Argumentation Artefacts in a Real Online Discussion System</b>	<b>176</b>
7.1	Description of BCause Deliberation Platform . . . . .	177
7.1.1	Design and Rationale . . . . .	177
7.1.2	Creation of a Low-fidelity Prototype . . . . .	179
7.2	UX Prototype Testing . . . . .	188
7.2.1	Assessing Synopsis Creation . . . . .	188
7.2.2	Assesing SciArgRecSys User Experience . . . . .	189
7.3	Study V - Assessing the Combined Impact of Automated Reporting and Scientific Arguments Recommendation with a Semi-Naturalistic Online Deliberation Experiment . . . . .	193
7.3.1	Research Question . . . . .	193
7.3.2	Experiment Design . . . . .	194
7.3.3	Results . . . . .	203
7.3.4	Discussion . . . . .	219
7.4	Summary . . . . .	220
<b>8</b>	<b>Conclusions</b>	<b>222</b>
8.1	Discussion . . . . .	222
8.2	Limitations . . . . .	228
8.3	Future work . . . . .	230

8.4 Final Remarks . . . . . 232

**Bibliography** . . . . . **233**

**Appendices** . . . . . **269**

A Study I - Interview Sketch . . . . . 269

B Study I - Codebook . . . . . 271

C Study II - Questionnaire . . . . . 276

D Study II - Focus-Group Interview Sketch . . . . . 281



# List of Tables

2.1	List of interviewees and overview of their expertise . . . . .	23
2.2	Interviews list of open-ended questions . . . . .	24
2.3	Example Interview Transcript Snippet with Coding Format . . . . .	26
2.4	The two main categories of themes identified (aspirations and concerns) regarding the use of Social Media in Collective Decision Making scenarios	27
2.5	Proposed technological artefacts for addressing aspirations and concerns	42
2.6	Technological artefacts correspondence to aspirations and concerns themes . . . . .	43
3.1	Public deliberation solutions taxonomy and review . . . . .	65
3.2	Online deliberation platforms support for Sensemaking and quality of deliberation . . . . .	72
3.3	Argumentation mining correspondence with NLP/ML tasks . . . . .	82
4.1	Kruskal-Wallis statistical test of Sensemaking features . . . . .	107
4.2	Kruskal-Wallis statistical test of Quality of Deliberation features . . .	107
4.3	Dunn's test post-hoc examination - pairwise comparisons . . . . .	107
4.4	Focus group analysis themes, corresponding design criteria and quote excerpts . . . . .	110



5.1	Different nuances on generated summaries on the same discussion using different prompts . . . . .	125
5.2	Descriptive stats of the debates used in the creation of the recommendation corpus . . . . .	127
5.3	Example summary using the <i>Mix<sub>g</sub>generationsgenerationstrategy</i> . . .	131
5.4	Example summary using the <i>Follow_sequences</i> generation strategy . .	132
5.5	Automatic summarisation metrics scores . . . . .	132
5.6	Human evaluation factors and corresponding given prompt to Mechanical Turk crowdworkers . . . . .	133
5.7	Sensemaking evaluation factors and corresponding prompts to Mechanical Turk crowdworkers . . . . .	134
5.8	Human evaluation metrics scores . . . . .	135
5.9	Summarisers models scores in Sensemaking dimensions . . . . .	136
6.1	Experiment results - accuracy, individual classes F1 and macro-F1 . .	146
6.2	Details of embeddings used in SciArgRecSys . . . . .	152
6.3	Example showing the transformation of an abstract to its main argument summary representation . . . . .	153
6.4	Quality of embeddings used in the three corpora for clustering . . . .	155
6.5	Evaluation factors per position item . . . . .	159
6.6	Overall evaluation factors of the recommendation set . . . . .	159
6.7	Evaluation of the overall recommendation set using random method .	170
6.8	Evaluation of the overall recommendation set using tf-idf method . .	170
6.9	Evaluation of the overall recommendation set using kNN method . . .	171
6.10	Kruskal-Wallis statistical test of human evaluation factors . . . . .	172
7.1	Contested score example values . . . . .	187
7.2	Descriptive statistics of the debates formed in UNINA use case . . . .	190

7.3	Group comparison statistical test - UNINA use case . . . . .	192
7.4	2x2 Factorial experimental design . . . . .	194
7.5	Evaluation factors for inclusion of artefacts study . . . . .	202
7.6	User contributions statistics over conditions A-D . . . . .	204
7.7	Invocations of computational artefacts per trial . . . . .	205
7.8	Appreciative and reflective feedback statistics . . . . .	206
7.9	Evaluation factors descriptive statistics . . . . .	207
7.10	Evalaution factors Kruskal-Wallis statistical test and post-hoc com- parisons . . . . .	208
7.11	Social network analysis results . . . . .	213
7.12	Topic diversity and coherence over conditions A-D . . . . .	217
7.13	Top topic models of experiment corpus . . . . .	218



# List of Figures

1.1	Research diagram . . . . .	15
2.1	Example interview memos of a sample interviewee . . . . .	27
2.2	Research diagram depicting sub-research questions and corresponding studies . . . . .	48
3.1	Pirolli and Card sensemaking model . . . . .	57
3.2	Deliberation systems examples . . . . .	64
3.3	Simple argument . . . . .	77
3.4	Tulmin’s model of argumentation diagram . . . . .	78
3.5	Argumentation illustration example based on Tulmin’s model depicted from Toulmin (2003) . . . . .	78
3.6	IBIS - Issue-Based Information System elements and interactions . . .	79
3.7	Argumentation mining NLP pipeline . . . . .	81
3.8	Evolutionary tree of modern LLMs . . . . .	88
4.1	Examples of automatic generated reports . . . . .	100
4.2	Sensemaking (SM) and Quality of Debate (QoD) evaluation factors extracted from literature (as detailed in Section 3.2.2) . . . . .	103
4.3	Discussion interface under conditions CA,CB,CC,CD . . . . .	106

5.1	Example discussion as it was rendered for the mechanical turk crowd-worker . . . . .	128
5.2	Example summary as it was rendered for the mechanical turk crowd-worker . . . . .	128
5.3	Examples of synoptical summaries on the debate topic : “Should all humans go vegan?” using (i) GPT-3 (2nd var.) and (ii) t5-large . . .	130
5.4	Pearson Correlation matrix of computational metrics (ROUGE-x, METEOR, BLEU, BERTScore) and human evaluation metrics . . . .	139
6.1	Examples of automatic argument annotation . . . . .	144
6.2	Example of recommendation query . . . . .	149
6.3	Recommendation API response . . . . .	150
6.4	High level view of argument corpus creation . . . . .	151
6.5	Abstract corpus embeddings visualisation . . . . .	152
6.6	Argument corpus embeddings visualisation . . . . .	152
6.7	Summaries corpus embeddings visualisation . . . . .	152
6.8	Examples of recommendation units . . . . .	163
6.9	Recommendation rating HIT interface . . . . .	164
6.10	Counts of participant responses in Accuracy, Novelty, Diversity and Adequacy . . . . .	167
6.10	Counts of participant responses in Perceived Usefulness, Satisfaction and Trust . . . . .	168
6.11	LLM based retrieval . . . . .	174
7.1	BCause discussion interface . . . . .	178
7.2	Argument-centric structure in BCause . . . . .	180
7.3	Argument input prologued by agreement slider . . . . .	182
7.4	Reflection card two stage interaction . . . . .	182

7.5	Reply dialog box and rendered “quoted” text within argument . . . .	182
7.6	Early prototype of BCause agreement slider and reflection card . . . .	184
7.7	Early prototype of BCause auto-generated arguments with reflection card . . . . .	185
7.8	SciArgRecSys instantiation in UNINA use case . . . . .	190
7.9	BCause UI in four different experimental conditions . . . . .	197
7.10	User task steps of interaction and corresponding state of discussion .	198
7.11	Participants’ gender . . . . .	203
7.12	Participants’ level of education . . . . .	203
7.13	Participants’ age groups . . . . .	203
7.14	Interaction model . . . . .	210
7.15	User interaction graph . . . . .	211
7.16	User/Position interaction graph . . . . .	212
7.17	Intertopic distance map in four different experimental conditions . . .	217



# Chapter 1

## Introduction

### 1.1 Problem Scope

In an era of increasingly complex challenges, such as climate change, nuclear proliferation, and the recent COVID-19 crisis, it is becoming more and more necessary to resort to large-scale consultation processes to identify and select solutions that best address organisations' diverse needs.

While consultation processes typically involve soliciting input or feedback from a large number of people encompassing a shallow level of engagement, deliberation in large organisations is a deeper process that often involves face-to-face discussions where a genuine exchange of views and active participation occurs.

Deliberation constitutes the process where collectives (i) identify potential solutions for a problem or resolutions to a conflict, and (ii) select the solution from this pool that most effectively meet their diverse requirements ([Klein, 2012](#)).

Since in-person deliberation is expensive, a plethora of socio-technical systems supporting this process to happen online exist, encompassing contemporary platforms such as social media, wikis, task management software (e.g. Github<sup>1</sup>, Jira<sup>2</sup>), mes-

---

<sup>1</sup><https://github.com/>

<sup>2</sup><https://www.atlassian.com/software/jira>



saging services, e.g. Slack<sup>3</sup>, and others, but even more traditional communication mediums including email, videoconferencing (e.g. Skype<sup>4</sup>, MS Teams<sup>5</sup>), and discussion forums.

The application of these deliberation technologies in organizations has demonstrated many benefits. These include the synergy of ideas (idea synergy) (Klein, 2012), enhanced diversity (Heitz et al., 2022)(allowing the underrepresented voices that usually reside in the “long tail” (Albrecht, 2006)), high-quality results solely due to the extensive verification by multiple actors, and overall better collective judgements - a concept commonly referred as “wisdom of the crowds” (Surowiecki, 2005).

Nonetheless, deliberation carried out online, with social media and online discussion technologies, often exhibit shortcomings in several aspects.

Social media, while rich in engagement features, is quite limited in features to support decision-making, leading to polarisation, division and conflict (Sunstein, 2018; Bozdag, 2020). This is reinforced by the fact that online spaces can be filled with low-quality comments, hate speech, or non-constructive remarks (Faddoul et al., 2020; Golbeck et al., 2017; Matias et al., 2015). This is due to the creation of ill phenomena, such as the “echo chamber” effect, in which a person is exposed predominantly to opinions that align with their own, thus creating an amplified and unchallenged view of their beliefs. This is closely related to filter bubbles (Pariser, 2011), a result of algorithmic filtering that tailors content to an individual’s existing preferences, limiting exposure to diverse perspectives (Bakshy et al., 2015). Social media platforms, therefore, can inadvertently contribute to the creation of echo chambers due to their design and underlying algorithms, such as personalised recommendations that match previous user interactions aligning with user’s known preferences, potentially limiting exposure to diverse viewpoints (filter bubbles) and

---

<sup>3</sup><https://slack.com/>

<sup>4</sup><https://www.skype.com/en/>

<sup>5</sup><https://www.microsoft.com/en-gb/microsoft-teams>

forming of communities solely by like-minded individuals (echo-chambers). This implies that “by design”, the diversity of opinions and content variety is damaged, leading to degraded quality (e.g. lack of critical thinking (Ribeiro et al., 2019) or even safety (Guntuku et al., 2017)). Online deliberation is not immune to false information and deliberate attempts to manipulate public opinion (Lazer et al., 2018). Regardless of the polarity or authenticity of the content, online deliberation often results in a large amount of information, making it difficult for users to sort, analyse, and understand all data points (Perez, 2008; Jonsson and Åström, 2014), therefore *make sense* of the deliberation.

Pirolli and Card (2005) present a model of the Sensemaking process, focusing on how intelligence analysts use technology in their work. While Pirolli and Card sensemaking model does not directly discuss deliberation, it provides insights into the cognitive processes that can help understand its implications in deliberative contexts. Sensemaking involves several iterative stages, such as data gathering, structuring, and interpretation, which are crucial in deliberation. Participants collectively seek and analyze information, negotiate meanings, and ultimately generate shared understandings. The Sensemaking process can be facilitated or hindered by the design and use of technological solutions, affecting the efficiency and quality of deliberation. As such, incorporating Sensemaking in the design of deliberation technologies is critical for enhancing their usability, supporting effective decision-making and fostering collaboration (Griffith, 1999).

Stromer-Galley (2007) proposed a set of qualitative characteristics of effective deliberation, such as: (i) expression of reasoned opinion; (ii) disagreement - as a sign that there is a problem that needs a solution and the existence of distinct views on a particular issue; (iii) equality - the affordance of each participant to participate on equal footing, or otherwise, no participant to dominate the conversation; (iv) sourcing - the appropriate provision of evidence or reference, whether personal or

external, subjective or not; (v) being on-topic - as for a discussion not going off-topic and meeting its objective of deepen understanding of an issue; and (vi) engagement (or reciprocity) where people respond to given claims of others and do not engage in parallel monologues.

Many of the aforementioned issues, which are prevalent in most social media and online discussion technologies, can be traced to **inadequate support for Sense-making** processes and **poor endorsement of the quality characteristics of deliberation content** as mentioned above.

Current deliberation support technologies include all tools that facilitate thoughtful consideration and discussion of issues. These include (i) decision-making support systems (van Hillegersberg and Koenen, 2016); (ii) argument visualization tools, e.g. DebateHub (Quinto et al., 2021), Argunet (Schneider et al., 2007), DebateGraph (Baldwin and Price, 2008), CmapTools (Cañas et al., 2004); (iii) platforms in general that facilitate collaborative work and discussion among large groups (e.g. Slack, Google Workspace, etc.), online discussion platforms (e.g. Reddit, phpBB forums), and digital democracy platforms (e.g. Pol.is<sup>6</sup>, Decidim<sup>7</sup>). Most of these platforms serve very well some stages of Sensemaking, but none actually succeeds in serving holistically all stages of the information-seeking loop, especially in the later sensemaking stages. For example, Quora thrives in the Search and Filter stage, bringing a high amount of external evidence to the discussion. However, it offers limited support for the schematise (in the visualisation of the answers), build case (to build a theory to support previously crafted hypotheses) and tell-story stage (to share the developed theory from the previous step). Similarly, Argunet, for example - a technology specifically designed to tackle the provisioning of reasoned opinion - naturally excels only on this aspect of deliberation content quality. However, it is not so good in other quality aspects, e.g. engagement (reciprocity), as there is no

---

<sup>6</sup><https://pol.is/home>

<sup>7</sup><https://decidim.org/>

function of posting an argument as a reply to someone else position or mechanism to ensure equality among participants.

In general, current deliberation support technologies, while efficiently supporting the early stages of *Sensemaking*, fail when approaching the latter most crucial stages of it (Llinas, 2014). In parallel, regarding the support for *quality deliberation*, while some platforms excel in some aspects of deliberation quality - particularly when specifically designed for such purposes, they often fall short in providing a more comprehensive coverage of the aspects of quality discourse (Klein, 2015).

## 1.2 Motivation

From the outline of ill phenomena occurring in social media and online discussion technologies in the previous section, we identify human sensemaking (the cognitive process through which people give meaning to their experiences) and quality of deliberation content as the primary domains that need better technological support. We seek to leverage recent advancements in Artificial Intelligence (AI) and Natural Language Processing (NLP) to develop solutions that improve Sensemaking and the quality of online deliberation.

Our focus on argument computation is motivated by recent advancements in Computational Argumentation, which have shown promise for the capability of such technologies to support Sensemaking in a variety of ways.

- i **Supporting the Search and Filter Phase:** NLP techniques can be employed to retrieve and summarise relevant information from large volumes of data, assisting the users in the initial information foraging stages of Sensemaking (Pirolli and Card, 1999).
- ii **Text understanding and analysis:** NLP is capable of analysing text and extracting important entities and relations, aiding humans in the Read and

Extract stages of Sensemaking.

- iii **Fact-checking and verification:** AI and NLP can uphold the quality characteristics of sourced information, ensuring the reliability and validity of the information in deliberation context ([Shu et al., 2017](#)).
- iv **Argument mining and evaluation:** automatically identifying argumentative components within text allows the analysis of argument structures and assessment of argument quality ([Lippi and Torroni, 2016a](#)).
- v **Improved recommendations:** the design of recommendation and timeline algorithms can be tuned to expose users to diverse viewpoints and challenge their beliefs, thereby encouraging content diversity and counteracting echo chambers ([Nguyen et al., 2014](#)).
- vi **Improved content quality:** AI can assist in moderating low-quality content and facilitate the creation of high-quality content; therefore, supporting the build case and tell-story stage of Sensemaking ([Gillespie, 2020](#); [Lee et al., 2020](#)).

Essentially, the use of AI can be appropriated to facilitate the support of Sensemaking and argumentative thinking (also known as evidence-based reasoning). This ensures that decision-making processes are informed but also “provides a foundation for rational argumentation and critical thinking in deliberation systems” ([Habermas, 1985](#)); fostering in this way valid argumentation to ensure that decision-making processes are grounded in reason, logic and sound evidence. Promoting valid argumentation in decision-making can have a profound impact in multiple areas. An argumentative approach can foster critical thinking and rational discourse ([Paul and Elder, 2006](#)), increase transparency in decision communication ([Association et al., 1990](#)), facilitate mutual understanding between participants ([Mercer, 2000](#)) and reduce the influence of cognitive biases, leading therefore to more robust outcomes ([Kahneman, 2011](#)).

This thesis aims to examine what and how computational elements can support technology-mediated deliberation and to assess their impact on improving participants' Sensemaking and enhancing the overall quality of the deliberation. Despite the considerable potential impact of tackling this issue and therefore enabling healthier deliberation - as highlighted in the previous section 1.1 - only partial or limited solutions have been achieved to date. Great potential is anticipated in incorporating recent advancements in Artificial Intelligence (AI), specifically in the domain of Natural Language Processing (NLP) and the recent development of Large Language Models (LLMs), into online tools supporting deliberation. My PhD research will examine incorporating computational argumentation tools in online deliberation platforms to benefit the quality of the deliberation itself and improve the sensemaking of participants involved in the deliberation.

## 1.3 Research Inquiry Rationale

### 1.3.1 Main Research Hypothesis

The main hypothesis underpinning this research can be phrased as follows:

**Main Research Hypothesis:** *Argument computation methods integrated into online deliberation systems can improve participants' perception of the quality of the online discussion, enhance their capability to make sense of the deliberation process and produce healthier social networks dynamics when compared to approaches that do not use any argument computation support.*

To test this hypothesis, first, we identified a series of deliberation problems in need of argument computation (2.2.5), considered state-of-the-art NLP and argument mining approaches to solve them (2.3), and identified two argument computation support

elements to focus the remainder of our investigation, which are: automated summary and arguments' recommender. Third, we designed concrete argument computation support solutions and integrated them into an online deliberation platform to test the impact of such solutions. Finally, we tested our argument computation solutions (both in isolation and in combination) in a series of controlled and *in the wild* experiments to assess their impact on the identified problems/target capabilities (7). Further details on the proposed research and evaluation methodology can be found in section 1.4 and in a separate section of each study addressing each of the above stages.

### 1.3.2 Research Questions

The research rationale and workflow described above raised a series of research sub-questions which drove our investigation.

***RQ1:*** *What are the higher level guidelines for the design of deliberative platforms for organisational decision-making?*

Although social media and online discussion platforms are useful for creating a network of social exchange, when the group's goal is to decide something or advance a collective understanding of a complex issue, these platforms are inadequate. It is critical to this research to understand how to improve the overall quality of online discussions, particularly when the goal of the group is to make collective decisions with evidence-based reasoning and promote a broader understanding of issues for all participants.

In the following Chapter 2, we address this question by conducting interviews with researchers and practitioners with diverse backgrounds, with the aim of analysing how quality discussions currently take place in ordinary organisations, with what platforms or social media they are facilitated, and how and when they succeed

in leading to effective collective decision-making. From the analysis of interviews with experts, we derive a number of aspirations and concerns about the use of deliberative platforms for collective decision-making. We have then systematically analysed deliberation problems and explored how argument computation would be most needed. From this analysis, we identified automated summarisation and recommendation of scientific arguments (as elements of robust argumentation) as the two most promising argument computation artefacts to address the largest number of issues. This led to the following four research questions that drive the rest of this thesis.

***RQ2:** What automated reporting approaches are more appropriate for online deliberation?*

We examined automated reporting as a promising means of improving Sensemaking in discussion platforms. Through comparison of three approaches to automated reporting: an abstractive summariser, a template report and an argumentation highlighting system, we observed improvements in the Sensemaking of participants and the perception of the overall quality of deliberation content. Through this examination, we suggest that both argument mining technologies and abstractive summarisation are particularly promising computational aids to improve sensemaking and perceived quality of online discussion, thanks to their capability to combine computational models for automated reasoning with users' cognitive needs and expectations of automated reporting. Still, abstractive summarisers present challenges in generating a comprehensive yet coherent summary of extremely long discussions in large crowds. We, therefore, proceeded to the next Research Question:

***RQ3:** To what extent can an AI-generated abstractive synopsis of an online discussion provide a quality summary of the discussion and significantly improve Sensemaking?*



We investigate the use of state-of-the-art generative large language models in summarising long online discussions and whether they can attain a harmonious balance between adequacy and coherence. For that, we compare the results of human and computational evaluation metrics and explore the effect on participants' Sensemaking. We conclude that a minimal loss of accuracy can be tolerated -and actually preferred- in favour of fluency and thus comprehensibility.

We have tested the performance of various state-of-the-art summary models for online discussions. We performed a hybrid evaluation approach, to measure both the computational performance using standardised metrics and also the quality of summaries when judged by humans. In addition to the intrinsic evaluation of the quality of the output summary, we performed an extrinsic evaluation to measure the impact on the overarching task for which the summary is intended. We, therefore, compared the impact on Sensemaking of the summary when presented alongside the original discussion. We conclude that prompting large Language models (LLMs) is the best method (human evaluation) for generating quality Summaries, though other methods are comparable from an NLP perspective. However, in terms of extrinsic evaluation, Large Language models based summary have the highest positive effect on Sensemaking compared to other methods.

These results showed improvements in our solution on one of the 2 target capabilities (*Sensemaking*). In order to target our second target metric (improved *quality of deliberation content*) we looked at the opportunity to foster Evidence-Based reasoning (EBR) in online discussions, with the integration of a recommender system for providing high-quality scientific arguments sourced from the scientific literature.

We hypothesise that recommending accessible scientific evidence (to support and oppose) discussion posts will improve participants' perception of the quality of the discussion and produce healthier social interaction between participants. Hence, we asked:

***RQ4:** To what extent does the provision of quality external scientific arguments improve sensemaking and the overall quality of the online deliberation process?*

To address this question, we developed a recommender system designed to provide scientific arguments in online discussions. We carried out an initial investigation exploring whether extracting arguments from scientific literature can be executed accurately at scale. We then evaluated different methods of recommending scientific evidence taking into account distinct granular levels of argument: i. short quoted extracts (excerpts) from research papers, ii. research paper abstracts and iii. summarised abstracts depicting the main argument of the research paper.

We undertook a comparative analysis of diverse methods pertaining to recommending scientific arguments, with a specific focus on identifying those that demonstrate the greatest efficiency at scale - as the intention is to integrate them into a large-scale online discussion system. We showed that recommending abstracted arguments (main argument abstraction via LLM transformation) from the scientific literature is better than recommending argument excerpts or paper abstracts, as they provide better-perceived usefulness, relevance, argumentation, and polarity identification of the argument recommendation.

The ultimate objective of this research is to devise argument computation support for online deliberation systems, to improve participants' perception of the quality of the online deliberation content, enhance their capability to make sense of the deliberation process and produce healthier social network dynamics.

The investigation of the research questions above showed the potential that abstractive summarisers and recommender systems of scientific arguments have to improve intrinsically the quality of the deliberation process.

To assess the effect of integrating those two artefacts in a live online discussion platform, evaluate their combined effects, and examine a larger variety of evolution

metrics, we carried out an exhaustive evaluation study examining an array of quality variables such as participants' Sensemaking, Mutual Understanding, Aesthetics, Engagement and Social Network dynamics.

**RQ5:** *To what extent does the automated reporting and provision of scientific arguments in combination improve Sensemaking and the quality of the online deliberation process?*

Through systematic analysis and interpretation of the data, we examined the effect of the two computational artefacts presence (automatic *synoptical summariser* and *SciArgRecSys*) in Sensemaking (SM), Engagement (Eng), Mutual Understanding (MU), Aesthetics (Aes), and Social Dynamics in online discussion and conclude on the interplay among these variables.

## 1.4 Methodology

### 1.4.1 Approach and Logic of Enquiry

The main purpose of the work in this thesis is to improve Sensemaking and the quality of deliberation within online discussions. To this end, we design a general methodology that we avail in different parts of this thesis. The methodology akin to Design Science for Information Science is oriented towards creating successful *artefacts* (Peffers et al., 2007). It consists of the following steps: (i) Problem identification and motivation, (ii) Definition of the objectives for a solution, (iii) Design and Development, (iv) Demonstration, (v) Evaluation and (vi) Communication. We solely focus on Research through Design (RtD) (Zimmerman et al., 2007); an approach to knowledge generation that emphasizes the role of design as a means of inquiry (Zimmerman and Forlizzi, 2014). We use the design process to investigate and explore the complex problem of how to improve online discussion Sensemaking and to develop

a deeper understanding of the issue.

Key aspects of Research through Design include:

- Exploration: Researchers actively engage with the design process to explore and understand the problem space, identifying new questions and potential solutions.
- Iteration: The design process typically involves multiple iterations, allowing researchers to refine their ideas, learn from failures, and converge on more effective solutions.
- Materialisation: By creating tangible artefacts or systems, researchers can more effectively communicate their ideas and insights and engage with their target audience.
- Reflection: Throughout the design process, researchers engage in critical reflection, considering the implications of their design choices and the broader context in which their work is situated.
- Evaluation: The outcomes of research through design are often evaluated in real-world contexts, helping researchers to assess the effectiveness of their design solutions and gather insights for future work.

As such, RtD is more closely aligned with the philosophical stance of *constructivism* (Crotty, 1998). Constructivism posits that reality is socially constructed and that knowledge is produced through human activity. In RtD, the design process is seen as a form of inquiry, where designers actively construct knowledge through their design activities. The design artefacts, systems, or environments created in this process are the embodiment of the designers' understanding of the problem space, reflecting their assumptions, beliefs, and values. This aligns with the constructivist view of knowledge as being constructed rather than discovered. On the other hand,

*positivism* holds that the only authentic knowledge is scientific knowledge, and that such knowledge can only come from positive affirmation of theories through strict scientific and empirical methods (Donaldson, 2005). Positivism tends to value objectivity, quantifiable research, and generalisable results, and it often seeks to identify universal laws or principles.

However, through RtD approach, we do not mutually exclude either of the two paradigms. Indeed, we use positivist methods, namely controlled experiments, to evaluate the artefacts produced through the constructivist process of RtD. So, while RtD is primarily an inductive approach, it is also iterative and flexible, allowing for a continuous cycle of theory generation (induction) and theory testing (deduction).

### 1.4.2 Exploring the Problem

Research through Design suggests that things should be informed by current theory and practice while spawning new theories and practice through the design and evaluation process (Zimmerman and Forlizzi, 2014). To explore the problem posed by the central research question 1.3.1), we carry a contextual inquiry, to understand the needs, behaviours, and perspectives of experts in this domain. This study, presented in Chapter 2, frames the problem and leads to the generation of potential design artefacts; that each holds a research question, as depicted in the research diagram (Figure 1.1).

This exploration survey (in Chapter 2) delineates concerns and aspirations of utilising online deliberation within Collaborative Decision Systems. As an output of this survey, we proceed to craft a set of guidelines for the development of online deliberation platforms that we use as the set of objectives for our solution. Also, the four main research questions RQ2, RQ3, RQ4 and RQ5 (see section 2.4) governing the rest of this work are unveiled. Subsequently, we deconstruct the primary research question into a set of more focused sub-questions that we address in the following chapters.

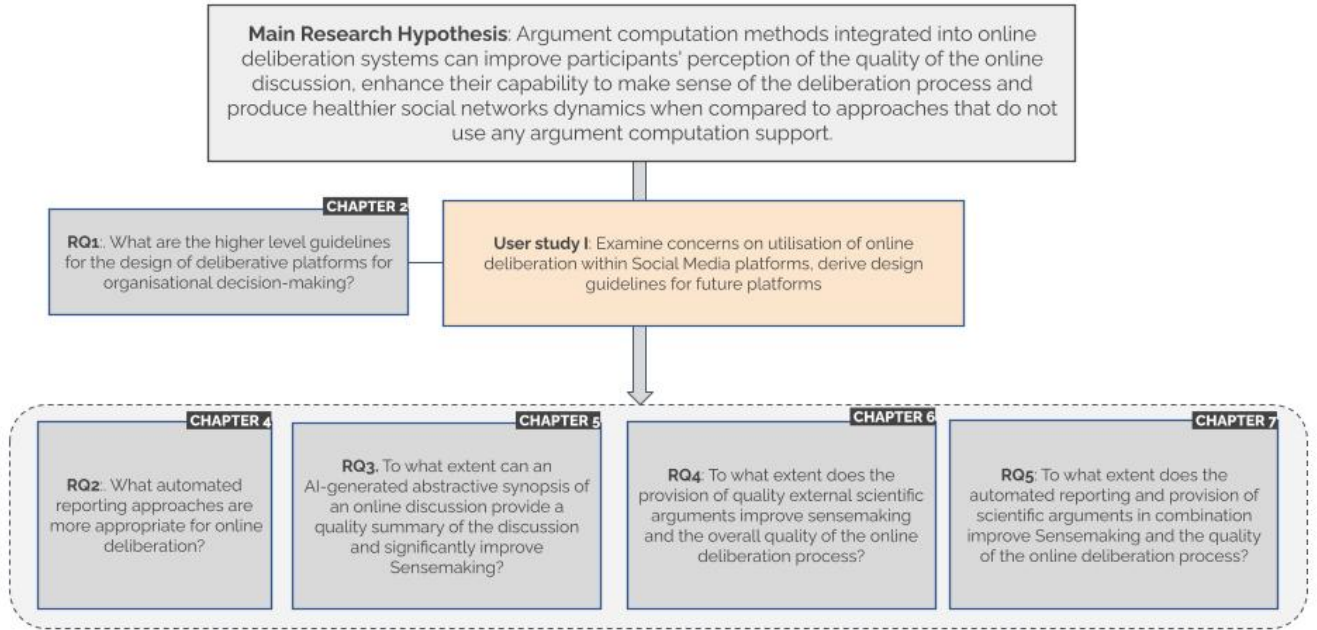


Figure 1.1: Research diagram

Out of a multitude of potential solutions, we funnel down to two specific artefacts (a discussion synopsis automatic generator and a scientific argument recommender) that we design and evaluate (in Chapters 5 and 6 respectively). As a final step, we assess the effect of including the two artefacts in an online discussion platform in a holistic study in Chapter 7.

## 1.5 Thesis Outline

We have already presented the research problem, the motivation and the research inquiry rationale in this introductory chapter. We also have deduced the research questions as shown in section 1.3.

We present an exploratory study (*Study I*) of the fears and aspirations of online deliberation in the current ecosystem of social media tools in Chapter 2.

After presenting background work and reviewing the current state of the literature on online deliberation and computational Sensemaking in Chapter 3, we will examine RQ2 (1.3.2): “What automated reporting approaches are more appropriate for online deliberation?” in Chapter 4 and RQ3 (1.3.2) “To what extent can an AI-generated abstractive synopsis of an online discussion provide a quality summary of the discussion and significantly improve Sensemaking?” in Chapter 5, respectively. Then we address RQ4 (1.3.2): “To what extent does the provision of quality external scientific arguments improve Sensemaking and the overall quality of the online deliberation process?” in Chapter 6.

In Chapter 7, we finally examine for RQ5 1.3.2 the effect of both artefacts’ (the automated synopsis and the scientific argument recommender) presence in a real deliberation platform. We conclude this thesis with a final chapter outlining our overall conclusions on the topic, limitations and proposals for the continuation of this work.

## Chapter 2

# Online Deliberation and Argumentation Support

“Begin at the beginning,’ the King  
said gravely, ‘and go on till you  
come to the end: then stop.”

---

Lewis Carroll, Alice’s Adventures in  
Wonderland

Social media and online discussion platforms are widely used by individuals, groups and organisations for group discussions and to facilitate participation in more inclusive decision-making processes. However, as discussed in Chapter 1, these technologies have some well-recognised limitations and pitfalls that undermine healthy group deliberations and often produce polarisation, division, and conflict. Although these tools are useful for creating a network of social exchange, when the goal of the group is to decide something or advance the collective understanding of a complex issue, these platforms are inadequate. It is critical to understand how to improve the overall quality of online discussions, particularly when the goal of the group is to make collective decisions, with evidence-based reasoning and promote a broader



understanding of issues for all participants.

In this chapter, we describe considerations and guidelines for creating online platforms that enable more inclusive, accountable, and evidence-based public deliberation and collective decision-making. To identify those, we conducted interviews with researchers and practitioners with diverse backgrounds, with the aim of analysing how quality discussions currently take place in ordinary organisations, with what platforms or social media they are facilitated, and how and when they succeed in leading to effective collective decision-making.

From the analysis of interviews with experts, we derive a number of aspirations and concerns about the use of deliberative platforms for collective decision-making. We proceed to propose a set of guidelines for the design of deliberative platforms for organisational decision-making.

## 2.1 Social Media for Collective Decision Making Scenarios

In recent years, social media has drastically changed the public discourse and political communication in society ([Suran et al., 2022](#)). They have unlocked the barriers of communication and provided new exposures by providing an open communication channel over the globe. As a consequence of this, people are now more actively participating in deliberations, argumentation, and explications of public issues and public interest debates ([Jennings et al., 2021](#)).

The use of social media as a medium or discussion platform, in general, has emerged as a critical component in collective decision-making scenarios. These digital platforms are capable of facilitating rapid communication and democratising the process, providing accessibility to a vast array of perspectives and enabling individuals to participate regardless of geographical boundaries ([Boyd and Ellison, 2007](#)). This

has transformed the nature of online interactions, allowing for asynchronous communication and a more inclusive dialogue and engagement. As an example, in [Treem and Leonardi \(2013\)](#), social media use within organisations was examined, arguing that social media are of important consequence to organisational communication processes because of their affordances (visibility, persistence, editability, association), altering the way that employees communicate, share information and collaborate. However, the efficacy of social media platforms in such scenarios is not without limitations. These technologies have some well-recognised limitations and pitfalls that undermine healthy group deliberations and often produce polarisation, division, and conflict.

In an early study of Online Social Network Sites (SNS), [Ellison et al. \(2007\)](#) investigated how SNSs can impact users' social capital, which in turn may influence collective decision-making. Later, [Conover et al. \(2011\)](#) explored the role of Twitter in political discussion and its impact on polarisation, later reconfirmed in various studies, e.g. ([Boxell et al., 2017](#); [Barberá et al., 2015](#)), suggesting though, a more complex and multifaceted relationship between social media use and polarisation. A more recent study on Facebook by [Bakshy et al. \(2015\)](#) examined the extent to which Facebook users are exposed to ideologically diverse content, demonstrating the appearance of echo chambers and filter bubbles, which are a significant factor in poor decision-making. Moreover, early signs of misinformation spread were investigated in [Vosoughi et al. \(2018\)](#), sharing insights on the role of misinformation in collective decision-making.

In summary, on the one side social networks are providing a vibrant space for online public or political discourse ([Suran et al., 2020](#)), on the other hand, they are not supporting the right conditions for meaningful discussions and informed decision-making and therefore have adverse effects on the formation of public opinion ([Rose and Øystein Sæbø, 2010](#)), such as the polarisation of ideas, the spread of

misinformation and formation of echo chambers ([Suran et al., 2022](#); [Pasquetto et al., 2020](#)).

It is becoming, therefore, increasingly imperative to understand the impacts (both positive and negative) of online discussion technologies on deliberation and decision-making, as it is a vital process to promote evidence-based reasoning and consensus-building in collective decision-making processes in the public sphere.

## **2.2 User Study I - Online Deliberation within Social Media Platforms, Fears and Aspiration**

In order to get distinct perspectives on the use of social media and online discussion tools for collective decision-making scenarios, we ran an initial exploratory and scoping study aimed at understanding the current use of existing social media and online discussion technologies in organisations, with the objective to capture both fears and aspiration for the use of online deliberation technologies. From the analysis of those, fears and aspirations, we then identified a series of problems in need of argument computation, which will be the object of the remainder of this thesis. From the study, we also distil a set of guidelines for the design of future and healthier online deliberation technologies.

### **2.2.1 Method**

In order to gather qualitative insights, different methods were considered. One of those was through focus groups that involve the facilitation of group discussions in order to explore participants' shared experiences, attitudes, and perceptions on a specific topic ([Krueger, 2014](#)). Focus groups provide a social context within which the phenomenon is experienced, and through synergy with other participants

provide opportunities for reflection and stimulate deeper expression of ideas that may not surface in individual interviews ([Barbour, 2018](#)). Essentially, focus group participants can listen to others' opinions before forming their own views. However, this introduces group dynamic bias and it focuses more on the collective perspective, offering less space to capture the individual perspective on a topic.

Structured interviews were another method considered to provide participants' insights, opinions, motives and experiences ([Gregar, 1994](#)). They provide a systematic and standardised approach to data collection, and they are characterised by predetermined, fixed-format questions, which while they promote greater reliability in the data collected, uniformity in questioning minimising data complexity and variability ([Cohen et al., 2017](#)), however, in the case of fully structured interviews, there is little space for the researcher to probe into given answers and follow on those; limiting the depth and richness of data collected ([DiCicco-Bloom and Crabtree, 2006](#)).

Another common survey method is by questionnaires, a valuable tool for researchers seeking to explore subjective human experiences, attitudes and perceptions ([Creswell and Creswell, 2017](#)). As a means of inquiry, it can administered, through various formats (e.g. online survey, mail, face-to-face) providing, therefore, flexibility to the data collection strategy - especially if combined with other methods ([Ponto, 2015](#)). However, the quality of the data collected may be compromised by participants' self-reporting inaccuracies ([Fowler Jr, 2013](#)).

In our case, we chose to do semi-structured interviews for qualitative data collection because this is the appropriate method if you want rich perspectives on a specific topic. They present a balanced approach, combining the advantages of structured and unstructured interviews ([Whiting, 2008](#)). They offer the opportunity for participants to elaborate on their responses, as researchers can probe deeper into specific topics and themes, thereby generating richer and more nuanced data ([Galletta, 2013](#)). This leads to more complex and variant data and it enables the exploration of unexpected

or emergent findings.

In our inquiry, a collective discussion among the sampled experts (e.g. a post-interview focus group) could have surfaced more explicitly the patterns of disagreement between the participants. Nevertheless, a deliberate effort was taken to emulate this process by incorporating targeted inquiries during the interviews and subsequently identifying such patterns in the analysis phase. Doing solely interviews was deemed as a pragmatic compromise and response to limitations imposed by time and resource constraints.

### **2.2.2 Semi-Structured Expert Interviews: Preparation and Recruitment**

To carry out the Expert interviews we recruited experts in a variety of fields, including privacy and accountability, quality of discussion, collective intelligence, argumentation, and sensemaking. The empirical study reported below consists of a series of interviews conducted over 1 month with 14 of such experts (hereafter noted as E01, E02, ..., E14).

Interviewees were given a sample sketch of the interview a week in advance and agreed on their data being shared via a consent form.

The one-on-one interviews were held over an online conferencing tool and lasted on average 45 minutes. The original sketch of the interview was loosely followed, as in a semi-structured interview format ([Wilson, 2012](#)), and the interviewees were allowed and encouraged to divert the conversation to emphasise what in their opinion were important topics to discuss.

Participants were recruited based on their expertise, allowing an equal share of different angles, namely, Interaction Designers (ID), Data Ethics (DE), AI Practitioners (AP) or Strategists (ST), Deliberation Specialist (DS), and Collective Intelligence

Table 2.1: List of interviewees and overview of their expertise

EX	ID	DE	AP/ST	DS	CI
E1(F)		✓			
E2(M)		✓			✓
E3(F)					✓
E4(M)	✓				✓
E5(M)				✓	
E6(M)			✓		✓
E7(F)	✓				✓
E8(F)			✓		
E9(M)			✓		
E10(F)			✓		
E11(F)		✓			
E12(M)				✓	
E13(M)				✓	
E14(M)				✓	

Scholar (CI) (see Table 2.1) and ground their viewpoints to proceed and proposed guidelines for the design of online discussion platforms.

### 2.2.3 Interviews' Questions

To elicit fear and aspirations on the use of technologies for online deliberation in real organisations, interviewees were first asked to think in very general terms, about how deliberation currently happened in their organisation, from there moving to a discussion on problems and opportunities in current practices. We, therefore, asked the following exploratory question:

*What are the main deliberation processes in the context of your organisation, what issues do you identify within those and what opportunities or solutions do you foresee technology to provide?*

Having this backbone question, several open-ended questions were then asked as prompts for further reflection. As in a semi-structured interview format, the line of

questioning was adapted, added, or removed according to the interviewee's responses. This flexibility to follow up on interesting topics or themes that emerge from the conversation led to the discovery of unexpected insights and richer data. The set of open-ended questions we used is shown in Table 2.2 while the interview sketch initially provided to participants can be found in Appendix A.

Table 2.2: Interviews list of open-ended questions

Qs(Q)	Description
Q1	What do you think online discussion tools are for? What is your experience with online discussion technologies? What tools/platforms have you used for such purposes in the past, as an individual or in your organisation?
Q2	How important do you think they are to make good decisions? What do you know about the current practice for making “big” group decisions? (in your organisation or in your general knowledge)
Q3	Do you have any need to realize group decisions in your organisational role or in your current research or everyday life? (or is just decisions behind closed doors or in small groups usually enough?)
Q4	Do you identify any problems in the current processes to carry on public consultation and deliberation in your organisation/research or general experience?
Q5	Do you have an established protocol or process for collective deliberation?
Q6	Can you give us some practical examples of some cases in which you personally needed to consult or involve a larger group of people in a decision? How did you carry on the decision? Did you use any technology to support it?
Q7	In what way do you think more participatory, more inclusiveness in this process will benefit your organisation's strategy and large-scale deliberation?
Q8	Do you have any further recommendations or insights that you consider important for the enhancement of online collective deliberation tools?

### 2.2.4 Analysis

In order to analyse the interview data, an inductive approach was employed. This method is grounded on the theory of qualitative data analysis ([Willig and Rogers, 2017](#)). Our approach, consisting of a shallow theory-driven evaluation ([Chen, 1997, 2012](#)), in combination with a Grounded Theory method ([Glaser and Strauss, 2017](#)), facilitated the identification, analysis, and interpretation of patterns of meaning embedded within the responses of the interviewees. This combination enabled the use of structured-based methods, like a theory-driven method with in-depth analysis and, in parallel, to take advantage of the profound analysis inherent in idiographic studies. We find that this combination approach is ideal for our exploratory study, as shallow theory-driven evaluation may only partially integrate theoretical perspectives to inform the analysis or draw limited connections between the evaluation findings and the underlying theory.

The inductive approach is a bottom-up method where the data analysis starts with specific observations and gradually progresses towards broader, more abstract themes ([Thomas, 2006](#)). This approach allows researchers to build their understanding of the data through the iterative process of coding, categorising, and theme development, rather than imposing pre-existing theoretical frameworks ([Braun and Clarke, 2006](#)). Initially, the data collected from interviews and focus groups were transcribed, as accurately as possible. In the process of data transcription, we opted not to employ a verbatim style, as it would be excessively meticulous for the purpose of this study. The transcription enabled us to familiarise ourselves with the nuances of the participants' responses, facilitating a comprehensive understanding of the data and enabling the consequent analysis.

Subsequently, the process of coding was initiated, which involved identifying significant and recurrent segments in the data that were relevant to the research question. During this phase, we maintained an open and flexible mindset to allow the emergence



Table 2.3: Example Interview Transcript Snippet with Coding Format

Data	Codes
... uh emails i find it very important because also you you need some things to be traceable ...	traceability
... by contrast, we have the experience with all the social media tools that got a lot of uptake. They were super strict and extraordinarily simple. Yeah, you know, people like email. ...	adoption familiarity simplicity
... So sometimes a group might need actual intervention by somebody who is more like a mediator that helps resolve disputes or conflicts amongst the participants so that they can even work together in the first place ...	interaction complexity facilitating discussion facilitated decision-making
... they need some help in being forward looking. They're, you know, to kind of avoid traps and things. The facilitator mediator could be extremely helpful there. There's a lot of elements of that that could be kind of technologised. I think, you know, send signals or identify potential themes, right? Or just reminders, That kind of stuff. I think that could work. ...	facilitated discussion technology support aspiring

of novel patterns and concepts. Essentially, codes are gerunds - the noun form of a verb ending in -ing - allowing us to capture “data in a mature way” (Glaser, 1978).

We offer a snippet example of the coding process in Table 2.3.

Further to the coding process, for each interview collections of memos were gathered. Memo-writing is a reflective and analytical process to enable the researcher to document their thoughts, insight and interpretation of the data collected. Within the context of an interview, memos are valuable to capture emerging ideas and themes, and enhancing the understanding of the data. We offer an example of a collection of memos written for one of the interviews in Figure 2.1

Following the coding process, the identified codes were grouped into broader categories, representing the underlying patterns and themes within the data. This phase

Experience: The interviewee has extensive experience in community-based research and working with online discussion tools. They discussed their experiences with various platforms, including social media, forums, and email.

Observations: They observed that people’s preferences and requirements often dictated which platform was chosen, as well as organisational, legal, and political factors. The interviewee highlighted the distinction between discussion and deliberation, and the importance of finding an equitable way to ensure all voices are heard.

Challenges identified: The inertia of using familiar tools, such as email, even when better-suited alternatives are available. The interviewee also noted the influence of organisational culture and decision-making processes, as well as the importance of making tools easy to use and adding value rather than working for users.

Figure 2.1: Example interview memos of a sample interviewee

Concerns	Aspirations
System abuse	Automated analysis
Inertia	Direct democracy
Adoption difficulty	Evidenced information
Shallow discussion	Automate workflows
Accountability	Computer-supported discussion
Trust and privacy	Computer-supported decision making

Table 2.4: The two main categories of themes identified (aspirations and concerns) regarding the use of Social Media in Collective Decision Making scenarios

involved constant comparison, ensuring each category was distinct and accurately reflected the data. The full codebook during the coding process can be found in Appendix [B](#).

### 2.2.5 Aspirations and Concerns

A summary of the aspirations and concerns of participants regarding the use of Social Media as tools in Collective Decision Making scenarios is presented in Table [2.4](#).

## Aspirations

By aspirations, we wanted to recognise measures that can help us to overcome aforesaid challenges. In particular, we wanted to know how our experts envision the use of technology for improving the quality of discussions and assistance in individual or collective Sensemaking.

**Automated analysis** Interviewees highlighted the potential to provide analytics on top of discussion that could assist understanding but also keep track of discussions (accountability).

For example E07 says:

(i wish) there was some sort of rating scale (to tell us) "did they spend all of their time arguing?" Or "did they pay any attention to anyone else's point of view?". "Were various points of view reconciled?" [...] "Did people even acknowledge?" [...] (aspiring about quantification of quality of discussion) if you could actually, at the end of the day, say this type of motion was used eighty percent of the time, this motion was used three percent of the time, we could actually compare meetings too, conclude that this group performed 30% more [...]

While E09 states that:

I think automated analysis tools are really helpful [...]. Now the new developments around, like with natural language processing, all the social network analysis tools that help us see patterns that were otherwise undetectable [...].

E12 believes automated analysis helps to detect malicious content:

(reflecting on the state of automated tools) I don't think is still there, [...]

um, they definitely help but they also raise red flags like is this accurate,  
is this real or fake news

**Direct democracy** In alignment with the principles of deliberative democracy, many experts shared the aspiration of direct democracy. Direct democracy represents a paradigm where decision-making processes are driven directly from the stakeholders involved. E14:

(reflecting on how participatory is the process of decision-making in their organisation) No, is not very participatory. Sometimes people that are the ground are invited in the meetings but they just need to communicate their professional opinion and understanding to their seniors and they will go ahead (with the decision) [...] is like indirect democracy, like voting for the parliament, right? [...] I have also worked in companies in the past, that people that actually do the job every day and were never even invited to give their opinion, (not even) to contribute not to make decisions

This aspiration is deeply rooted in the democratic ethos of participatory decision-making. Achieving equal participation in a debate (or in general in a deliberation process) is often a requirement to pursue direct democracy. However, this is difficult to implement in real-life situations and often degrades to pseudo-equality. E10 remarks:

(talking about organisational decision making) the challenge is often to find an equitable way to help voices to be heard and also is often a goal to try and hear all the voices in a conversation from all the stakeholders (before reaching a decision) (..)

**Evidenced information** Providing evidence has a direct impact on the quality of the discussion. According to E10, it provides justification and strengthens the rationale behind a decision:

(asked about what makes a good discussion) what makes a good discussion, is the quality of the outcome. [...] In the end, you have a decision that is made and you have the reasons for it, the cases for it, and the cases against it and why you went with the case you went with.

E14 suggests that providing evidence or hard facts as they call it, reduces subjective or opinionated decisions:

It makes an informed statement that is backed by evidence [...] it doesn't necessarily mean that you will have a conclusion. [...] On the opposite, if you have people contributing anecdotal feedback then is a problem.

E05, suggests that the design of the interface of a discussion platform should be able to address various levels of depth in a discussion:

[...] (discussing whether bringing evidence to discussion influences engagement) I'm still hoping someone has that idea for a system that is able to do that so that people that want more depth can get it people that want the shallow thing can also get it, so an interface that doesn't force depth; deeper people to go shallow or shallow people to pretend that they are deep

**Automate workflows** E01 states some of the problems regarding the quality of discussion, such as information overload, difficulties tracking the flow of conversation, and the propagation of false information or 'fake news'. However, they later state that:

[..] technically I don't know how we can do this (solve the previously mentioned problems) but we can in some way try to imagine that bringing machines to do the little things is going leave some space for us -humans- to think about the important stuff [..]

While E11 states that automating workflows may be crucial for accountability reasons and the need to have an automatic tool to support deliberation workflows:

(talking about the problems of large-scale deliberation) [..] if I am engaging 100 thousands of people, I need a tool that manages to summarise for the organiser, there might be an accountability moment (someone asks for a previously mentioned point), in which we show the summary, otherwise the complexity would be so high that -you know- they don't care looking in 100k different parts.

**Computer-supported discussion** E05 aspires on the potential use of computer-supported discussion to abate the problem of large-scale discussion:

[..] Discussion tools work well if you have two people but it does not work if you have lots of polylogues - lots of people talking about different things. I would love a tool that somehow allows you to structure, slash, visualise, whatever you want to call it, arguments in a way that lets you keep track of what's going on with lots of people talking about different things all at the same time

E06 believes that the key to achieving this is a good interface

[..] (talking about computer support discussion tools) I've come to realize that it's the interface that's going to make or break any of these tools

Then E02 is going visionary to solve the problem of too many meetings, by saying:

(envisioning tools) perhaps have some tool to do some automatic transcription of your meetings and then build the argumentation trees and classify things into opinions and evidence or you have at the end of your meetings a very nice summary of what happened, you get related facts and evidence while you type on your slack [...]

**Computer-supported decision making** Most of the interviewed experts stated that online discussion technologies provide open space for debates and deliberation, thus it is important to enhance the process of healthy discussions to create an informed opinion and achieve more constructive outcomes. They argue that they can be used to enact decision-making, as stated by E08:

[...] (answering how decision-making is exercised in their organisation) I think your question about how big decisions are made [...] is a question of organisational structure and politics and philosophies so they might be enacted through certain tools [...]

However, there are limitations on what technology can do and whether it can solve inherent problems of democratic processes in society. In a sense, technology is just implementing the deliberation protocols of the physical world in an online environment - but is not inventing a new way of doing things. As E07 states:

(replying to the high aspiration of technology for e-democracy) technology is part of the solution but is never the solution, e-democracy will not solve the problems of democracy, democracy will solve the problems of democracy

Giving a historical perspective on the emergence of social media as deliberation support tools, E09 states:

[..] what I see with all these social media tools and in general with all airBnB-ish platforms, things, they took a lot of the things done with Group Support Decision Making tools but they did it in a different way - it was all organised around the thing that people try to do!

## Concerns

The aggregated themes about concerns regarding implementing online deliberation with the use of social media are outlined below:

**System abuse** Allowing for a rich and diverse opinion environment comes with its limitations. For example, banning extreme ideologies or hate speech may conflict with the freedom of speech principle. As conveyed by E07, talking about a previous experience with a platform they faced the dilemma of them where to put a barrier to extreme ideologies:

(talking about a tool they have developed) somebody asked me: will we allow postings from neo-nazis? and my response -and the guy really applauded me for it- you know it's kind of an interesting thing [..] if they weren't doing anything criminal you know posting stolen credit card information [..] taking [..] the free speech approach we would let them do it, but that would be a loss for us.

System abuse may be originated in the design rationale of social media platforms.

As E11 illustrates:

(reflecting why Twitter creates polarisation) Twitter, it's a very narcissistic tool in terms of the way discussions are breaded and even if people do it, it's a very aggressive, kind of media, similarly, on Facebook discussion of facts and polls or whatever can be lost because it doesn't have the



remit [...]. Who says what and how right, quality of discussions, origin of the statements in the framing, the framing over agendas and issues, and so they're not because they are threats [...], So, yes, it does create polarisation [...].

Implementing inclusiveness is a great challenge as it may be abused by certain people:

[..] had a hard time reconciling them with freedom of speech, I had to go with inclusiveness but at the same time [..] these things will always clash with freedom of speech -which I want (said emphatically) to preserve at all costs- but you don't want to enforce moderation or enforcement of ideas [..]

However, it is suggested that the problem of system abuse is not a technological problem per se, e.g. in [Wright and Street \(2007\)](#) is stated that "This evidence suggests that we should view deliberation as dependent on design and choice, rather than a predetermined product of the technology.", and reiterated by other experts, e.g. E05 says:

e-democracy will not solve the problems of democracy democracy will solve the problems of democracy

**Inertia** A big challenge of designing tools that add value without creating additional work and encourage users to adopt new tools instead of relying solely on familiar ones. Overcoming the inertia of users is the most difficult task, to defeat it a future platform needs to contain an enormous amount of added value rather than adding work [Gaved et al. \(2019\)](#). As conveyed by E08:

(talking of how to overcome inertia)[..] the trick is to find ways of added value over added work [..]

In that sense is advisable for platforms to focus to excel in a single feature, therefore generate a unique point of differentiation rather than adopting and imitating all elements common in all social media platforms.

[..](online deliberation) platforms are still at an infancy state, not really like physical deliberation - like citizens' assemblies or participatory budgeting- for them, the design and the technology is the real focus now on getting one thing right [..]

**Adoption difficulty** Replicating the flow of face-to-face deliberative exercises on online platforms for more effective discussions is a great challenge. As E09 underscores:

[..] when people talk [..] they work out a lot of the discussion norms on the fly [..] there's sort of a different sort of speech, community or cultural expectations as to what counts in decision making or even what arguing with each other is. They can differ quite a bit and it goes beyond ethnic and racial things.

while E11 reflects on how comparison to other physical deliberation settings should be made:

[..] sometimes there is this classic mistake to compare how bad the discussion is on Facebook or Twitter with a well-crafted face-to-face deliberation but again Facebook and Twitter should be compared to a discussion in a bar or a bench [..] is a little bit of an unfair comparison, right? if you are on a bench or in a park and you're having a conversation it will never be the same as a highly structured conversation in a work effect[..] so that comparison is totally pointless

This theme is recurring as a reflection of how natural human dialogue happens in real life, which is actually desirable, however, it is only suitable for casual conversations. In such non-argumentative discussion, populism thrives often with the aid of rhetoric skills that does not necessarily equates with argument validity. E08 exemplifies this:

[..] if we met in a park and we're having a quick discussion about what we thought about the government's response to the pandemic or football or whatever um this might be a very natural conversation but equally it could be um we could be really bad at it we you only have to turn on the TV to see populist politicians and populist political supporters making some sometimes really really not very clever statements and they're very good at arguing but um maybe this doesn't you know even though they have a good skill at arguing [..]

Helping participants understand the content and process of discussions may not be enough to adapt online discussion for decision-making. As noted by E11:

(technology) is absolutely like the starting point right? However, the technology per se is often useless and it's really a mix between procedure online culture like training of people, the onboarding is extremely important, and so on and on [..] (..)

**Shallow discussion** The discourse occurring in social media is described by our participants as “shallow”. This may be due to lack of argumentation as E14 states:

(commenting about the lack of argumentative discussions online) [..] so [people] state that this is it. And that's it! There is no further analysis or argumentation.. and very often the real time discourages any further continuation! Respond quickly and move on!

And reiterated by E06:

(stating what would be the ideal) [...] to provide the reasons why your opinion is like this to be a reasoned opinion, well argued, to offer a distinct opinion of someone else not just repeat again the same...but engage into a meaningful back and forth conversation instead of just -you know- (as the state of discussion is currently) I'm saying my own thing and that's it and I don't care what you say...

Or it may occur due to trolling, as stated by E05:

(talking about dealing with trolls) are just people who are destructive, the only role for them is to distract at a very shallow level any discussion [...] at the level that I started believe that is a lost battle

**Accountability** Multiple experts emphasised an extremely significant concern i.e., accountability, and stated that in recent times accountability has become an issue of debate because it is presently lacking on crowd-oriented web-based platforms. As a consequence, clarity and responsibilities are lacking in user actions. Additionally, they stated that social media and online discussion platforms don't promote accountability and thus encourage the spreading of unlawful content, including misinformation and propaganda. Regarding this context, E3 elaborates:

There is main three issues that comes with social media platform first: privacy, second: accountability, and third quality of discussions and data ethics; so let's take one by one. The problem is that there is no regulation if you want, for example, newspapers are accountable on what is reading there [...], this kind of response, the social responsibility is important for you. Let's move on how discussion is actually affecting good quality off collective decisions [...], I see a there is a level of polarisation [...]. So, you need to have provided facts and evidence but this has been given so many definitions, and there's so many elements that [...].

**Trust and privacy** One noteworthy recommendation given by all experts was that platforms should provide a safe space for online discussion which is currently lacking in contemporary online discussion platforms. And should allow the selection of different degrees of privacy with respect to personal information or conversation data. As it will foster the trust and bring more transparency. For instance, E9 emphasises on that, by saying:

The platforms that manage our data don't respect our privacy, its basically a disease, a big problem, it's like that somebody is stealing your property. Too bad! I can say it makes me frustrated and angry. So that's what you don't want to see.

E08, iterates a classic concern of privacy regarding the use of your personal data, as reported extensively by literature (for example in [Pelteret and Ophoff \(2016\)](#)):

[..] you know the classic one is, uh, me having a conversation with my friends and this conversation then being used to sell my data and potentially affect my personal circumstances in the long term so you know you have the the classic situation of me maybe talking to my friend Lucas about my health isn't so good and then finding out that either an internet platform is trying to sell me health products or i have a problem getting a job or healthcare declined by insurance because of some sort of condition that may negatively affect my opportunities later. So i think in terms of privacy there is the concern that it's a tricky one but the the concern that the purpose to which i believe i'm sharing, um, the information

On top of the challenges identified by our participants was the element of trust, both towards social platforms but also towards other participants in online discussions.

This can be addressed with critical thinking, in general with advancing media literacy of social media users. As stated by E07:

(talking about recent challenges in tool she develops) some of these problems I'm also struggling [...] it's about trustworthiness, online trustworthiness, [...] to engage people into critical thinking before they share online information, to assure them that what they see is real, etc. [...].

### 2.2.6 Design Guidelines

We have devised a set of guidelines based on the set of aspirations and fears presented in the previous Section 2.2.5. We centre the design guidelines around the cross-cutting themes of *accountability*, *argument-structure*, *collective decision-making* and *privacy*.

- *Accountability* is often perceived by users in a restrictive way. Users tend to narrowly associate accountability only with the notion of traceability; paying little attention to the broader social consequences/implications (see *Accountability* concern theme and *Direct democracy* aspiration theme). To manage to cover the wider concept of accountability:
  - *DG1*. Design processes that allow users to inspect, confirm, dispute and correct past conversations. This includes mechanisms that enable users to scrutinize, verify, contest, and rectify historical records - fostering a greater sense of responsibility for an individual's behaviour.
  - *DG2*. Promote transparency, especially in crucial aspects of information processes. This will enhance clarity in the way data is collected, stored and disseminated; therefore increasing user's trust in the system.
- Pure *argument structure* is not ideal in live scenarios.

- *DG3* An argumentative structure of conversation is perceived as harming the natural flow of discussion, so refrain from employing purely argument-centric solutions (see *Adoption difficulty* concern theme). Instead, employ a hybrid approach that maintains the scrutiny of the argumentative structure but does not hinder the natural progression of the discourse (see *Computer-supported discussion* aspiration theme).
- *DG4*. Employ hybrid interfaces that retain the temporal sequencing of the conversation and loosely visualize argument structures. Interfaces should preserve the organic flow of discussion but also hold the affordance to reveal underlying connections and patterns (see *Adoption difficulty* and *Inertia* concern theme).
- *DG5*. Allow for different modes of online discussion, e.g. informal, and goal-based. By facilitating diverse modes of online discussion, a platform can foster a more inclusive and engaging environment for discourse (see *Computer-supported decision making* aspiration theme).
- *DG6* Implement dynamic interfaces that reduce the cost of argumentation; therefore addressing the challenge of adapting familiar social media for deliberative purposes (see *Direct democracy Computer-supported discussion* aspiration theme)).
- *Collective Decision Making* (CDM) Important issues or decisions are not always addressed in a collective manner; removing the opportunity for organisations to cultivate a diverse perspective culture (see *Computer-supported decision making* aspiration theme). Therefore, the following design guidelines are proposed:
  - *DG7*. Integrate collective decision-making techniques into business and enterprise workflows and procedures.
  - *DG8* CDM Systems needs to be transparent and interpretable so that

users can question assumptions and motivate decisions. Good decision-making requires people questioning processes; allowing users to scrutinize assumptions and motivations regarding a decision.

- *DG9* CDM Systems has to be agile and adaptable to community needs, serving the collective interests and promoting the sense of shared ownership over the decision-making process.
- Privacy is a paramount concern for users (see *Trust and Privacy* concern theme).

As such two additional design guidelines are proposed:

- *DG10*. CDM Systems may and should use any personal information that is useful for the decision-making process; in a way though, that retains individual autonomy, i.e. retain privacy but also individual reputation. Systems should strike between meaningful engagement and respect for personal boundaries.
- *DG11*. Corporate processes should have privacy embedded into CDM Systems design. This can be done by embedding privacy safeguards and will reinforce trust and transparency within the organisation.

## 2.3 How can Argument Computation Help?

To focus our investigation, we systematically looked at all aspirations and concerns and mapped to what extent each of the envisioned support technologies mentioned by the experts address each concern and aspiration. The objective of this systematic mapping was to identify technological gaps and problems that would be best served by argument computation enhancements.

We present in Table 2.5, the list of tools mentioned and proposed by the interviewed experts and in Table 2.6 the result of the systematic mapping which shows the



	Artefact	Brief description
A1	Summariser	Summarise discussions - make them more compact and digestible
A2	Highlighter	Use AI-powered algorithms to analyse discussions, detect patterns, and highlight relevant information.
A3	Moderation	Combine with technology and human moderation
A4	Information filtering	Automated detection of harmful content, misinformation, disinformation
A5	Interfaces	Adaptive, navigational affordance
A6	Recommender	Recommendation systems to suggest relevant content, participants, or discussion threads based on users' interests, expertise, and prior contributions
A7	Visual analytics	Visualisation tools that present complex data and discussion insights in an easily digestible format
A8	Gamification	Incentive mechanisms to encourage users to participate in high-quality discussions
A9	Feedback mechanisms	Offering participants real-time reactions, ratings, comments or other forms of automated feedback
A10	Software for facilitation of decision-making	Navigate the complex information, analyse diverse perspectives and reach consensus in challenging issues

Table 2.5: Proposed technological artefacts for addressing aspirations and concerns

correspondence of each technology to the identified themes of aspirations and concerns.

	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10
<b>Concerns</b>										
System abuse				●		●				
Inertia	●				●		●	●		
Adoption difficulty	●	●			●			●		●
Shallow discussion	●	●				●	●			
Accountability	●	●				●	●		●	
Trust and privacy				●		●			●	
<b>Aspirations</b>										
Automated analysis	●	●					●			
Direct democracy			●							
Evidenced information	●					●	●			
Automate workflows	●									●
Computer-supported discussion	●		●					●		●
Computer-supported decision making	●	●						●		●

Table 2.6: Technological artefacts correspondence to aspirations and concerns themes

Upon examining Table 2.6, it becomes evident that none of the proposed artefacts emerges as a panacea - capable of addressing all concerns or realising all aspirations identified by the interviewees in the previous Study I (2.2). Actually, we observe considerable gaps, with certain themes lacking implementation by any or in the best case only by a few of the proposed artefacts. For that, our attention is drawn towards artefacts A1 (Summariser) and A6 (Recommender), as these demonstrate the most potential to address the identified issues. We elaborate further below on the rationale for the selection of these two artefacts.

### Summariser artefact

A synoptical summariser, as an AI artefact, can potentially play a crucial role in mitigating some of the concerns associated with using social media for decision-making. It offers a robust solution to help users navigate the complex landscape

of online discussions by analysing, organizing, and presenting information in an easily digestible format. A synoptical summariser can be designed to be adaptable and customisable to various organisational requirements, legal constraints, and personal preferences. This flexibility makes it an ideal tool for diverse audiences and contexts, minimizing the challenges of tool selection. By intelligently summarizing and highlighting credible information sources, a summariser can reduce the cognitive load to analyse a discussion, counter the spread of unverified information, isolate subjective opinions and misinformation, and promote instead objective and reasoned digest to the user.

### **Recommender artefact**

Further to the use of a summariser, provisioning reliable arguments can significantly contribute to addressing the concerns of using social media for decision-making described above. Similarly to the summariser, those concerns and aspirations can be addressed with the integration of a recommender of scientific arguments. A recommender may assist to filter out any unverified information, subjective opinions, and potential misinformation, offering well-researched and evidence-based arguments. This fosters higher-quality online discussions, thereby promoting informed decision-making. Regarding the concern of familiarity and difficulty in adopting such a tool, this can be mitigated by incorporating user-friendly interfaces and integrating seamlessly with existing tools to reduce the learning curve.

## **2.4 Research Design**

In Chapter 1, I have introduced the main research questions RQ1, RQ2, RQ3, RQ4, and RQ5 (see Section 1.3.2). With the help of the study presented in the current chapter addressing RQ1, we can now proceed to distil the remaining high-level

research questions into more fine-grained sub-research questions. The rationale is to operationalise each remaining research question (R2-5) into a corresponding study. In the following, we describe each RQ and how it will be addressed.

*RQ2. What automated reporting approaches are more appropriate for online deliberation?*

In the previous section, I have motivated the focus on automated reporting as one of the most promising technological enhancements to address experts' fear and aspirations for the use of online deliberation technologies for organisational decision-making. Recent advances in computational tools (namely NLP, LLM and argument mining) have generated a rapidly evolving and expanding landscape of automated reporting methods that can effectively support and enhance the automated reporting of online discussions. Therefore, to address RQ2 a study (*Study II*) has been conducted (reported in Chapter 4 to: (i) identify and evaluate automated reporting methods and (ii) design the integration of automated reporting methods within an existing online discussion platform).

Similarly for RQ3:

*RQ3: To what extent can an AI-generated abstractive synopsis of an online discussion provide a quality summary of the discussion and significantly improve Sensemaking?*

In the creation of an AI-generated synopsis of an online discussion, is crucial to consider several aspects that contribute to the quality of such summary. To answer this question we therefore executed a research study (*Study III*) that tested the performance of various state-of-the-art models for automated summarisation of online discussions, by using both human and machine assessment metrics of the quality of automated summaries (reported in Chapter 5). We then compared the impact

of the different summarisation approaches on participants' Sensemaking, when the automated summary is presented alongside the original discussion.

Addressing RQ4 requires splitting into the following sub-RQs:

*RQ4a. How can you provide high-quality scientific arguments recommendation to deliberation processes?*

The effective integration of high-quality scientific arguments into an online deliberation processes is crucial for fostering evidence-based reasoning and decision-making. The key methods of selecting sources, extracting and evaluating scientific arguments, and finally developing a Scientific Argument Recommender System (SciArgRecSys) are examined in a two-legged study (*Study IV*) reported in Chapter 6.

*RQ4b. What effect does the provision of external scientific argument recommendations has on Sensemaking (SM) and Quality of Deliberation?*

This research sub-question aims to assess the impact of scientific argument recommendations on our two main improvement metrics. To assess sensemaking we used the same 9 metrics used to assess the summariser, while in order to assess the quality of deliberation we used a mix of quantitative metrics of perceived quality in terms of Engagement (Eng), Mutual Understanding (MU), Aesthetics (Aes) together with standard metrics of Social Network dynamics. To examine the effect in those dimensions, we carried out a comparative study with and without SciArgRecSys artefact deployed in a live online discussion platform. We report this study (*Study V*) in Chapter 7.

The same study was also used to address RQ5, which asks what is the effect in the aforementioned dimensions of both synoptical summariser and SciArgRecSys artefacts in combination when they are deployed in a live online discussion platform. Having split research questions into sub-research questions and assigned an experimental study for each, the research diagram presented in 1.1, is now reshaped as

depicted in Figure [2.2](#)

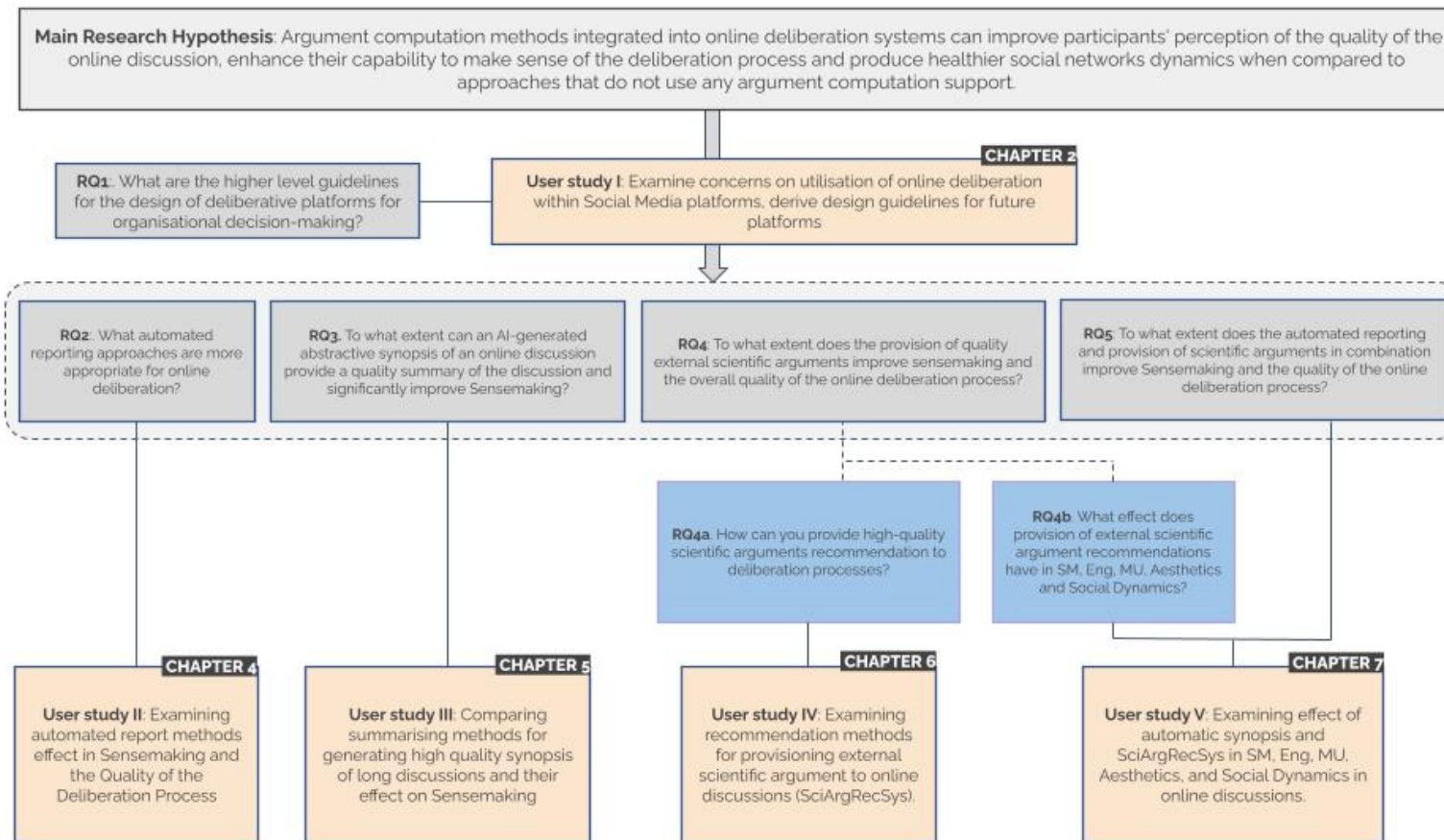


Figure 2.2: Research diagram depicting sub-research questions and corresponding studies

Each of the research questions presented in the previous Section 1.3 will be addressed using an individual investigation in order to provide distinct answers. Specifically, as shown in diagram 1.1, we carried out the following studies:

- *User study II* (targeting **RQ2**): A cross-comparison of automated reports methods measuring their effect in Sensemaking and the Quality of the Deliberation process.
- *User study III* (targeting **RQ3**): An examination of different methods for artificially generating summaries of long discussions and their degree of proximity to human-generated summaries.
- *User study IV* (targeting **RQ4a**): An evaluation study of the quality of mining and recommendation of scientific arguments and the impact of scientific arguments recommendation on discussion participants' sensemaking.
- *User study V* (targeting both **RQ4b** and **RQ5**): An evaluation study of enabling automatic synopsis, reflective nudges and scientific argument recommender in an online discussion platform and their effect on participants Sensemaking and the quality of the deliberation (in terms of Mutual Understanding, Aesthetics and Social Dynamics)

## 2.5 Summary

In this chapter, we took an inductive approach to explore the main aspirations and fears of experts for the use of technology in deliberation settings. From those, we funnelled down to the four research questions that drive the rest of this thesis. We will specifically look into the use of automated reporting in online deliberation contexts to improve sensemaking (RQ2), the quality characteristics of a synoptical summariser of online discussion (RQ3), the provision of reliable scientific arguments



to promote evidence-based reasoning and the overall quality of the deliberation (RQ4) and the impact of both in Sensemaking and Quality of Deliberation when deployed in a real deliberation platform (RQ5). Before proceeding to discuss the creation of corresponding artefacts and their evaluation studies, we provide in the following chapter a review of the pertinent literature. That chapter includes a section delineating the motivation underpinning this thesis. Furthermore, we elucidate the rationale for concentrating on the four aforementioned research questions.

# Chapter 3

## Literature Review

### 3.1 Online Deliberation as a Mean of Deliberative Democracy

Deliberative democracy refers to the mode of democratic governance that focuses on the process of deliberation - that is reasoned discussion and the consideration of different arguments - as the mean to reach a decision. It posits that political decisions should be the product of fair and inclusive deliberation and not just aggregated preferences ([Gutmann and Thompson, 2004](#)).

Public consultation and public deliberation are both processes of deliberative democracy. Public consultation typically involves soliciting public opinions, comments, or feedback on specific proposals or issues ([Fishkin, 2009](#)), whereas public deliberation involves processes that encourage discussion and dialogue among members of a large community for any civic policy issue ([Shane, 2004](#)). In this thesis, we solely focus on public deliberation.

Large organisations employ public deliberation methods and techniques in their decision-making processes. It is well understood that they benefit from the large scale of contributions ([Fuchs, 2008](#)), the idea synergy and synthesis that occurs in this

collective environment (Smiraglia, 2014; Aladalah et al., 2016), harnessing, in this way, the “wisdom of the crowds”<sup>1</sup> - a central idea in collective intelligence research (Malone et al., 2009). Moreover, the attainment of transparency in the process is crucially beneficial in effective government *accountability* (Schaeffer and Yilmaz, 2008) and *trust* (Wang and Wan Wart, 2007), and leads to better designed strategies, for instance in designing mitigative strategies in hazard, risk and vulnerability (HRV) analysis (Pearce, 2005).

## 3.2 Public Deliberation for Wicked Problems

Organisations will often face difficult questions regarding their strategy and operations, e.g. their financial planning (Ehrgott et al., 2004), design of marketing strategies (Kwak et al., 2005) or personnel evaluation and selection (Liang and Wang, 1994). There is an inherent difficulty to reach a consensus in these questions as multiple alternatives exist, multiple criteria to evaluate solutions exist and various actors with different magnitudes of influence exist (Mahyar et al., 2017). Originally within city planning, Rittel and Webber (1973) introduced the notion of *wicked problems*. They argued that urban planners deal with different problems than natural scientists and engineers face, which are definable and have solutions that have well-defined solutions (*tame* problems). Contrarily, wicked problems appear in environments with heterogeneous viewpoints, with participants holding different agendas and often disparate -if not contradicting- purposes. While “wicked” problems cannot be formally defined, there are 10 common characteristics -according to Rittel and Webber’s definition (Rittel and Webber, 1973)- they share:

- There is no definitive formulation of a wicked problem

---

<sup>1</sup>The term “The Wisdom of Crowds” first introduced by James Surowiecki in the same-titled book in 2004 Surowiecki (2005), received much attention in management and forecasting domains (either applying it to prediction markets, Delphi methods or human swarming methods

- Wicked problems have no stopping rule
- Solutions to wicked problems are not true-or-false, but good-or-bad
- There is no immediate and no ultimate test of a solution to a wicked problem
- Every solution to a wicked problem is a “one-shot operation”; because there is no opportunity to learn by trial-and-error, every attempt counts significantly
- Wicked problems do not have an enumerable (or an exhaustively describable) set of potential solutions, nor is there a well-described set of permissible operations that may be incorporated into the plan
- Every wicked problem is essentially unique
- Every wicked problem can be considered to be a symptom of another problem
- The existence of a discrepancy representing a wicked problem can be explained in numerous ways. The choice of explanation determines the nature of the problem’s resolution
- The planner has no right to be wrong

The importance of Rittel and Webber’s introduction of “wicked problem” is the linkage of design rationale to problem-solving. They introduced an argumentative approach to understanding the problem and the importance of the process of doing so. Wicked problems cannot be *solved* using operations research methods such as Multiple-criteria decision analysis (MCDM) as those depend on quantification of the criteria and the effect of each decision - which by definition is non-applicable to such problems. Looking into ways of discovering meaning and rationalising peoples’ actions, [Weick \(1995\)](#) introduced the concept of *Sensemaking* in the same domain of organisation science. His work focuses more on human intuition in decision-making

by creating a coherent framework to define perception, cognition, memory and action (Weick et al., 2005). We discuss further the concept of Sensemaking, its importance and relatedness to deliberation processes in the following Section 3.2.1.

### 3.2.1 Sensemaking in Online Deliberation Systems

Sensemaking examines how people make sense out of their experiences in the world. It analytically examines how people construct inner representations of the outer world and how they utilise those to solve specific tasks (Russell et al., 1993). From a psychological perspective can be seen as the expression of concerns of the individual for creativity, curiosity, comprehension, mental modelling and situation awareness (Dervin, 1999).

Depending on the alignment to a human science philosophy or theory, there are many sensemaking models in the literature, each establishing a different perspective on sensemaking. The models draw insights from distinct research areas such as Human-Computer Interaction (HCI), Organisational Communication theory, Library and information sciences, Cognitive Engineering Science, Social Psychology and Sociology. However, regardless of the field examining sensemaking, a common ground element is the attempt to understand what happens when people face situations that have not been anticipated, could not be predicted, or for whatever reason prevent the natural and predictable progression of events.

Below, we outline the major cognitive models for sensemaking that have been proposed in the literature:

**Dervin's Model** Brenda Dervin's model defines sensemaking as a time-space system where people live in a world which changes and with gaps at any given time-space (Dervin et al., 2003). At any point in time-space, for whatever reason, we have to make sense of information/knowledge and bridge gaps, therefore, Dervin's

model is often referred to as the gap-bridging model. Dervin's SenseMaking Model consists of an ongoing process that involves moving from a state of "equilibrium" (where everything makes sense) to a state of "disequilibrium" (where something doesn't make sense). In this model, individuals actively seek information to bridge the gap and restore equilibrium. The main differentiation of Dervin's model is the proposed methodology, called Sense-Making Methodology (SMM) ([Dervin and Frenette, 2000](#)), and its emphasis on the individual's perspective and the subjective nature of sensemaking. It places a strong focus on the user's information needs, cognitive gaps, and the continual process of making sense of information

**Russell's model** Learning loop context: [Russell et al. \(1993\)](#) argue that sensemakers develop and refine representations, e.g. they develop frameworks (schemas) to structure and organise the information being gathered and then when they believe that the framework is sufficient, i.e., that there is no significant data (residue) remaining from the searching, they fill-out or encode the representation with detailed content. Russell's model ([Russell et al., 1993](#)) was constructed while researchers were examining the cost aspect of sensemaking, i.e. the cost of extracting information from located information resources. For example, in one of their first use cases, they examined the cost of digesting large amounts of information for creating repair manuals for printer technicians.

**Pirolli and Card's model** [Pirolli and Card \(2005\)](#) extended on Russell's proposed model of internal learning loop and provided a notional model of sensemaking loop for intelligence analysis. They argue that information processing is happening in distinct steps and is overruled by an "information foraging" loop and a parallel "sensemaking" loop. The main premise of this process is "the production of novel intelligence from massive data". The steps involved in the bottom-up (from data to theory) Sensemaking process according to this model are:

- *Search and Filter:* External evidence and information are gathered in an information store (*shoebox*) and filtered according to relevance.
- *Read and Extract:* Information inside the *shoebox* is read and snippets of potentially useful for theorising are extracted.
- *Schematise:* Information can now be visualised or schematised and can be built into little stories for interrogative questions (what, why, what, etc.)
- *Build case:* A theory is built based on evidence from previous step to support hypotheses
- *Tell story:* The theory developed in the previous step is shared with the audience of interest.

In the reverse process (top-down or from theory to data), the processes of re-evaluate, search for support, search for evidence, search for relations and search for information are followed.

From the models of sensemaking reviewed above, in the remainder of this PhD research, we will focus on one of the most influential models, the Pirolli and Card's model of Sensemaking. [Pirolli and Card \(2005\)](#) extended on Russell proposed model of internal learning loop and provided a notional model of sensemaking loop for intelligence analysis. As such this model provides the right components to conceptualise and analyse computational support and human sensemaking as two separate and parallel processes, and therefore is the most suited to investigate computational approaches to improve Sensemaking.

As introduced in the Motivation Section [1.2](#) of the Introduction chapter, Sensemaking support for online deliberation processes is limited and comes with challenges. One of the two aims of this PhD research is to investigate computational argumentation to improve participants' sensemaking in online deliberation processes. From the

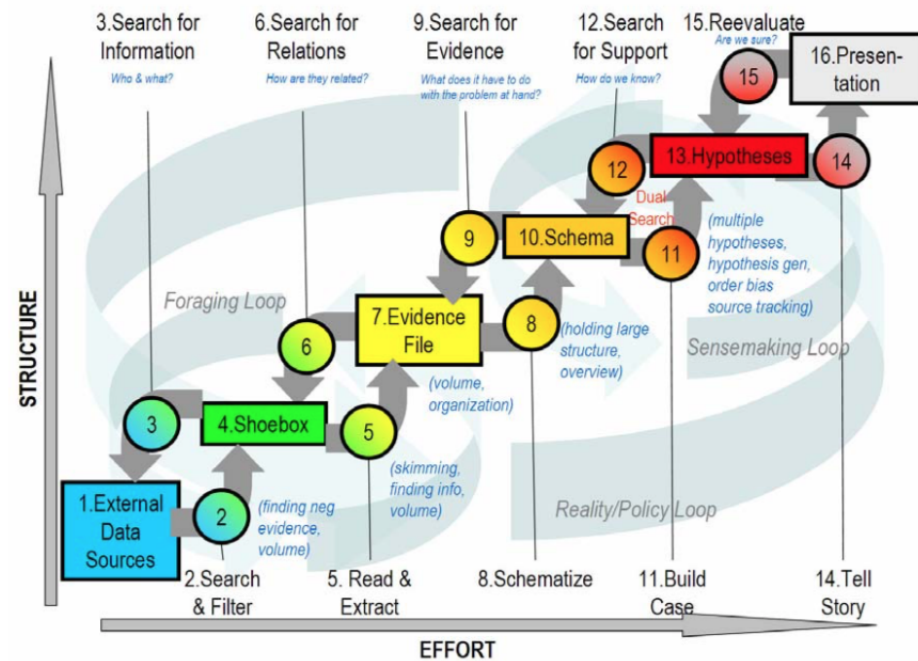


Figure 3.1: Pirolli and Card sensemaking model

models of sensemaking reviewed above we have selected Pirolli and Card’s model as the most appropriate to scaffold our review of online deliberation platforms, and to identify what sensemaking support is lacking in today’s deliberation systems (see Section 3.3.2). This will help further specify the research gap and focus our research questions.

Turning now our attention to the second aim of this PhD research, which is, improving the overall quality of the online deliberation processes, in the following we proceed to review metrics of quality of deliberation. This will provide selected metrics that we’ll use, again, to scaffold the review of online deliberation platforms and identify what deliberation quality metrics are not adequately supported by today’s deliberation systems (see Section 3.3.2). The review also informs the deliberation metrics set that we will then use in evaluation studies II 4 and V 7.



### 3.2.2 Deliberation Quality Metrics

There have been research attempts to develop a metric to measure the quality of deliberation, however, no consensus has been achieved as to what set of measures shall be used to assess the quality of a debate or a deliberative process. For example, [Stromer-Galley \(2007\)](#) proposes a coding scheme for qualitative analysis of deliberative discourse. The coding scheme is operationalised in the following identified elements of what can be considered healthy deliberation:

- Reasoned Opinion: Examines whether opinions are reasoned, expressed as claim with evidence that can be confirmed or denied.
- Sourcing: The type of evidence or reference; whether personal or external, trustworthiness, subjective or not is important to assess the expressed opinion.
- Disagreement: is a sign that participants have distinct views and that the problem is indeed significant and needs a solution.
- Equality: A healthy deliberation assures equal participation to everyone, with no participant dominating the discussion.
- Topic: The conversation stays on topic and does not diverge from the defined issue.
- Engagement: Examines whether participants engage with others, e.g. respond to their comments, or just engage in parallel monologues.

Similar qualitative methods use similar coding schemes. [Graham and Witschge \(2003\)](#) draw from Habermas' Theory [Habermas \(2015\)](#) for democracy and introduce the notion of *understanding* in different stages: Reciprocity (similar to Strommeyer-Galley's Engagement) as an initial step, a deeper level of understanding via reflection of other's opinion called *reflexivity* and as last stage *empathy* - to put yourself in

other shoes. Also, prior bias and beliefs could damage the quality of a debate. It is well established that people are often unable to escape their prior biases and beliefs. [Taber and Lodge \(2006\)](#) demonstrated that people are highly biased when they evaluate political arguments - especially when those are highly polarised. In their experiment, they observe that participants, when reading pro and con arguments, counter-argue the contrary arguments and uncritically accept supporting arguments. Following these coding schemes, [Muhlberger and Stromer-Galley \(2009\)](#) attempt to create automated measurement techniques of deliberation quality, still basing their input on manual coding. Generalizing to online discussion systems, there have been attempts to measure quality of such systems. For instance, [Kay \(2006\)](#) created a set of variables assessing the quality of a discussion board in 12 distinct areas (such as social learning, cognitive processing, quality of discussion, navigation issues and others). Within Computer Supported Collaborative Learning (CSCL) field, there have been models that try to explain certain participation behaviours, e.g. [Goggins and Xing \(2016\)](#), while others attempt to build participation and engagements metrics [Zhu et al. \(2016\)](#) using network analysis techniques in an attempt to build a predictive model for the quality of discussion boards.

We can also identify a different approach for evaluating the quality of argumentative discussions, based on Frans H. van Eemeren and Rob Grootendorst pragma-dialectical theory [Van Eemeren and Grootendorst \(2016\)](#). The pragma-dialectical theory, an evolution of Arne Naess's [Naess and Hannay \(1968\)](#) principles for effective discussion, regards an ideal discussion where argumentation has a central role, consisting of four stages:

- *Confrontation Stage*: A difference of opinion emerges between parties. The issue at hand and standpoints are identified.
- *Opening Stage*: The parties establish a shared set of rules and definitions

to facilitate the discussion. Commitment to resolve the dispute critically is affirmed.

- *Argumentation Stage*: Each party puts forth arguments to justify their standpoint and responds to the other party's arguments. The goal is to systematically test assertions.
- *Concluding Stage*: The parties determine if the difference of opinion has been resolved based on the arguments and evidence presented. If not, the discussion may continue or be postponed.

The ideal model of critical discussion as outlined by van Eemeren and Grootendorst in their pragma-dialectical approach establish a framework consisting of the following rules:

- *Freedom rule*: Parties must not prevent each other from advancing standpoints or casting doubt on standpoints.
- *Burden of proof rule*: A party that advances a standpoint is obliged to defend it if asked.
- *Standpoint rule*: A party's attack on a standpoint must relate to the opponent's actual formulation of that standpoint.
- *Relevance rule*: A party may defend standpoints only by advancing argumentation related to that standpoint.
- *Unexpressed premise rule*: Discussants should not falsely attribute unexpressed premises to the other party.
- *Starting point rule*: Discussants agree on certain statements as points of departure for the discussion.

- *Argumentation scheme rule*: Arguments used in the discussion must be logically valid and have premises acceptable to the participants.
- *Validity rule*: The reasoning in the argumentation must be logically valid or capable of being rendered valid by adding missing premises.
- *Closure rule*: Failures to properly apply relevant rules of discussion should be pointed out.
- *Usage rule*: Parties must not use expressions in an unusual way without due explanation.

We identify [Stromer-Galley \(2007\)](#), [Graham and Witschge \(2003\)](#) and [Kay \(2006\)](#) debate quality metrics as the most appropriate to measure the quality of deliberation in this PhD research.

[Stromer-Galley \(2007\)](#) and [Graham and Witschge \(2003\)](#) debate quality metrics will be used in the technology review (in the next Section 3.3) to annotate if a deliberation platform is promoting a given quality of discussion feature. In *Study II* (Chapter 4), we will use a merged set of variables proposed in [Graham and Witschge \(2003\)](#) and [Kay \(2006\)](#) to construct a questionnaire to quantitatively assess the quality of the deliberation by measuring participants' perceptions of discussion quality.

### 3.3 Online Deliberation Current State and Issues

In the previous sections, we outlined the value of public deliberation for organisations in general; we now focus on technology-mediated communication and proceed to examine mechanisms of how public deliberation is currently implemented in online environments, with a focus on the *consolidation* and *reconciliation* phases.

Over the past 30 years (since the advent of the world wide web), Information and Communication Technologies (ICT) have been integrated into public consultation

initiatives. We review here those software technologies that support online deliberation. Many technologies were not designed for this purpose, e.g. blogs, forums, and message boards. However, this technological review aims to capture purposed or not tools that have been used extensively in public participation activities.

For clarity reasons, we define the concept of deliberation as the process in which participants engage in a reasoned opinion expression about an issue in an attempt to identify solutions to a stated problem and evaluate these suggested solutions [Roberts \(1997\)](#). [Bächtiger and Parkinson \(2019\)](#) extend this definition to its democratic aspect: “Democratic deliberation is about using that mode in an inclusive and equal manner, oriented towards an effective, collective decision point and on into implementation”. Deliberation on a given issue of a community spans and progresses through a number of phases ([Velikanov and Prosser, 2017](#)). Initial phases correspond to *ideation* and *consolidation* where ideas are proposed, discussed, edited and evaluated. Later phases correspond to *reconciliation* phase where proposals are aggregated and iteratively reevaluated and finally the *selection* phase - where a winning proposal is selected for implementation.

Existing solutions for public consultation and online deliberation can be arranged in the following three categories, according to the anchoring concept of the participant contribution:

- Time-Centric Systems: Content is organised on a temporal basis (*when* it was contributed). Typical examples are email and chat rooms where usually posts appear in timely order, most recent first (or opposite). In general, time-centric systems thrive when it comes to the scale of participants but lack efficiency for public consultation purposes due to the scattering of information (as evidenced in [Aragón et al., 2017a](#)).
- Question-Centric Systems: Contributions aim to answer a central question, the

most representative example of such systems are *Question-Answering* systems, e.g. [stackoverflow.com](#). They usually focus on one domain and thrive in answering questions that are easily verified for correctness, e.g. *what is* type of questions. However, they have weak mechanisms to show the rationale or narrative of the contributor. Often answers contain duplicate arguments (pieces of information that have been mentioned in other answers) and do not promote collaboration on the level of each answer but rather are usually flooded by many shallow and overlapping comments (shown at [Murphy \(2004\)](#)).

- Issue-Centric Systems: Participants interact by not only providing their ideas, comments and therefore arguments but also explicitly linking those, creating deliberation argumentation maps ([Kirschner et al., 2012](#)). Such augmentation in the deliberation process enables more systematic and structured discussion leading to healthier participation ([De Liddo et al., 2012](#)), harnessing the collective opinion and intelligence ([Bonabeau, 2009](#)). Also, the provision of evidence in arguments - evidence-based reasoning - is directly linked to better decision making ([Evans et al., 1993](#)) (this link consists of the spine idea of this doctoral work). In addition, they help to build a shared understanding of the discourse ([Conklin and Begeman, 1988](#)) which, as discussed in Section 3.2, is essential when tackling wicked problems. The key weaknesses of such systems orbit around the inherent complexity of the user interfaces and argumentation technologies for conversation, e.g. vagueness of concepts require -in advance- definition of the argumentation scheme in place and training in using the argumentation diagramming tools.

Typical examples of the above systems can be found in Figure 3.2.

According to [Klein \(2015\)](#), the above simple taxonomical approach for deliberation systems can be expanded by adding 2 categories of debate-centric and argument-

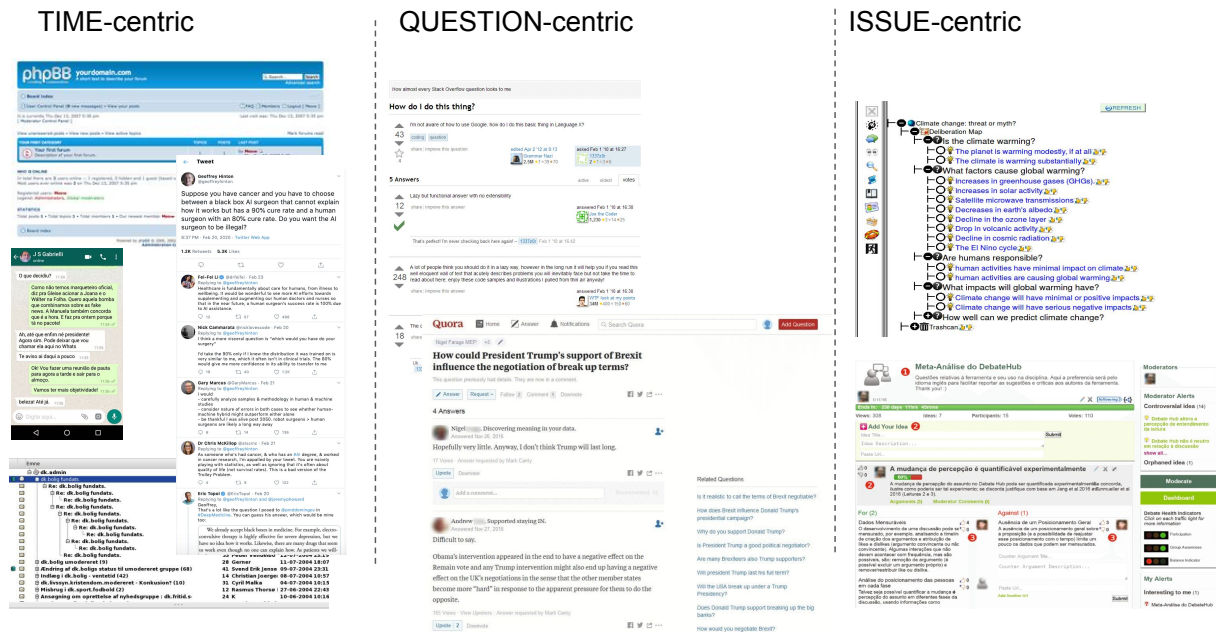


Figure 3.2: Deliberation systems examples

centric systems. This works on the premise that issue-centric systems are essentially topic-centric systems, without the argumentation enactments in place. Following this classification scheme, debate-centric systems are topic-centric systems with a structured argumentation in place (usually organised around pro and con arguments columns) and argument-centric systems are topic-centric systems that allow for complex deliberation to be captured with the use -again- of argumentation schemes (contributions are organised around ideas, issues and arguments). We present this expanded taxonomy in Table 3.1 with typical example platforms, along with details on where they thrive and where they are disadvantaged.

Centrality	Technologies	Platforms example	Advantages	Disadvantages
Time	Email Forums micro-blogging chat	Twitter WhatsApp GMail phpBB /vBulletin forums	Crowd-scale participation No learning-curve Quick ideation	Low signal to noise Insular ideation Balkanization Non-comprehensive coverage Dysfunctional argumentation Opacity
Question	QnA	StackOverflow Quora	Excellent at close-ended questions Attraction of domain experts	Domain scope Weak in open-ended questions High level of redundancy Non-collaborative
Topic	Wikis	Wikipedia Loomio	Collaborative authoring Quality content Up-to-date content	No transparent argumentation No deliberation mechanism Problematic conflict resolution
Debate	Debate platform	ProCon.org  debate.org	Argumentative structure  Can cover controversial topics Broad coverage of argumentation	No mechanism of connecting arguments No depth in open-ended questions
Argument	Debate platform	DebateHub Consider.it	No duplication Transparent relation Encourages evidence-based Collaborative argumentation	Steep learning curve Low engagement Excessively complex argumentation networks

Table 3.1: Public deliberation solutions taxonomy and review



### 3.3.1 Overview of Online Deliberation Systems

We proceed to examine the results presented in Table 3.1. Regarding support for collaborative work, Topic-centric systems perform better than others, while Argument-based systems due to their non-intuitive interfaces perform poorly regarding this aspect. For the same reason, Argument-based systems also have low engagement, while other systems, e.g. time-centric systems, achieve high levels of participation. However, the kind of participation carried out on time-centric platforms is inferior. We observe phenomena such as platform islands or Balkanisation (where users self-assemble to subgroups that reinforce their original biases) and solo ideation (where ideas are shallow contributions by solo individuals rather than the intellectual outcome of fruitful argumentative process) (Helm, 2017). Furthermore, anonymity in these platforms encourages a sense of impunity that leads to toxic behaviours (Hardaker, 2010) such as trolling or flame-wars. In contrast, other platforms and especially topic-centric, through mechanisms such as moderation achieve reasonable levels of respect and reciprocity among their participants (Dencheva et al., 2011). However, this is done by sacrificing other aspects of good deliberation such as the lack of transparency in argumentation and decision-making (“why this decision was taken?”). Also, apart from debate and argument-centric platforms, there is no native deliberation mechanism, e.g. polling or voting. Furthermore, whereas debate and argument platforms enforce (either a weak or strong) argumentative structure to facilitate discussion, time and question-centric systems do not, leading to highly redundant content, dysfunctional argumentation and superficial discussions. The weakness of such systems is that they do not support the thorough examination of complex problems nor offer depth in the discussion, however, they are most suitable for non open-ended questions (that can easily be verified with factual evidence). Having reviewed and analysed the technological types of deliberation platforms, we will proceed in Section 3.3.2 to review technological solutions (platforms) that

implement online deliberation systems. This review will be carried out in terms of support of sensemaking as defined in Section 3.2.1 and according to the set of elements of what comprises a good quality of debate as introduced in Section 3.2.2.

### 3.3.2 Computer Supported Public Deliberation Platforms

#### Review

#### Selection of platforms

Several technological platforms can be used in an angled way to serve online deliberation. Below we are examining platforms that have been explicitly designed to serve deliberation, i.e. omitting generic purposed platforms like email, forums, chat, etc. We also reviewed platforms that explicitly support decision-making while we excluded from this review platforms that are just reconfigured forum platforms (e.g. a dry list of statements pro and con about a topic).

**DebateHub**<sup>2</sup> A tool for online communities to raise issues, and debate ideas with pro and con arguments. It follows a minimised IBIS model, aiming to provide an intuitive interface to users that require no training and minimal time for familiarisation.

**Truthmapping**<sup>3</sup> A tool for diagramming argument maps. The logical structure of an argument needs to be stated explicitly to a premise and conclusion, e.g. A stands, therefore B. Organises the created maps into high-level categories (e.g. Education, Religion, Law, etc.) and invites collaboration via agreeing and contributing critiques or rebuttal to critiques to established maps.

---

<sup>2</sup><https://debatehub.net/>

<sup>3</sup><https://www.truthmapping.com/>

**Debategraph**<sup>4</sup> A tool for creating argument maps. Provides a unique way of exploring maps by “bubble view” - a navigation by diving in and out nodes of argumentation. Allows the linking of ideas with several types of links with different semantics, e.g. Advocacy, Citation, Causation, Explanation, Inconsistency, and many others.

**Argunet**<sup>5</sup> A desktop application for creating and editing argument maps. It constructs arguments as a series of propositions allowing through a rich set of semantic relations among them. It is not a pure online application but allows collaborative work only with the client application. The created argument map, though can be rendered for online use and embedded in third-party websites. Argunet as of October 2018 is no longer supported but rather authors of the software now focus on the creation of Argdown<sup>6</sup>: a lightweight markup language for defining argumentation.

**Cohere**<sup>7</sup> A visual tool to create, connect and share ideas. It organises information in an argumentative diagrammatic structure while it enables the use of argumentation schemes. It uses an augmented IBIS model semantically connecting Ideas with (external) evidence and Users, creating in this way, a three-layered network of Ideas, Groups of Users and Documents. It is aimed to be used as a tool for Contested Collective Intelligence (De Liddo and Buckingham Shum, 2010).

**Coggle**<sup>8</sup> is a mind mapping platform, with a focus on collaborative work and has been used in learning setups (CSCL<sup>9</sup>) (Papushina et al., 2016). It supports

---

<sup>4</sup><https://debategraph.org/>

<sup>5</sup><http://www.argunet.org/>

<sup>6</sup><https://argdown.org/>

<sup>7</sup><http://cohere.open.ac.uk/>

<sup>8</sup><https://coggle.it/>

<sup>9</sup>CSCL: Computer Supported Collaborative Learning (Halavais, 2016) is the area about technologies that can support the pedagogical approach that claims that knowledge can be acquired

the integration of external media (images, audio, video) and has some shared collaborative work features, e.g. historical log of changes, differentiation between editions, branching and merging user's editions.

**DebateArt**<sup>10</sup> is a debate platform. It follows the familiar structure of a discussion forum, with the unique characteristic that it splits the deliberation process into rounds, that in each round presented arguments are filtered down before going to a final round of voting.

**Debatebase**<sup>11</sup> host hundreds of debates that are commonly used in debate competitions. It offers a flat pro and con argument structure and also encourages the inclusion of scientific evidence (each argument has its bibliography).

**Kialo**<sup>12</sup> is an issue-based debate platform. It aspires to create a civilised and constructive debate space with the use of moderation, promoting reasoned opinion (through a minimal IBIS-style argumentative structure), with the aid of synopsis visualisations and implementing mechanisms for conflict resolution (Beck et al., 2018).

**Pol.is**<sup>13</sup> is a platform that primary goal is to collect feedback from large groups of people. It supports highly expressive but also highly scalable communication to inform decision-making. It can be instantiated for politics, business and classrooms. The issues discussed by the participants are supported via an underlying AI system that analyses the distributions of votes, performs topic modelling on the opinions shared and clustering (via K-means) to group users of the same opinions together.

---

collaboratively through interaction

<sup>10</sup><https://www.debateart.com/>

<sup>11</sup><https://idebate.org/>

<sup>12</sup><https://www.kialo.com/>

<sup>13</sup><https://pol.is/home>

**Loomio**<sup>14</sup> is an online tool for deliberative decision making. Through collaborative practices it creates communities around a topic and offers mechanisms for consensus building around it (Jackson and Kuehn, 2016).

**Quora**<sup>15</sup> is a social question-answer forum. Though main focus is to build a question-answering community, it can be used as structured deliberation system. As it attracts a great heterogeneity of users, questions and users' opinions, it produces a high-quality knowledge base (Wang et al., 2013).

**Debate.org**<sup>16</sup> follows a classic pro and con argument comparison outline around generic topics. It differentiates by constraining the discussion time period and defining "rounds" of debate here, top arguments qualify for the next rounds.

**ProCon.org**<sup>17</sup> is a platform to collaboratively collect pro/con arguments as those are stated by public figures around public issues (e.g. election candidates political stances).

**Citizenlab**<sup>18</sup> is a citizen participation platform that supports co-creation and consultation on city-planning issues, e.g. budgeting, urban transportation and others. The platform enables ideation, voting, commenting and discussion among registered users, and an overview dashboard for the sponsor of the activity (Deibert et al., 2019).

**Consider.it**<sup>19</sup> is an opinion-sharing platform where opinions are organised as pro and con arguments of the debated issue. Its unique characteristic is the opinion

---

<sup>14</sup><https://www.loomio.org/>

<sup>15</sup><https://www.quora.com/>

<sup>16</sup><https://www.debate.org/>

<sup>17</sup><https://www.procon.org/>

<sup>18</sup><https://www.citizenlab.co/>

<sup>19</sup><https://consider.it/>

analytics feature that groups people (and their corresponding arguments) in a horizontal axis of support - enabling to quickly assess whether a debate is polarised, and what is the common ground and navigate easily to the opposing arguments.

**Parlia**<sup>20</sup> hosts questions and arguments (opinions) about a range of topics. It enables its users in a collaborative way (wiki-fashion) to add a set of possible ideas (answers) to questions and complete with a range of supporting arguments on that idea. It is encouraging the contribution of factual evidence (and plans to have factual validity checking mechanisms) and offers a recommendation system to enable argument reuse and avoid duplication.

### 3.3.3 Technological Gap Identification Analysis

We reviewed the platform described above with a specific analytical goal, which is to identify the level of support provided by each platform to sensemaking and the overall quality of debate. We, therefore, proceeded to describe each platform according to its support in sensemaking and debate quality characteristics.

For the analysis of sensemaking support, we used Pirolli and Card's sensemaking model 3.2.1 as the most applicable in large-scale deliberation processes and because it is focused on technology mediated data processing. For each stage in this model, we denote with an X whether each platform sufficiently supports this step. Next, we used (Stromer-Galley, 2007) and (Graham and Witschge, 2003) debate quality metrics (as discussed in 3.2.2) to annotate if the platform in review is promoting each feature and denote it with a V. The results of this annotation phase can be found in Table 3.2.

---

<sup>20</sup><https://parlia.com/>

Name	Arg. scheme / CT support	Support for Sensemaking bottom-up					Debate quality				
		S&F	R&E	Sch.	BC	TS	RO	Eq	So	Und	Eng
DebateHub	Minimal IBIS, visual analytics, dashboards	●	●	○	○	○	●	○	●	●	○
Truthmapping	assumptions, conclusions, critiques, supporting thoughts, rebuttals, argument diagramming	●	○	●	○	○	●	○	○	○	○
Debategraph	I-nodes(e.g. ideas) linked with a rich set of relations	○	○	●	●	○	●	●	○	●	○
Argunet	Argument diagramming	○	○	●	●	○	●	●	●	○	○
Cohere	Augmented IBIS, argument diagramming	●	●	●	○	○	●	○	●	●	○
Coggle	Mind mapping	○	○	●	○	○	○	○	●	●	○
DebateArt	-	○	○	●	●	○	●	●	●	○	●
Debatebase	Pro/Con arguments, external scientific evidence	●	●	○	●	○	●	○	●	●	○
Kialo	Minimal IBIS, topic arguments, visual analytics, sunburst contribution diagram, flat argument	●	●	●	●	○	●	●	●	●	○
Pol.is	Opinion based	○	○	●	●	○	●	○	●	○	●
Loomio	Decision-making	●	●	●	●	●	○	●	●	○	●
Quora	Question-centric	●	●	●	○	●	●	●	○	●	●
Debate.org	Pro/Con argument, rounds of argumentation with a final decision round	●	●	●	●	○	○	●	○	●	●
ProCon.org	Database of public stated opinions	●	●	○	○	○	●	○	●	●	○
Citizenlab	Participation platform	○	○	●	●	●	○	●	●	●	●
Consider.it	Opinion based aggregator	○	●	●	●	●	○	●	●	○	●
Parlia	Opinion based aggregator	●	●	●	●	○	●	●	●	○	●

Table 3.2: Online deliberation platforms argumentation inclusion and support for Sensemaking processes five stages: Search and Filter (S&F), Read and Extract (R&E), Schematise(Sch), Build Case (BC), Tell Story (TS) and Quality of debate: Reasoned Opinion (RO), Equality (Eq), Sourcing (So), Understanding (Und), Engagement (Eng). Support level is given in 3 levels: full support ●, partial support ●, no support ○

Analysing the table, we make the following observations:

- While there is sufficient coverage of the early stages of Sensemaking process - 8 out of 17 (47%) full support of Search and Filter, 7 out of 17 (41%) full support of Read and Extract - and relatively high coverage in Schematize step 11 out of 17 (64%) full support, 3 partial support and only 3 with no support, there is lower support of the latter stages of sensemaking - 6 out of 17 (35%) has no support for Build case and only 1 out of 17 for the last step of Tell story.
- The most well-supported element of debate quality is Reasoned Opinion (9 out of 17) while the least supported element is Engagement (9 out of 17 with low level) and Understanding (7 out of 17 with low-quality support).
- No platform fully supports all stages of sensemaking, with only 3 supporting 4 out of 5 stages (Parlia, Consider.it, Kialo). At the same time, 9 platforms support up to just two stages of sensemaking.
- Same for debate quality, no platform covers all 5 selected elements of quality of debate. Only two platforms (Parlia, Kialo) support up to 4 elements of debate quality.

### 3.3.4 Defining the Research Gap

From our technological review of deliberation platforms, we observe that while there is a great interest and effort to integrate argumentation support technologies in online deliberation systems, these technologies lack support for the various stages of participants' sensemaking. Most platforms scarcely support the latter stages of the sensemaking process, which are the most challenging to enhance, and at the same time lack support for all aspects of a good debate.

A quote of [Weick \(1995\)](#) says: "How can I know what I think until I see what I



say?”. We argue that the current state of online deliberation systems does not enable participants to reflect on their own ideas.

At the same time, we observe that computational argumentation mining is flourishing, achieving significant advances over the past few years that enable what was considered not doable before. In this research, we hypothesise that the efficient use of computational argumentation mining has the potential to considerably improve online deliberation processes. More specifically, we hypothesize that argumentation mining can be utilised in a novel way to produce automated reports of the state and progress of online discussions; therefore improving the sensemaking of participants and the overall quality of the deliberation.

Thus far, we have discussed online deliberation technological types and platforms, their role in public consultation and have identified limitations in their use. We have proposed that utilisation of argumentative technologies can benefit online deliberation, and we have analysed in details what aspects of sensemaking and quality of debate need enhancements. We now proceed with the final section of the literature review, where we focus on our specific research hypothesis that “argumentation mining can be utilised to produce automated discussion reports thus improving sensemaking of participants and the overall quality of the deliberation”. We therefore review relevant literature on argumentation theory, prevalent argumentation models, argumentation mining, provide an overview of automated reporting mechanisms, text generation with the use of Large Language Models which will then inform both the design and evaluation of our technical solution.

## 3.4 Argumentation Theory, Models and Mining Approaches

Producing coherent arguments is vital to justify actions, especially in ill-structured problems ([Cho and Jonassen, 2002](#)). Or as Rittel states: “If you can tell me why you say plan A is great, and I understand your judgements, you have succeeded in objectifying your space of judgement to me. And although I might not share your judgement and might not be convinced, I understand you now.” ([Rittel \(1972\)](#), p. 394). As there is an inherent relationship between argumentation and problem-solving, the main argument for utilising argumentation and integrating it into deliberation processes is that it helps and actually enables people’s reasoning for better problem-solving ([Mercier and Sperber, 2011](#)).

Argumentation as defined by [Van Eemeren et al. \(2019\)](#), is “a social, intellectual, verbal activity serving to justify or refute an opinion, consisting of a constellation of statements and directed towards obtaining the approbation of an audience”. In simpler terms, we define argumentation as the exchange of arguments between parties with the goal of shaping or reforming an opinion.

In communicative contexts, the notion of argument is centred around the idea of providing considerations in support of or against a claim, rather than just asserting it ([Pinto, 2010](#)). In Pinto’s notion of argument, an argument consists of a claim (conclusion, standpoint) and considerations (premises, reasons) intended to provide support, justification, or grounds for that claim. The claim being argued for may be explicit or implicit, and the considerations need not deductively entail the conclusion, but should be reasonably taken to support it. The function of argument is rationally persuasive rather than logically probative. Arguments aim to provide good reasons for thinking the claim is true, likely, or acceptable, not irrefutable proofs.

We therefore, within public deliberation, define argumentation as the communicative

activity to mediate exchange and critical evaluation of pro and con reasons for problems and disagreements (Mohammed, 2016).

Argumentation theory has long-standing history that can be traced back to ancient Greek philosophers (e.g. Aristotelian logic (Lear, 1986)), and throughout history, philosophers, logicians and theorists explored the question of what makes an argument sound and correct, looking for logical fallacies and examining errors in reasoning.

Walton (2009) tries to pin down the tasks involving argumentation as follows:

*Identification*: here the premises and conclusions of an argument are identified and attempted to fit them to an argumentation scheme (a known form of argument)(Walton et al., 2008).

*Analysis*: in this task, implied arguments are discovered and made explicit as a preparation for the evaluation step

*Evaluation*: is the determination of whether an argument is weak or strong.

*Invention*: is the construction of new arguments to support or prove the conclusions of the argument.

It is useful to make the distinction between argument as a product and argument as a process (Reed and Walton, 2003). *Argument as a product* refers to the static structure of an argument - the conclusion and the premises that support it. This views an argument as the end result or finished product of reasoning. *Argument as a process* refers to the dynamic reasoning activity involved in constructing, presenting, interpreting, criticizing, and revising arguments. This views argumentation as an ongoing social and cognitive process.

Along the same lines, a minimally viable definition of an argument would be a statement (proposition) consisting of a premise, a conclusion and an inference (a link) between the premise and the conclusion. This definition often occurs within academic writing context, e.g. Potter (2006). Using diagramming, as a simple yet effective way to visualise and analyse argumentation, a single argument would look

like in Figure 3.3.

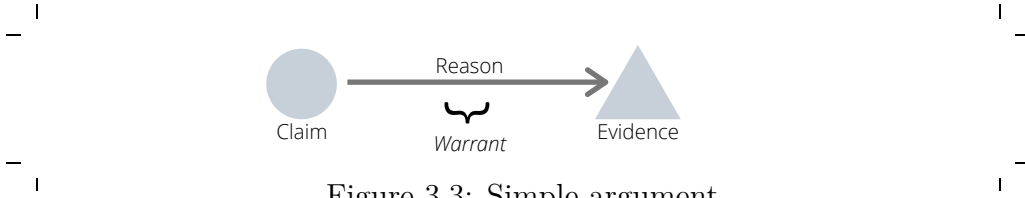


Figure 3.3: Simple argument

Argumentation can be then seen as a chain of arguments (Bentahar et al., 2010), where arguments can be used as one of the essential parts of a greater argument, connected to external arguments either to support or refute other arguments. A different but highly influential and popular model of argument would be Toulmin's model of argument (Toulmin, 2003). In this model, you have an incontrovertible *datum* **D** that is connected via a *warrant* **W**, to a -perhaps subjective, therefore controvertible- *claim* **C**. In this case, the warrant is the inference rule while data can have a variety of functions such as to support a claim, as an explanation, a justification, or a rebuttal. Figure 3.4 depicts a diagrammatic representation of Toulmin's model. Below we define the key components of it:

- Claim (*C*): An assertion that can potentially have controversial nature
- Data (*D*): Facts or established beliefs related to the claim
- Warrant (*W*): The inference rule that connects data with claim.
- Backing (*B*): The foundation on which the warrant is based and therefore be trusted
- Qualifier (*Q*): The degree of certainty associated to the claim.
- Rebuttal (*R*): A statement that covers situations that the Claim can be defeated.

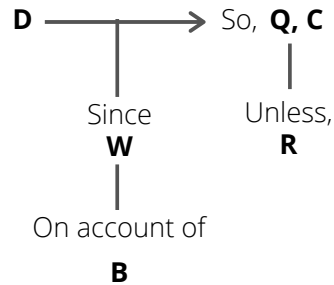
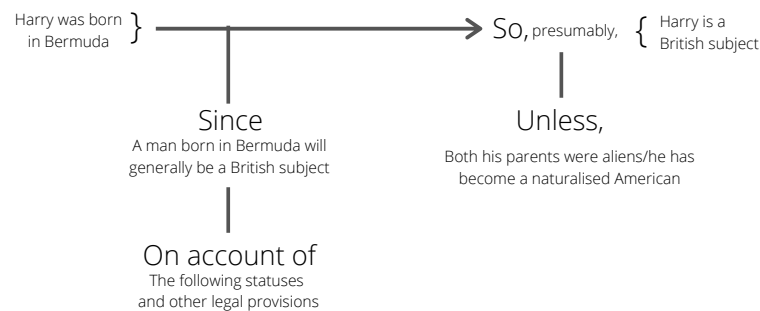


Figure 3.4: Tulmin's model of argumentation diagram

Figure 3.5: Argumentation illustration example based on Tulmin's model depicted from [Toulmin \(2003\)](#)

Though highly effective, such a model requires a representational fit to practically model a certain field. For example, Tulmin's model originates from the philosophy of law and has been the principal model followed by early attempts of building computer-supported collaborative argumentation software for teaching problem-solving to undergraduate Law students ([Carr, 2000](#)).

Attempting to contextualise argumentation on a different domain -in discussion systems- and as a proposal to deliberate around “wicked” problems (see paragraph

3.2), [Kunz and Rittel \(1970a\)](#) proposed a model called IBIS (Issue Based Information System), depicted in Figure 3.6, which is composed of three concepts (Arguments, Positions, Issues) and nine relationships like supports, objects-to, replaces, temporal-successor-of or more-general-than. It conceptualises the main interactions happening in problem-solving tasks, but more than a conceptualisation schema, Issue-based information systems (IBIS) is essentially a method for structuring and documenting design rationale in complex design problems, such as those arising for example in projects of collaboratively constructing large networks. Early transfers of IBIS model to the web consists of gIBIS ([Conklin and Begeman, 1988](#)) which in turn led to the evolvement of argumentation-based deliberation tools, such as MIT deliberatorium ([Klein, 2011](#)) and all other tools reviewed in Section 3.2 and 3.3.2.

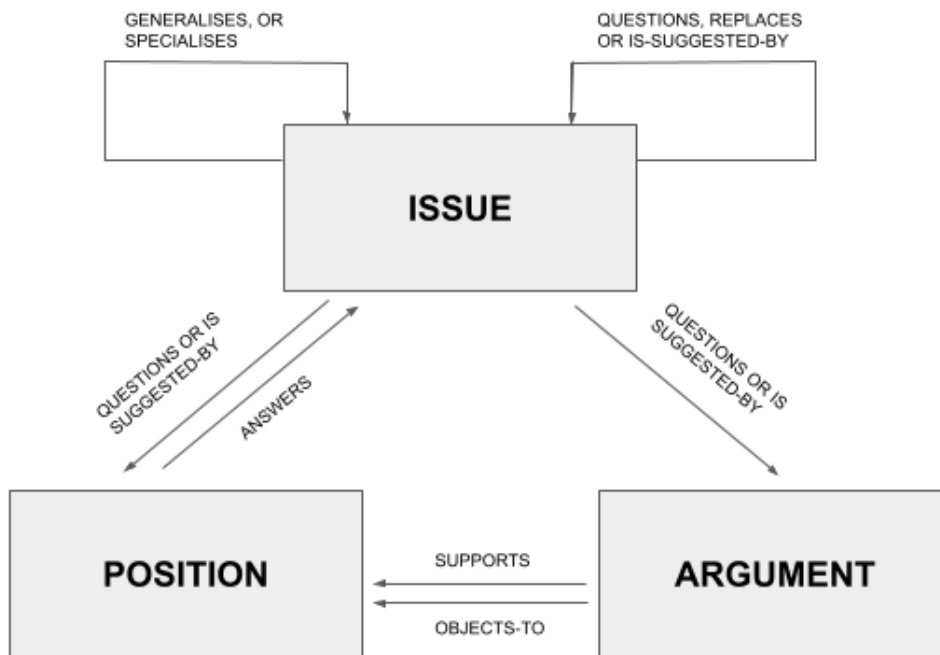


Figure 3.6: IBIS - Issue-Based Information System elements and interactions

Public debates on “wicked” problems or online discussions in general and deliberative dialogues where group decision-making takes place are some examples of *argumentative polylogues* ([Lewiński and Aakhus, 2014](#)). Argumentative polylogues refer to complex argumentative discussions involving multiple participants and viewpoints.

They have different features that make dyadic dialectical analysis problematic for the purpose of evaluating argumentation, for example may not adequately perceive the strategic rationality behind the choices of arguers engaged in polylogical encounters. It is worth noting that this network of discussions involving different players, positions and places usually sparse on the web presents an additional challenge in computationally detecting argumentation patterns. Nevertheless, early endeavours have been made, such as [Musi and Aakhus \(2018\)](#); [Perret-Clermont et al. \(2019\)](#); [Aakhus and Lewiński \(2017\)](#), which involve a fusion of manual analysis and automated argument mining techniques.

**Computational approaches for argumentation mining** Argumentation mining (AM) aims to extract structured arguments from unstructured natural language text. A more extensive definition would be: “analyzing discourse on the pragmatics level and applying a certain argumentation theory to model and analyze textual data at hand” ([Habernal and Gurevych, 2017](#)). A relatively young field of Computational Linguistics ([Green et al., 2014](#)) has attracted the attention of researchers from the broad Artificial Intelligence area, more specifically the Natural Language Processing and Knowledge Representation and Reasoning areas. It differs from other well-established areas of Opinion mining or Sentiment analysis, as it looks at a deeper analysis of the public discourse. For instance, opinion mining looks to detect *what is* the opinion of the public towards a topic, e.g. a product, while Argumentation Mining looks to answer the *why* is that certain opinion.

Computationally, Argumentation Mining involves a set of variety of interconnected NLP tasks. However, we can group those tasks into fundamental steps that comprise the following processing pipeline:

- Argumentative sentence detection (sentence classification): This first step in the argument mining pipeline involves the extraction of sentences from the

input text that contain an argument. Usually, this is formulated as a sentence classification task, either with the use of a binary classifier (to distinguish argumentative sentence or not), a set of binary classifiers (one for each type of component) in the case that we assume that a sentence contains more than one type of argument component or a multi-class sentence classifier.

- Argument component boundary detection (boundaries detection): In this step, the exact boundary of the argument is detected and decided. As a segmentation problem, it can be applied on a sentence level, e.g. an argument spanning among several sentences, but can also be applied on a token level e.g. in the case that an argument clause is contained within a portion of a sentence.
- Argument structure prediction (relations prediction): This step aims to predict links between the arguments of the previous set of discovered arguments.

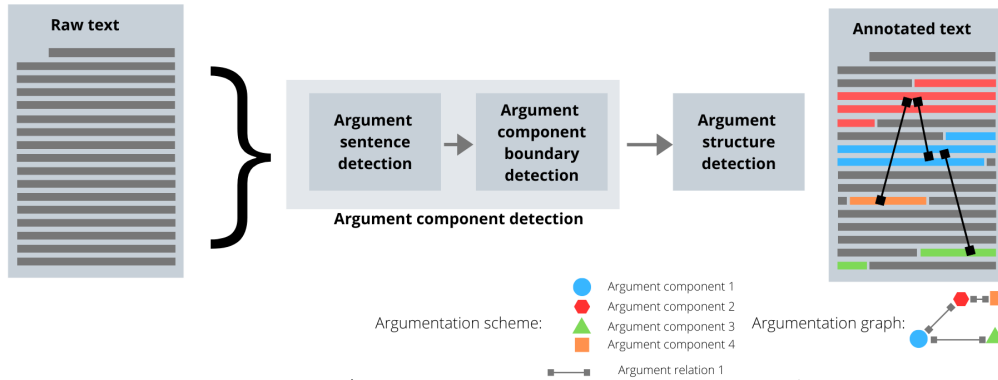


Figure 3.7: Argumentation mining NLP pipeline

The output of the last step of the above pipeline is a structured graph, comprising nodes of argumentation components, e.g. claim, premise, and edges representing different relation entailments according to the argumentation scheme selected. For example, if a monological model is followed, the graph's nodes would be claims and evidence connected with support relations. The steps of the above AM pipeline draw



some similarities with NLP tasks. Table 3.3 displays the correspondence of AM pipeline task to traditional NLP and Machine Learning tasks.

Argumentation mining task	NLP/ML tasks
Sentence detection	<p>Sentence classification (categorizing or labeling individual sentences into predefined categories or classes)</p> <p>Hedge cue detection (identifying linguistic cues or expressions that suggest uncertainty, vagueness, or qualification in the statement being made)</p> <p>Sentiment analysis (determining the sentiment or emotional tone expressed in a piece of text)</p> <p>Question classification (categorising or labelling a given question into predefined classes or categories based on its intended meaning)</p> <p>Subjectivity prediction (determining whether a piece of text, is subjective or objective)</p>
Boundary detection	<p>Sequence labeling (assigning a label or category to each element in a sequence of tokens)</p> <p>Named entity recognition (identifying and classifying entities within a text)</p> <p>Text segmentation (dividing a continuous stream of text into meaningful segments or units)</p>
Relations prediction	<p>Link prediction (analysis of textual documents to predict connections between them)</p> <p>Discourse relation classification (determining the type of relationship or connection between different segments of text within a document or discourse)</p> <p>Semantic textual similarity (quantifying the degree of similarity or relatedness between two pieces of text based on their meaning or semantics)</p>

Table 3.3: Argumentation mining correspondence with NLP/ML tasks

The first attempts of computational argumentation mining can be traced back to 2010. A few notable approaches would be for argumentation in legal texts (Palau and Moens, 2009; Mochales and Moens, 2011) and in analysis of argumentative and rhetorical structure in scientific literature texts (Teufel et al., 2009).

Various models have been proposed for Argument component detection, i.e. the composite step of argumentative sentence detection and boundary detection, including early attempts using Naive-Bayes (Moens et al., 2007) and SVMs (Mochales and Moens, 2011). More recently approaches using LLMs have been proposed such as AMPERSAND (Chakrabarty et al., 2020) or BERT-based (Lugini and Litman, 2021). Relations prediction is a necessary step to detect argument structures. Early approaches included textual entailment methods (Cabrio and Villata, 2013) and SVMs (Habernal and Gurevych, 2017), while more recent work focuses on fine-tuning LLMs, e.g. (Reimers et al., 2019).

As shown in literature surveys on Argument Mining (Lippi and Torroni, 2016a; Cabrio and Villata, 2018; Lawrence and Reed, 2020), some common challenges faced by early AM models and persist to the present day are: (i) Argumentation Modeling - the selection of model (argumentation scheme) according to the domain, (ii) Subjective Interpretation - an argumentative text may have multiple valid interpretations of its structure, (iii) Lack of large quantities of appropriately annotated arguments to serve as training and test data, and (iv) Automatic identification of argumentation schemes remains a major challenge with state-of-the-art systems having low accuracy. Indicatively, we mention that most recent systems, e.g. (Chernodub et al., 2019), achieve -a relatively low- micro-F1 score of 64.54% (on their corresponding corpora and over their corresponding scheme). In the case of Margot web service (Lippi and Torroni, 2016b) -which we use in Study II in Chapter 4 - the task of claim detection reportedly achieves F-measure of 66.6% and evidence detection 90.7%.

## 3.5 Automated Reporting

We examine below methods that can be utilised to generate automated reports.

**Text simplification** is the NLP task that modifies the syntax and lexicon of natural language to reduce its length and complexity, therefore improving its understandability and readability. Automating this process comes with great challenges. Typical approaches for text simplification include:

- *Lexical approaches*: in this approach difficult words are replaced with easier (more readable or understandable) words while preserving the underlying meaning of the original text. This is done with the aid of external lexicons, e.g. WordNet (Miller et al., 1990), to discover synonyms (used e.g. in (Carroll et al., 1998)) or by using word vector space models to discover semantically close words, e.g. (Sahlgren, 2006; Ma et al., 2018).
- *Syntactic approaches*: in this approach sentences with difficult syntactic idioms (e.g. subordination, relative clauses, irregular word order) are converted to simpler ones that convey the same message without hindering their readability. Initial approaches (e.g. (Chandrasekar et al., 1996)) require hand-crafting simplification rules that, assuming an accurate parsing of syntactic features, can lead to precise systems; it is however time-consuming and lack coverage. In this approach, a concern is also the order and readability of the generated text, for example, avoid applying transformation rules blindly as they break text cohesion (Siddharthan, 2006).
- *Explanation generation*: in this approach, difficult concepts are identified and then augmented with external information before putting it back within the same context. Examples of this technique would be SIMTEXT (Damay et al., 2006) that provides dictionary definitions for medical terms or FACILITA (Watanabe et al., 2009) that uses Wikipedia references for some difficult identified terms.
- *Statistical machine translation*: Borrowing from the technique of automated

machine translation ([Koehn, 2009](#)), this technique situates text simplification as a text-to-text generation. In contrast with rule-based methods mentioned above, it demands less rule handcrafting and the output is more fluent (readable), however, the creation of the parallel corpora is usually costly. Recent advances in machine translation utilising neural sequences ([Bahdanau et al., 2014](#)) have been transferred to text simplification, e.g. in ([Nisioi et al., 2017](#); [Chopra et al., 2016](#); [See et al., 2017](#)), obtaining better performance compared to traditional methods.

A different way to categorise text simplification algorithms would be whether they generate new phrases (abstractive summarisation), e.g. in ([Rush et al., 2015](#); [Liu et al., 2018](#)), or just extract pieces of text from the given input (extractive summarisation), e.g. in TEXTRANK ([Mihalcea, 2005](#)).

**Templates** A different method to produce a report that does not necessarily involve text simplification would be via the use of a template-based Natural Language Generation (NLG) system. These systems map data input directly to a linguistically structured template. Though simpler than text simplification approaches, it still requires syntactic and lexical operations in the realisation phase to produce coherent output ([Reiter and Dale, 1997](#)). Though not considered as pure-NLG systems and criticised as more difficult to maintain and of poorer and less coherent outputs, they are heavily used in industry, e.g. in journalism for automatic news generation ([Caswell and Dörr, 2018](#)) or in business intelligence for automatic creation of financial reports ([Mishra et al., 2019](#)).

## 3.6 Natural Language Processing in the Era of Large Language Models

In recent years<sup>21</sup>, the advent of Large Language Models (LLMs), such as openAI’s GPT-3 (Brown et al., 2020), GPT-4 (OpenAI, 2023), Google’s BARD<sup>22</sup>, Anthropic’s Claude (Azzollini and Pomponio, 2019) and others have revolutionised the field of Natural Language Processing.

Characterized by their ability to generate human-like text, these models can process vast amounts of information and predict subsequent text based on a given input, thereby demonstrating a nuanced understanding of various semantic, syntactic, and discursive elements in language (Bommasani et al., 2022). They have so far demonstrated great potential in various fields of NLP, such as *text generation*, *machine translation*, *sentiment analysis* and *information extraction* (Raffel et al., 2020). In machine translation, for instance, LLMs have facilitated the automated conversion of text from one language to another with remarkable accuracy, addressing the persistent challenge of maintaining contextual semantics, e.g. achieving human parity in Chinese to English translation (Hassan et al., 2018).

### 3.6.1 Evolution of LLMs

Language models are probabilistic models that estimates the likelihood of a sequence of words or tokens in a given language. They learn to predict the probability of a word or token based on the context of the preceding words or tokens in a sequence. Historically, the concept of encoder-decoder models gained popularity with the advent of sequence-to-sequence (seq2seq) (Sutskever et al., 2014) models

<sup>21</sup>We can originate the start of *Large* language models in 2017 to the publication of the “Attention is all you need” paper (describing the transformer architecture) (Vaswani et al., 2017) and the subsequent release of BERT (Devlin et al., 2018) in 2018

<sup>22</sup><https://bard.google.com/>

for tasks like machine translation, often based on RNNs and LSTMs. Attention mechanisms (Bahdanau et al., 2016) were introduced to address the limitations of seq2seq models, helping them focus on relevant parts of the input sequence when generating the output. In 2017, Vaswani et al. (2017) had a profound impact on the field of NLP as it introduced the *transformer* architecture that addressed much of the limitations of recurrent and convolutional neural network architectures. It also introduced the concept of *self-attention*, which allows the model to weigh the importance of words in an input sequence when generating an output sequence and has been widely adopted in subsequent models. Some of the foundation models based on the transformer architecture were the Generative Pre-trained Transformer (GPT) (Radford et al., 2018) introduced by openAI, a decoder-only Transformer model focused on unsupervised pre-training for language modelling tasks, BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2018), introduced by Google in 2018, an encoder-only Transformer model that leverages bidirectional context for various NLP tasks through masked language modelling, and in 2019 T5 (Text-to-Text Transfer Transformer) (Raffel et al., 2020), an encoder-decoder Transformer model that unifies NLP tasks into a single text-to-text format. A breakthrough was attained with the release (through an API) of GPT-3 (Brown et al., 2020) in 2020, a massive LLM with 175 billion parameters, demonstrating impressive few-shot learning capabilities and strong performance on a wide range of NLP tasks. Overall, we can classify LLMs into three main categories: Encoder only (e.g. BERT, RoBERTa, Electra), Decoder only (e.g. GPT-x, BARD, PaLM) and Encoder-Decoder (e.g. T5, Flan-T5). A diagram depicting the current state of LLM families is shown in Figure 3.8<sup>23</sup>, sourced from Yang et al. (2023).

---

<sup>23</sup>In light of the rapid evolution of LLMs, we anticipate the likelihood that this diagram may quickly become outdated

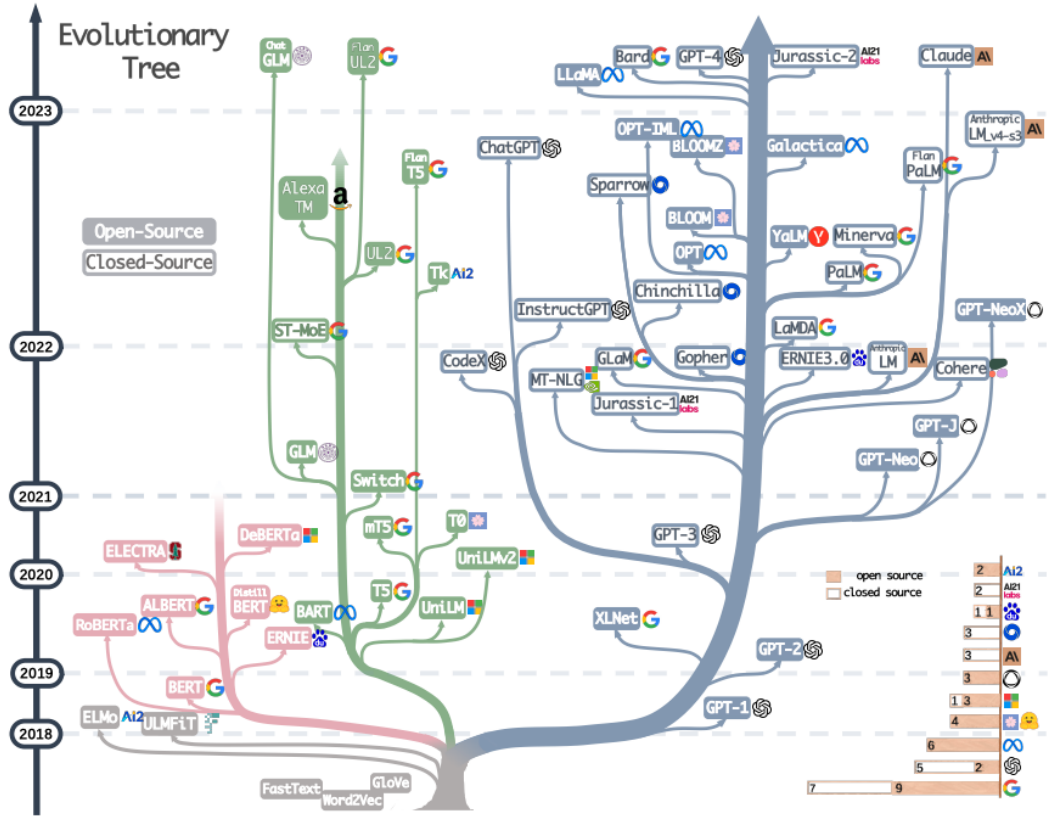


Figure 3.8: Evolutionary tree of modern LLMs, decoder-only models are in the blue branch, encoder-only models in the pink branch, and encoder-decoder models in the green branch. Reprinted from *Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond*, by Jingfeng Yang et al, 2023

### 3.6.2 Text Generation Strategies with LLMs

In its simplest form Language Models can be seen as next word predictors (Ganai and Khurshheed, 2019). However, following a *greedy search* approach (at each generation step, the model selects the token with the highest probability) to generate a sentence, will lead to highly focused and repetitive text, lacking in diversity and creativity (Welleck et al., 2019). To tackle this, practitioners use *Beam search* where the model generates multiple candidate sequences in parallel (called beams) and selects the next token based on the highest combined probability of the sequence until that point. Beam search aims to find the most probable sequence but similarly to greedy search

can suffer from a lack of diversity and may generate repetitive or overly focused text (Hokamp and Liu, 2017). It is important to note that high-quality human language does not follow a distribution of high probability next words (as demonstrated in Holtzman et al. (2019)). In other words, as humans, we want generated text to surprise us and not be predictable. To tackle this issue of predictability, you can introduce *frequency penalties*, i.e. control the repetition of tokens based on their frequency in the generated text e.g. avoid repeating 2-grams or re-adjust probability on each generation step according to the appeared tokens frequency (Klein et al., 2017). By penalizing tokens that have already appeared frequently in the generated text, making them less likely to be selected again, overly repetitive text is prevented but this affects the readability of the generated text.

Another method to decrease repeatability is to perform *sampling*, i.e. to select the next token randomly based on the probability distribution of the vocabulary (Fan et al., 2018). This approach can produce highly diverse and creative text but may sacrifice coherence and focus. Popular sampling techniques are top-k where the top k tokens with the highest probability are considered and top-p (or nucleus sampling) where only a dynamic set of tokens -where their cumulative probability is at least p- is considered. The parameter p controls the trade-off between diversity and focus: higher values of p (e.g., 0.9) lead to more diverse and creative text, while lower values (e.g., 0.5) result in a more focused and coherent text. Another way you can control the creativity of the text is with the temperature parameter, which controls the softmax function behaviour when it normalises the tokens probabilities (Bishop and Nasrabadi, 2006). For example, a high temperature makes the probability distribution of the next token "flatter," giving lower-probability tokens a higher chance of being selected. The generated text may be less focused and coherent, as the model is more likely to explore different paths in the generation process; but it will be more creative.



Another prompting technique that allows LLMs to understand the desired pattern or structure of the output is few-shot learning (also known as contextual learning) (Wei et al., 2022). In few-shot learning, the model is provided with a small number of input-output example pairs as context. We distinguish according to the number of examples given to: (i) 0-shot learning, which is convenient yet ambiguous and challenging as the prompt’s selected words matter a lot, (ii) 1-shot: a single example output removes ambiguity yet is still challenging to get the desired response, (iii) few-shot: it has stronger performance, however, you need to carefully select the examples. Further to prompting strategies to extend the performance of an LLM you can fine-tune it, which will achieve the strongest performance in a specific task but is costly, needs a curated labelled dataset, and there is a high probability to overfit your data. Also, LLMs can hallucinate false information due to their training objective of maximizing the likelihood of the next token given the context. Without grounding in common sense or world knowledge, they may generate plausible-sounding but incorrect or nonsensical text. To counter the phenomenon of hallucinations an approach would be to Fine-tune through embeddings, which simply entails leaving the pre-trained model frozen as a fixed feature extractor, adding an embedding layer on top of the pre-trained model, and then only training the added embedding layer on the downstream data. In this way, you are embedding domain-specific knowledge while retaining the general knowledge in the pretrained model and avoid overfitting to small downstream tasks.

Further to fine-tuning, a recent technique to increase the performance of LLMs is RLHF (Reinforcement Learning from Human Feedback) (Ziegler et al., 2020) where LLMs are fine-tuned using human-generated feedback. In the RLHF paradigm, human annotators are assigned the task of ranking various generated variations, which are subsequently employed in the training of a reward model designed to predict the quality of the output (Bai et al., 2022). Then, this reward model is

used as feedback to fine-tune (using Proximal Policy Optimization ([Schulman et al., 2017](#))) the original model. A variant of the GPT-3 model that was trained using the RLHF methodology played a significant role in the development of ChatGPT<sup>24</sup>, a chatbot application that has popularised the area.

LLMs can be used for Argument Mining by fine-tuning them on a dataset of argumentative text. This allows the LLM to learn the patterns and structures of argumentative text and can then be used to identify and extract arguments from new text. This approach has been used for example in [Behrendt and Harmeling \(2021\)](#) to calculate embeddings to measure the similarity of arguments. We use a similar approach in Chapter 6 to fine-tune BERT for scientific argument mining.

## 3.7 Automatic Summarisation Systems

An automatic summarisation system can be seen as the process of condensing a (potentially long) document into a short paragraph that conveys the core information ([Nazari and Mahdavi, 2019](#)). It differs from a compression system, as it attempts to reduce the dimension of data but without excessive loss of information ([Hahn and Mani, 2000](#)). As such, it needs to accurately identify the important bits of information and take care to include them in the final generated output while maintaining fluency and coherence ([Batista et al., 2015](#)).

The two main strategies for text summarisation are *extractive summarisation* where a subset of sentences from the input document is identified and directly copied into the summary (e.g. ([Xu et al., 2019](#); [Zhong et al., 2019](#))) and *abstractive*, where the important subset of sentences are detected and paraphrased to generate the output summary (e.g. ([See et al., 2017](#))). Abstractive summarisation as a task, combines -in a challenging fashion- understanding the meaning of long documents (a Natural

---

<sup>24</sup><https://chat.openai.com/>

Language Understanding - NLU (Allen, 1995) subtopic) and subsequently generating a readable natural language output of this (a Natural Language Generation - NLG (Gatt and Krahmer, 2018) subtopic).

The latest state of art approaches for automatic summarisation comprise of Sequence-to-Sequence learning (Sutskever et al., 2014) and specifically transformers. Transformer architecture, introduced in (Vaswani et al., 2017) has proven to be outstandingly effective in Natural Language Processing tasks (Tay et al., 2022). It entails an attention mechanism that uses the context of a word to weigh the importance of each token and can handle long-term dependencies (relationship of distant words or sentences), which is particularly important for summarisation (Tay et al., 2022). It is therefore logical, given modern LLMs large scale and generative capabilities, that they hold significant potential for the task of summarisation. For example, BART (Lewis et al., 2019) and PEGASUS (Zhang et al., 2020) achieve state-of-the-art performance in known datasets such as CNN-DailyMail (Hermann et al., 2015). As this time writing this thesis, the leading performance in summarisation task over CNN-DM<sup>25</sup> has been demonstrated by variants of PEGASUS fine-tuned for this dataset (Zhao et al., 2022). Notably, a version of fine-tuned BART occupies the second position (Liu and Liu, 2021).

### 3.7.1 Summary as a Text-to-text Task

Automatic summarisation is the natural language processing (NLP) task that involves generating a shorter version of a longer text while retaining its essential meaning. This task is considered a text-to-text task because it involves generating a summary text that is based on the input text (Gambhir and Gupta, 2017).

Prompting is a technique used in automatic summarisation that involves providing a specific prompt or topic sentence to guide the summarisation process (Widyassari

---

<sup>25</sup><https://paperswithcode.com/sota/abstractive-text-summarization-on-cnn-daily>

[et al., 2022](#)). This technique differs from traditional summarisation methods, which generate summaries without any prior knowledge of the intended topic or purpose of the summary. The use of prompts can change the flavour of summarisation by providing a structured framework that can help ensure that the summary captures the key elements of the input text. Prompts can also help reduce bias in summarisation by guiding the summariser towards a specific goal or objective. Additionally, prompts can improve the coherence and fluency of the summary by providing a clear starting point for the summarisation process.

In Chapter 5, we outline a summarising method for long discussions based on LLM prompting.

### 3.7.2 Summarisation Evaluation

To indicate some notion of “text quality” in their evaluation, automatic summarisation systems employ some of the following human evaluation criteria ([van der Lee et al., 2021](#)):

- *Fluency*: The summary should be grammatically correct and easy to understand.
- *Readability*: The summary should be written at an appropriate level of difficulty for the intended audience.
- *Comprehensiveness*: The summary should cover the key points and main ideas of the original text.
- *Coherence*: The summary should be logically organized and should flow smoothly from one idea to the next.
- *Relevance*: The summary should only include information that is relevant to the main topic of the original text.

- *Concision*: The summary should be as concise as possible without losing important information.
- *Accuracy*: The summary should accurately represent the content of the original text.
- *Factuality*: refers to whether the summary accurately reflects the facts and information presented in the original text. A good summary should not include any false or misleading information and should be based on the content of the original text.
- *Adequacy* refers to whether the summary is sufficient to convey the main ideas and key points of the original text. In other words, the summary should provide enough information for the reader to understand the main points of the original text, without leaving out important details.

To be useful in practice, however, automated metrics should agree well with human judgment. The most established and commonly used automatic metrics are:

- BLEU ([Papineni et al., 2002](#)) (bilingual evaluation understudy) is primarily used in the machine translation domain and measures the correspondence (how similar) of a machine-generated summary to a reference human summary.
- ROUGE ([Lin, 2004](#)) (Recall-Oriented Understudy for Gisting Evaluation) examines the overlap of unigram (ROUGE-1), bigrams (ROUGE-2) and the longest common subsequence (ROUGE-L) of the machine summary to the reference human,
- METEOR ([Banerjee and Lavie, 2005](#)) (Metric for Evaluation of Translation with Explicit ORdering) focuses on the sentence level rather than corpus level, by creating alignments between the test and reference sentences and generally achieves better correlation to human judgements.

- BERTScore (Zhang et al., 2019) leverages the contextualized embeddings generated by BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2018), to compare the similarity between the generated text and the reference text. It is considered more robust against issues like paraphrasing and word reordering, as it captures semantic similarity rather than just lexical matching.

The above metrics mostly rely on n-gram overlap which promotes repeating identical text from the reference source and penalises novel expressive sentences that intend to communicate summary better; focusing more on the output text matching to references, which has a negative bias to human demand (Yang et al., 2018). Furthermore, they have under-informative and reliability issues (caused by different software implementations). Foremost, automatic metrics is well known that do not necessarily agree with human evaluations (Novikova et al., 2017), therefore, the recommendation is to use automatic metrics in the early stages of summary system development as a cheap repeatable measure, before proceeding to the costly and time-consuming human evaluation of the end product (van der Lee et al., 2021).

**Argumentation mining as a summarisation approach** In extractive summarisation (but also for intermediate steps of abstractive summarisation) a key point for the performance of the system is the identification of the most important segments of text (to be later used in the output generation). We propose (for crafting the summariser component depicted by RQ3 1.3.2) to follow a similar approach with the outputs of argumentation mining, i.e. to identify the most significant pieces of text using the most central arguments detected (via network analysis of the argumentation graph). We also propose the use of summarisation via LLM (prompting) as an efficient method to transform long research abstracts into condensed text snippets that retain the main argument of papers in Section 6.2.3; this serves as a

novel recommendation unit in our scientific argument recommender- SciArgRecSys in Chapter 6.

## Chapter 4

# Automated Methods of Reporting for Online Deliberation

The aim of argument, or of  
discussion, should not be victory,  
but progress.

---

Pensées de J. Joubert (1848)

In this chapter, we investigate automated reporting methods as a means to improve sensemaking - the capability of people to make sense of what they read in an online discussion platform - and improve the overall quality of the deliberation process. We proceed to make a comparison of three approaches to assist newcomers or latecomers in a discussion: (i) abstractive summarisation, (ii) template reporting and (iii) argumentation highlighting system. We evaluate these against factors for sensemaking and quality of the deliberation process.

**RQ2:** What automated reporting approaches are more appropriate for online deliberation?

**User study II:** Examining automated report methods effect in Sensemaking and the Quality of the Deliberation Process



## 4.1 User Study II - Argumentation-based Automated Reports

This chapter presents an investigation of computational aids to improve key stages of Pirolli and Card’s sensemaking model (Pirolli and Card, 2005): communicating results and story-telling. We start from the premise that automated reporting can tell a ”story” of the state and progress of an online debate, which can improve participants’ understanding and positively affect the overall quality of the debate by helping them to better inform their contribution. As people struggle to comprehend large or dynamic debates, computational aids such as an automatic summariser or a text-highlighter of argumentative structures, have the potential to inspire for improving the current status of online debate technologies by enabling participants to quickly grasp the overall state of a debate before or while they are contributing. This may lead to better quality debates, for instance, by enabling people that are new to the discussion to inform where they could better contribute. Automated reporting can also be particularly useful in highly dynamic debates, which tend to evolve fast, for returning participants to quickly grasp the progress made since they last contributed, thus reducing idea repetition.

This study (*Study II*) looks specifically at the capability of automated reports to improve newcomers’ sensemaking and the perceived quality of deliberation. We compared three different automated reporting mechanisms: Abstractive summarisation, Templated summary and Argumentation highlight - excerpts of each are shown in Figure 4.1. We aimed at understanding which one of the three approaches is more promising in improving new participants’ capability to make sense of an online discussion. We also aimed to investigate the extent to which providing a summary can improve newcomers’ perception of the overall deliberation’s quality. To this end, we carried out a cross-sectional between-subjects quantitative study where we

compared 4 different groups: (3 treatment groups), each with a different automated reporting method and one control group (deliberation with no automated report). Results were evaluated against a theory-based set of Sensemaking (*SM*) and Quality of Deliberation content metrics (*QoD*) features (Figure 4.2); as explained in Section 3.2.2. The above quantitative study was followed by a focus group to discuss with users each automated reporting method and to qualitatively assess their applicability in the context of a deliberation platform. In the following, we motivate and describe the research and draw preliminary conclusions which can inform the design of future automated reporting systems for online discussion.

## 4.2 Study Research Question and Experimentation Design

The assumption underpinning this investigational study is that an automated text-based report presented to a new user in advance or in parallel to his interaction with a discussion improves his understanding of the overall state of the debate and enables healthier participation in it. This assumption drives the research question RQ2 as described in Section 1.3.2:

**RQ2:** What automated reporting approaches are more appropriate for online deliberation?

We use a mixed methods approach with explanatory sequential design (two-phase design (Creswell and Creswell, 2017)) consisting of a quantitative study (Section 4.3) to identify results for a follow-up investigation, in phase two, which conducts qualitative data analysis (Section 4.4).

**(i) Abstractive summary**

- Students do not need government's funding ( aka loans and grants ) because they can learn for free ( with open courseware ) and get college credits for low costs through massively open online courses ( mooc's ) .
- alternatives may not be as easily accessible as public universities
- in most countries there is no right to study other than the right not to be hindered by the state
- automation and robotics will reduce the need for uneducated work force . more and more workers need a basic education and skills they get from further education
- children from a higher class background doesn't have more opportunity through the education system . inequalities would have been entrenched in primary and secondary schooling by then
- general public education improves the employ-ability of all youth beyond basic knowledge

**(ii) Templated summary**

This debate has drawn 646 contributions from 110 participants. There have been many claims centring mainly around 15 ideas. Some of the strongest ideas in favour of this proposition are that everyone has the right to study and educated, while those opposing the idea claim that the cost of attending makes students more accountable for their work. Also, the need for a college degree has increased, however, students do not need government's funding ( aka loans and grants ) because they can learn for free ( with open courseware ) and get college credits for low costs through massively open online courses ( mooc's ) . alternatives may not be as easily accessible as public universities

**(iii) Excerpt of Argument (claims or evidence) highlight report****claim - evidence**

Students do not need the government 's funding (aka loans and grants) because they can learn for free (with open courseware) and get college credits for low costs through massively open online courses (MOOC 's).

General public education improves the employ-ability of all youth beyond basic knowledge. In 1988, nominal median household income was \$ 27,225 with average college tuition for a public four-year university \$ 3190 per year . ... More and more workers need a basic education and skills they get from further education . Considering the tuition compared to income, it has only had a 38.4 percent real increase in 30 years when compared to median household income, resulting in a 1 percent real increase year-over-year when compared to median household

Figure 4.1: Examples of automatic generated reports on the debate topic: "Higher education should be publicly funded." using (i) abstractive summary, (ii) templated summary and (iii) argument highlight

### 4.2.1 Experiment Design

We concentrated our efforts on entirely automated reporting methodologies. For that, we chose for evaluation three methods of automated reporting that are operable on textual data and have the potential to unveil rhetorical structures within text, which makes them more suitable for the analysis of discussion data:

- *Abstractive summarisation*: In contrast to the simpler *extractive summarisation*, it does not just extract significant snippets from the original, but rather generates an entirely new text (See et al., 2017). It relies on advanced NLP techniques to identify lexical and discourse elements of the input text producing an intermediate semantic representation of it, before proceeding to synthesise important bits of information to a newly created document. Recent advancements with the use of encoding transformers have produced models that achieve close to human performances (Dou et al., 2020; Yan et al., 2020; Zhang et al., 2020). The abstractive summary that we used was generated with the model described in See et al. (2017) and an excerpt can be seen in Figure 4.1(i).
- *Templated NLG summary*: As the simplest form of Natural Language Generation, templated summaries are systems that are based on static or dynamic templates, a blend of human-crafted rules to “fill-in-the-blank” in a text template. An advantage of templated NLGs would be that they can easily combine heterogeneous data sources (and not just rely on the pure textual input), for example in our case we included some simple statistical indicators (number of participants in the overall debate, number of claims, number of total contributions). The template used in our study was realised using SimpleNLG<sup>1</sup> engine and can be seen in Figure 4.1(ii).

---

<sup>1</sup><https://github.com/simplenlg/simplenlg>

- *Argumentation highlight*: is a different kind of reporting, aimed at highlighting and revealing argumentation structures in the given text. It relies on Argumentation mining (see 3.4), an NLP method capable of analyzing people's argumentation, i.e. extracting arguments and their relations from text (Cabrio and Villata, 2018). An argument mining processed version of the debate (where arguments, claims and evidence are highlighted) can be seen in Figure 4.1(iii). The claim and evidence extraction used in this study was done using the MARGOT system (Lippi and Torroni, 2016b).

### 4.3 Quantitative Comparative Study

The objective of the quantitative study was to systematically compare the three automated reports described above in terms Sensemaking and perceived quality of deliberation improvements. As mentioned in Section 3.2.1, quantifying *Sensemaking* to measure it as an experimental variable is a challenging task. For our study we used Alsufiani's (Alsufiani et al., 2017) theoretical factors of SM, with an extra feature to assess improvements in Reflection, as proposed by De Liddo et al. (2020). This resulted in the 9 Sensemaking factors presented in Figure 4.2. Similarly, for measuring *Quality of Deliberation* we used a merged set of 11 factors proposed in Graham and Witschge (2003) and Kay (2006) to construct a questionnaire to quantitatively assess users' perceived quality of deliberation (also presented in Figure 4.2).

In order to address the research question we formed the following testable hypotheses:  $H_1$ : Any form of report improves internal sensemaking of participant and perception of the deliberation quality compared to no summary present.  $H_2$ : A templated summary report improves SM and perception of QoD compared to an abstractive summary report.  $H_3$ : An argument highlight report improves SM and perception

Code	Literature sensemaking factor	Adapted definition
SM1	<a href="#">Weick (1995)</a> Retrospect	Reflection
SM2	<a href="#">Alsufiani et al. (2017)</a> Gaining insight	Insights
SM3	<a href="#">Alsufiani et al. (2017)</a> Finding connections	Focus
SM4	<a href="#">Alsufiani et al. (2017)</a> Structuring	Argumentation
SM5	<a href="#">Alsufiani et al. (2017)</a> Reducing confusion,	Explanation
SM6	uncertainty, and ambiguity	Assess Facts and Evidence
SM7		Distinguishing
SM8	<a href="#">Alsufiani et al. (2017)</a> Gap-finding and	Assess assumptions
SM9	gap-bridging	Change assumptions

Code	Literature quality of debate factor	Adapted definition
QoD1	<a href="#">Kay (2006)</a> Message clarity	Message clarity
QoD2	<a href="#">Kay (2006)</a> Message quality	Message quality
QoD3	<a href="#">Kay (2006)</a> Presence of new knowledge	New knowledge
QoD4	<a href="#">Kay (2006)</a> External resources used	External resources
QoD5	<a href="#">Kay (2006)</a> Resolution of discussion	Discussion resolution
QoD6	<a href="#">Stromer-Galley (2007)</a> Reasoned opinion expression	Reasoned opinions
QoD7	<a href="#">Stromer-Galley (2007)</a> Sourcing	Trustworthy evidence
QoD8	<a href="#">Stromer-Galley (2007)</a> Distinct views - disagreement	Distinct opinions
QoD9	<a href="#">Stromer-Galley (2007)</a> Equality	Equality
QoD10	<a href="#">Stromer-Galley (2007)</a> On topic	On topic
QoD11	<a href="#">Stromer-Galley (2007)</a> Engagement	Engagement

Figure 4.2: Sensemaking (SM) and Quality of Debate (QoD) evaluation factors extracted from literature (as detailed in Section 3.2.2)

of QoD compared to a templated summary or abstractive summary reports. Our experimental conditions consist of the presence (or not) of an automated report, with the following four conditions:  $A$ =control/no report,  $B$ =abstractive summary report,  $C$ =templated report and  $D$ =argumentation highlight.

### Quantitative study - Pre exploration

Our quantitative study consisted of a comparative study of 1 independent variable with 4 different conditions, each given a different report interface:

- Condition A (CA): Abstractive summary of the main arguments discussed around an issue, is provided before and during engagement with the actual debate. The abstraction summary was generated with the model described in [See et al. \(2017\)](#) and an excerpt can be seen in Figure 4.1.
- Condition B (CB): A template-driven NLG summary of a discussion of an issue, is provided before and during engagement with the actual debate. The template was realised using SimpleNLG<sup>2</sup> engine.
- Condition C (CC): An argument-mining processed version of the debate (where arguments, claims and evidence are highlighted) is provided during engagement with the actual debate. The claim and evidence extraction was done using MARGOT system ([Lippi and Torroni, 2016b](#)).
- Condition D (CD): No summary of any form was given to the user prior to or during the interaction with the actual debate.

#### 4.3.1 Task

Users were presented with a simple user interface where an issue was debated with pro and con arguments and an automated report was presented along according to the

<sup>2</sup><https://github.com/simplenlg/simplenlg>

given condition forming four different groups correspondingly. The issue discussed was “*Should higher education be publicly funded?*” and data were scraped from the [kialo.com](https://www.kialo.com/should-higher-education-be-publicly-funded-7565?path=7565.0)<sup>3</sup> platform. Only the top 10 pro and top 10 con arguments were presented. The ranking was assigned according to the given order by Kialo platform which is deduced as a mixture of voting and user rating. The used dataset was extracted from Kialo as a platform that fosters reasoned discussion (focuses by design on arguments) as the interface implemented also adheres to discussion in a similar argumentative fashion. All interfaces follow simplistic claims with pro and con arguments structure which fits the structure of the underlying data.

The interfaces of the four conditions (CA,CB,CC,CD) can be seen in Figure 4.3.

Users were then asked to devote at least five minutes to study, read and comprehend the issue discussed and the arguments of both sides. As a completion criterion users were asked to (i) Identify the most significant (strongest) argument in their opinion, (ii) Contribute at least one argument for, or against, and (iii) Summarise their understanding of the debate in a short text of 50-100 words. We assigned each group 40 participants (160 total) and used Mechanical Turk to recruit users. A reward of equivalent 12\$ per hour was given for each completed task and users had 10 min to complete the task. For each task to be considered accepted, (i) a summary of more than 50 words, (ii) authentic style (not just copy-paste a long piece of text from the platform) and (iii) justification of answers consisting of more than five words was required.

Failed HIT tasks were resubmitted to new users until a valid response was received. In total, we had to assign 273 tasks before getting 160 valid completed tasks (an excess of 113 failed tasks or 58/42% success/fail rate)

---

<sup>3</sup>The full topic can be found at <https://www.kialo.com/should-higher-education-be-publicly-funded-7565?path=7565.0> 7565.1



MTurk Survey

Higher education should be publicly funded.

Debate summary - Use this summary report as a reference while exploring the rest of the debate, hover over highlighted phrases to see arguments in context

- students do not need government's funding ( aka loans and grants ) because they can learn for free ( with open courseware ) and get college credits for low costs through massively open online courses ( moocs ) , alternatives may not be as easily accessible as public universities
- in most countries there is no right to study other than the right not to be hindered by the state . in most countries there is no right to study other than the right not to be hindered by the state , automation and robotics will reduce the need for uneducated work force , more and more workers need a basic education and skills they get from further education , more ,
- public universities are far less likely to adapt to that change since there is no administration for whom it is in their self-interest to do so , publicly funded programs do work in certain areas , scientific endeavors have been successful in the past

Public higher education would not be subject to the pressure necessary to perform and improve itself, as the outcome influences the survival / advancement of a given program / its managers less directly, as is the case for private enterprise in the marketplace.

Public higher education would not be subject to the pressure necessary to perform and improve itself, as the outcome influences the survival / advancement of a given program / its managers less directly, as is the case for private enterprise in the marketplace.

(a) Discussion with abstractive summary - condition CA

MTurk Survey

Higher education should be publicly funded.

Debate summary - Use this summary report as a reference while exploring the rest of the debate, hover over highlighted phrases to see arguments in context

This debate has drawn 646 contributions from 110 participants. There have been many claims centering mainly around 13 ideas. Some of the strongest ideas in favor of this proposition are that **everyone has the right to study and educated** , while those opposing the idea claim that **the cost of attending makes students more accountable to their work** . Also, **the need for a college degree has increased** , however, **students do not need government's funding ( aka loans and grants ) because they can learn for free ( with open courseware ) and get college credits for low costs through massively open online courses ( moocs )** , alternatives may not be as easily accessible as public universities

Students would be able to find their passions as opposed to studying what will give them the fattest paycheck.

Public higher education would not be subject to the pressure necessary to perform and improve itself, as the outcome influences the survival / advancement of a given program / its managers less directly, as is the case for private enterprise in the marketplace.

Public higher education would not be subject to the pressure necessary to perform and improve itself, as the outcome influences the survival / advancement of a given program / its managers less directly, as is the case for private enterprise in the marketplace.

(b) Discussion with templated summary - condition CB

MTurk Survey

Higher education should be publicly funded.

Debate summary - Use this summary report as a reference while exploring the rest of the debate, hover over highlighted phrases to see arguments in context

35 claims have been found, 30 evidence have been found. **Claims** are displayed in **blue** , **evidence** is displayed in **red** . Words in **bold/italic** belong both to a claim and to a piece of evidence.

The military will pay for one's schooling if one signs up with a branch of service.

The internet and personal research are good examples.

**Universities do not fund the government's & business** ( aka loans and grants ) because **they can learn for free** ( with open courseware ) and get

Students would be able to find their passions as opposed to studying what will give them the fattest paycheck.

Public higher education would not be subject to the pressure necessary to perform and improve itself, as the outcome influences the survival / advancement of a given program / its managers less directly, as is the case for private enterprise in the marketplace.

Public higher education would not be subject to the pressure necessary to perform and improve itself, as the outcome influences the survival / advancement of a given program / its managers less directly, as is the case for private enterprise in the marketplace.

(c) Discussion with argument highlights - condition CC

MTurk Survey

Higher education should be publicly funded.

Students would be able to find their passions as opposed to studying what will give them the fattest paycheck.

Public higher education would not be subject to the pressure necessary to perform and improve itself, as the outcome influences the survival / advancement of a given program / its managers less directly, as is the case for private enterprise in the marketplace.

Public higher education would not be subject to the pressure necessary to perform and improve itself, as the outcome influences the survival / advancement of a given program / its managers less directly, as is the case for private enterprise in the marketplace.

(d) Discussion with no summary present - condition CD

Figure 4.3: Discussion interface under conditions CA,CB,CC,CD

### 4.3.2 Analysis

We used discrete variables to measure participants' SM and perceived QoD (Likert scale survey answers). We mapped the Likert scale responses to a linear scale paying attention to reversing the negated questions. An initial Shapiro-Wilk Test showed that the values do not conform to a normal distribution and the variance is not homogeneous, therefore, we used a non-parametric test to compare the conditions.

We chose a Kruskal–Wallis test (Vargha and Delaney, 1998), also known as one-way ANOVA on ranks, followed by a post hoc examination using Dunn's test (Dinno, 2015) to reveal which group was stochastically dominating.

	SM1 Reflection	SM2 Insights	SM3 Focus	SM4 Argumentation	SM5 Explanation	SM6 Assess facts	SM7 Distinguishing	SM8 Assess assum.	SM9 Change assum.
A (control)	3.925 (0.858)	3.675 (1.022)	3.25 (1.334)	3.5 (1.062)	3.25 (1.295)	2.75 (1.295)	3.85 (0.863)	3.3 (1.159)	2.95 (1.131)
B (abs. sum.)	4.325 (0.944)	3.45 (1.108)	3.6 (1.354)	3.925 (1.022)	3.475 (1.377)	2.55 (1.376)	4.05 (1.011)	3.55 (1.338)	3.2 (1.181)
C (temp. rep.)	4.05 (0.714)	3.5 (0.784)	3.1 (1.194)	3.5 (1.012)	3.125 (1.158)	2.875 (1.158)	3.75 (0.954)	3.35 (1.098)	3.125 (1.244)
D (arg. high.)	4.35 (0.921)	4.05 (1.011)	3.875 (1.244)	4.15 (1.026)	3.575 (1.393)	2.4 (1.373)	4.175 (0.812)	3.725 (1.395)	3.25 (1.255)
H-stat(3)	<b>12.036</b>	<b>11.271</b>	<b>9.960</b>	<b>12.800</b>	3.492	3.635	6.014	4.659	1.785
p-value	<b>0.0072</b>	<b>0.0103</b>	<b>0.0189</b>	<b>0.005</b>	0.3217	0.3036	0.1108	0.1985	0.618

Table 4.1: Kruskal-Wallis statistical test of Sensemaking features

	QoD1 Msg Clarity	QoD2 Msg quality	QoD3 New knowledge	QoD4 Ext. resources	QoD5 Resolution	QoD6 Reasoned opinions	QoD7 Trustworthy	QoD8 Distinct opinions	QoD9 Equality	QoD10 On Topic	QoD11 Engagement
A (control)	3.525 (0.997)	2.625 (1.102)	3.5 (1.012)	3.2 (1.114)	2.45 (1.06)	3.35 (0.892)	3.325 (0.729)	3.225 (0.973)	3.25 (1.006)	3.725 (0.784)	3.675 (0.997)
B (abs. sum.)	4 (0.905)	2.425 (1.318)	3.575 (1.129)	3.35 (1.026)	2.575 (1.152)	3.65 (0.975)	3.8 (0.966)	3.125 (1.158)	3.325 (1.227)	3.975 (0.861)	3.8 (0.992)
C (temp. rep.)	3.675 (0.858)	2.75 (1.235)	3.525 (0.846)	3.275 (0.986)	2.65 (1.051)	3.7 (0.822)	3.45 (0.875)	3.125 (0.911)	3.2 (1.042)	3.8 (0.992)	3.725 (0.876)
D (arg. high.)	4.15 (0.975)	2.125 (1.09)	3.7 (1.181)	3.125 (1.284)	2.625 (1.371)	3.85 (0.921)	3.775 (0.973)	2.975 (1.229)	3.125 (1.399)	3.85 (1.051)	3.65 (1.188)
H-stat(3)	<b>12.426</b>	6.818	1.535	0.508	0.679	6.187	<b>8.384</b>	1.002	0.493	2.34	0.442
p-value	<b>0.006</b>	0.0779	0.6741	0.917	0.8779	0.1028	<b>0.0386</b>	0.8007	0.9202	0.5047	0.9314

Table 4.2: Kruskal-Wallis statistical test of Quality of Deliberation features

We present the results of Kruskal-Wallis H statistical test ( $df=3$ ,  $\alpha=0.05$ ) of quality and SM features in table 4.1 and 4.2 containing a complete overview of means and standard deviations by condition, the Kruskal-Wallis test statistic (H-stat) and corresponding p-value of each feature. Our analysis of SM features showed a significant difference ( $p < 0.05 = \alpha$ ) between the four conditions for Reflection (SM1), Insights (SM2), Focus (SM3) and Argumentation (SM4) while no significant differences emerged for the rest of SM features (SM5-SM9). For the QoD features, a significant difference was observed for Message Clarity (QoD1) and Trustworthy evidence (QoD7) feature. The subsequent pairwise comparisons in the statistically significant features are shown in Table 4.3, where we used Bonferroni-adjusted significance level of  $\alpha^* = 0.05/6 = 0.00833$ .

We can confirm hypothesis  $H_1$  for the argumentation highlight report (group D), which outperforms the control group for better Message Clarity (QoD1), Reflection(SM1),

$H_i$	QoD1 Message clarity	QoD7 Trustworthy evidence	SM1 Reflection	SM2 Insights	SM3 Focus	SM4 Argumentation
$H_1$ A < B	$z=2.039$ , $p=0.041$	$z=2.542$ , $p=0.011$	$z=2.593$ , $p=0.009$	$z=0.879$ , $p=0.379$	$z=1.287$ , $p=0.198$	$z=1.819$ , $p=0.068$
$H_1$ A < C	$z=0.32$ , $p=0.748$	$z=0.878$ , $p=0.379$	$z=0.404$ , $p=0.685$	$z=1.246$ , $p=0.212$	$z=0.629$ , $p=0.529$	$z=0.076$ , $p=0.938$
$H_1$ A < D	<b><math>z=3.039</math>, <math>p=0.002</math></b>	$z=2.209$ , $p=0.027$	<b><math>z=2.682</math>, <math>p=0.007</math></b>	$z=1.825$ , $p=0.067$	$z=2.243$ , $p=0.024$	<b><math>z=2.917</math>, <math>p=0.003</math></b>
$H_2$ B < C	$z=1.719$ , $p=0.085$	$z=1.664$ , $p=0.096$	$z=2.188$ , $p=0.028$	$z=0.367$ , $p=0.713$	$z=1.916$ , $p=0.055$	$z=1.896$ , $p=0.057$
$H_3$ B < D	$z=0.999$ , $p=0.317$	$z=0.333$ , $p=0.738$	$z=0.088$ , $p=0.929$	<b><math>z=2.704</math>, <math>p=0.006</math></b>	$z=0.955$ , $p=0.339$	$z=1.098$ , $p=0.272$
$H_3$ C < D	<b><math>z=2.719</math>, <math>p=0.006</math></b>	$z=1.331$ , $p=0.183$	$z=2.277$ , $p=0.022$	<b><math>z=3.071</math>, <math>p=0.002</math></b>	<b><math>z=2.872</math>, <math>p=0.004</math></b>	<b><math>z=2.994</math>, <math>p=0.002</math></b>

Table 4.3: Dunn's test post-hoc examination - pairwise comparisons

and Argumentation (SM4) features. This confirms our assumption that argumentative technologies perform better regarding key SM features. We cannot draw a conclusion however on the effects of the remaining SM factors such as reducing confusion, uncertainty and ambiguity.

Hypothesis  $H_2$  cannot be confirmed i.e. we cannot draw a conclusion on which type of summary-based report (abstractive or templated) performs better in any of the features examined. Finally, we can confirm  $H_3$ , that Argumentation Highlight performs better than abstractive summary, by providing more Insights (SM2), and performs better than templated report in Message Clarity (QoD1), Insights (SM2), Focus (SM3) and Argumentation (SM4).

Argument mining techniques are explicitly designed with the goal to improve discourse analysis, but their impact in realistic application contexts is yet to be demonstrated. Our preliminary findings confirm that argument mining results, when presented along an online discussion, even in the simplest form (as basic highlights) perform better than other automated reporting approaches, and can help users to better make sense of the logical structure of a debate, gain insights and perceive a clearer message.

## 4.4 Qualitative study

We conducted this qualitative study to gather more insights on the value and applicability of automated reporting in online deliberation systems and investigate further the qualitative differences between the methods compared in the quantitative study. For that a small focus group was carried out which focal point was around the Sensemaking and Quality of the Deliberation process features that did not reveal any statistical significance in the quantitative study. We conducted the focus group at the university's premises in a meeting room equipped with external monitors

to display the different reporting interfaces alongside the online discussion. The focus group lasted one hour and gathered views from three participants with diverse backgrounds (Computer Science, Education and Biology). We elected to have diverse backgrounds as an attempt to limit biases stemming from educational background, as the debate topic used as a case study (“Higher education should be publicly funded”) is considered highly controversial and polarising. We asked exploratory questions regarding the qualitative characteristics of each of the automated reports (B, C, D) and asked participants to discuss, compare, identify shortcomings and suggest improvements.

We then transcribed the full length of the focus group and proceeded to analyze the text using the controlled vocabulary of the SM and QoD features described in the related work section (deductive coding). On a second round of coding, we identified six main themes, which are shown in Table 4.4.

We can deduce the following preliminary criteria for the design of automated reports of online discussion:

- T1: *Accuracy*: Users complained of low accuracy in the system C (argumentative highlight):

“I m just in the 3rd one, it says that the blue is evidence sometimes is evidence but sometimes is quite clearly just an opinion”

that introduces confusion and damages the trust-building of the system. Participants aspired really highly that such systems should provide bullet-proof accuracy. For example, in another instance a participant questions the argumentative structure used and the claim/evidence definitions:

“ People should pay for other people’s education because in a democracy everyone has an interest in others making the best decisions.” That’s not evidence. Or that’s not what I would call evidence ”

Themes	Design criterium	Excerpts
Confusion	Accurate	“..it says that the blue is evidence sometimes is evidence but sometimes is quite clearly just an opinion”
Trustworthiness	Credible	“...even if Piers Morgan tells you the sun comes from the East and sets in the West you would be ”Emmm””
Exploration	Intuitive navigation	“...for me is a bit chaotic because has everything you need so is not a sort of introduction but is interesting to see the things highlighted”
Practicality	Informative	“It is a bit terrifying when I see such a big text, it is overwhelming. So having like four lines and understanding those instead- or having seven bullet points is more helpful I think.”
Focus and in-sight	Evidenced	“I like the idea of highlighting the evidence [...], actually presenting in here actual undeniable evidence is part of the debate, I think it will be very very strong.”
Reflect and evaluate	Structured	“...you need to show the limitations of the debate,for instance the polarisation, political bias, which one of the three it better gives you an overview of how the quality of the debate is, if is equal, if is diverse, if is not dominated by one opinion...”

Table 4.4: Focus group analysis themes, corresponding design criteria and quote excerpts

- T2: *Credibility*: Our participants highlighted the credibility of information as the main driver, to improve that suggested supporting claims by linking them to evidence, otherwise such unsubstantiated claims should be ignored - even suggest they online debates should follow the same rigour like scientific text: “people link to evidence in these debates or do they cite or do they? They should I am making this claim and therefore ... because according Tim Berners Lee 2008 puff”
- T3: *Focus and insight*: The participants criticised the ability to gain insights and whether this process was damaged by confusing UI elements: “I liked the bulleted points but I found it a bit confusing going back a lot times”. However, they acclaimed the ability to enhance focus when (any) form of automated reporting was in place: “I like the idea of highlighting the evidence [...], actually presenting in here actual undeniable evidence is part of the debate, I think it will be very very strong.”. Comparing the three forms of automated reporting presented to them: “The first one (templated) did not give enough, the third (argumentation highlight) gave far too much. It is like is prrrrrr, it does not appear to be a summary it just appears to be everything. Where is the 2nd (abstraction summarisation) is actually to using , definitely does that (make sense)”
- T4: *Navigation*: participants expressed their appreciation for a navigational system, however, recognised that it introduces extra complexity to the discourse. “Third summary for me is a bit chaotic because has everything you need so is not a sort of introduction but is interesting to see the things highlighted”
- T5: *Reflect and evaluate*: “ I think we assert few times which one helped you to understand the debate more, it was the 2nd one. But to assess the overall quality of the debate, you need to show the limitations of the debate, for

instance the polarisation, political bias, which one of the three it better gives you an overview of how the quality of the debate is, if is equal, if is diverse, if is not dominated by one opinion...”

- T6: *Trustworthiness*: Regarding trustworthiness of evidence users found that the provenance of the information shown is important, for example giving an counter-example of a non-trustworthy source would make evidence equally non-trustworthy: “Even if Piers Morgan tells you the sun comes from the East and sets in the West you would be ”Emmm””
- T7: *Usefulness*: The main function of such reports is to provide a quick overview of the overall topic debated: “It is a bit terrifying when I see such a big text, it is overwhelming. So having like four lines and understanding those instead- or having seven bullet points is more helpful I think”

In addition to the aforementioned points, there are several other observations that warrant discussion:

- Participants were not much aware of deliberation-specific platforms. However, they tend to use non-purposed platforms, e.g. Reddit, Twitter for carrying discussions and partly for deliberative purposes.
- Anonymity is considered less important, while credibility and trustworthiness of participants are considered very important factors for the perceived quality of an online debate
- Social media (e.g. Twitter) are not suitable for deliberation as they are very polarised, inducing echo chambers
- Templated summary approach provision of a quick (“in a glance”) overview of the overall debate in a compact piece of text is highly appreciated by the participants.

- At the same time, it is acknowledged that this type of summary fails to cover the full set of arguments present in a debate (the “long tail”) but rather concentrates in two pro and con arguments.
- Highlighting arguments is considered rather “dubious”. Inaccuracy in claim or evidence disambiguation really hurts the user experience.
- People advocate a hybrid approach with a templated and argumentative highlight

## 4.5 Discussion

The findings of both parts (quantitative and qualitative) of this *Study II* corroborate our main research assumption of the thesis (1.3.1) that automated reports (as means of argument computation) have a positive effect on the internal Sensemaking of newcomers and perception of the quality of deliberation. The main trigger for the Sensemaking process, we believe is the affordance of automated reports of providing a “quick glance” to the user; offering a manageable amount of information while guiding him to the points of interest. This can be vital to newcomers (or late returning users) to find connections, gain insights, reflect on the information given, and ultimately better *make sense* of the debate.

Regarding which of the examined automated reports improves Sensemaking and quality of deliberation the most, we observed that argumentation highlight outperforms the other two -summary based- types of report. We believe that this is due to its affordance of distinguishing between claims, arguments and evidence, helping people shape their opinions on fairer grounds, a crucial need for users. The full integration of such reports in a fully developed online discussion platform consists of a harder design challenge which will be addressed in the remainder of the thesis, especially in



the development of the synoptical summariser in Chapter 5 and the evaluation of this artefact in a real live platform in Chapter 7.

For some of the quality elements of good deliberation, e.g. resolution of discussion or engagement, and the Sensemaking elements of reducing confusion, uncertainty and ambiguity we did not observe significant effects of the presence of automated reports. Similarly, message quality and presence of new knowledge in the text, as well as equality of participation, and engagement in the discussion are wider design challenges that cannot be addressed by a simple summary, they require future research on new user interaction paradigms, and design solutions that specifically target them.

**Study limitation** Some of the limitations of this study are:

- It does not take into account other factors that possibly affect our target variables, e.g. prior knowledge on the topic, political beliefs, etc.
- The experiment design assumes that Sensemaking and perception of quality of deliberation variables are independent and examines them together. However, this assertion may not stand as there is a strong interplay between those two variables.
- The quantitative study has some shortcomings as only one debate was tested - various debate topics of different thematic categories would have abated any bias coming from the choice of topic.
- The reported critique of reporting methods due to inaccuracy implies that the accuracy of the three reporting methods should have been verified prior to their utilization in the study
- The study design was based on summaries of equal length without investigating the relevance of length and other variables (e.g. amount of arguments, the

structure of those) and how those influence other variables (e.g. participation and decision-making) which can be examined in studies **III** (Chapter 5) and **V** (Chapter 7).

- The virtual execution of the study did not allow for follow-up questions. This was the main driver to set up a focus group afterwards, though we need to acknowledge that it had a small number of participants.

Nonetheless, via our quantitative study, we confirmed our hypothesis (H1,  $A < D$ ) that argument-mining approaches to automated reporting, significantly improve participants' Sensemaking, by increasing reflection (SM1) and idea structuring (SM4), and enhancing the perception of quality of deliberation, by improving message clarity (QoD1). We have also shown relative improvements of this method compared to other automated summary approaches. Our qualitative examination also revealed that accuracy is a weakness of such systems, that originates from the low performance of the underlying NLP system. Though NLP has seen tremendous growth in recent years, it is not yet perfect; actually, the tasks of automatic summarisation and argument component identification (used in our study) are considered a major challenge in the fields of NLP and Argument Mining respectively (see 3.4). Consequently, the generated reports lack coherence, which is vital when results are intended to be used in real environments. We posit that this enduring challenge could be effectively mitigated by introducing novel approaches that emphasise on producing coherent and human-naturalistic text. We present such an approach in Chapter 5.

## 4.6 Summary

Online Discussion platform designers often resort to computational aids to improve participants' sensemaking and quality of deliberation. In this chapter, we examine automated reporting as a promising means of improving sensemaking in discussion

platforms. Through comparison of three approaches to automated reporting: an abstractive summariser, a template report and an argumentation highlighting system, we observed improvements in the Sensemaking of participants and the perception of the overall quality of the deliberation. We suggest that both argument mining technologies and abstractive summarisation are particularly promising computational aids to improve Sensemaking and the perceived quality of online discussion, thanks to their capability to combine computational models for automated reasoning with users' cognitive needs and expectations of automated reporting.

In the next chapter (Chapter 5), we solely focus on abstractive summarisers and the challenges of generating a comprehensive yet coherent summary of the extremely long discussions produced by large crowds.

## Chapter 5

# Summarising Online Discussions - an Argumentation Approach

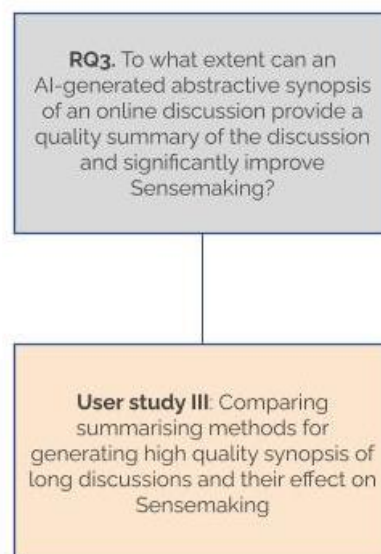
“Words can be like X-rays if you use them properly – they’ll go through anything. You read and you’re pierced.”

---

Aldous Huxley, Brave New World

Comprehending long online discussions poses a significant challenge, either for newcomers or late returning participants. Consequently, the use of automatic summarisation techniques has been investigated as a potential medium to quickly digest the discussion contents and enact sensemaking ([Sanchan et al., 2017](#); [Tigelaar et al., 2010](#)).

Automatically generated summarises, however, face the challenge to achieve an equilib-



rium between being fully comprehensive and maintaining human language natural style.

We investigate the use of state-of-the-art generative large language models (LLMs, see [3.6](#)) in summarising long online discussions and whether they can attain a harmonious balance between adequacy and coherence. To this end, we assess the quality of the generated summary (comparing human and computational evaluation metrics) and then explore how effective the summary is in improving the participants' ability to make sense of a large online discussion (sensemaking effects). We conclude that LLM approaches to summarisation outperform state-of-the-art alternative approaches in the majority of metrics of computational summary quality. We also found that a minimal loss of accuracy can be tolerated -and actually preferred- in favour of human fluency and comprehensiveness when such approaches are used. For what concerns sensemaking effects, GPT3-based summaries also provide better and more significant improvements in participants' Sensemaking compared to other methods considered, even though we observed that different GPT3 model variations affect different sensemaking factors.

## **5.1 Study III - Effect of Prompting Input in Quality Characteristics of Generating Summary**

Automatic text summarisation is the process of presenting one or more text documents while maintaining the main information content using a machine ([Widyassari et al., 2022](#)). In layman's terms, the idea of a summary is to provide a brief overview of the main points or ideas presented in a long text. Generating a summary, however, is a complex task that involves understanding and conveying the "essence" of a longer text.

The style and qualitative characteristics of each summary can be dictated by the

overall context or domain it appears. For example, summaries of clinical medical records need to be concise yet comprehensive; and though there is minimal lexical cohesion, accuracy is non-negotiable (Adams et al., 2021). On the other hand, executive summaries are not meant to list comprehensively all the points of interest but the main purpose is to direct the reader (help him to navigate a long text); for that, a personal (perhaps biased) view is expected (Yaffe, 2020).

Similarly, summaries of online discussions can take a stance rather than just cataloguing the posts in shorter form (essentially replicating the original content). Some approaches follow a purely extractive approach of selecting the main posts (Bhatia et al., 2014; Zhou and Hovy, 2006; Tigelaar et al., 2010), others however, follow an abstractive approach of reshaping the most relevant aspects incorporating the social context as a source of information (Tampe et al., 2022) or by reflecting user needs (Mehdad et al., 2014).

In general, though, any summary faces the challenge to attain a reasonable reduction rate without a high loss of background knowledge (Hahn and Mani, 2000). These shortcomings are amplified on applications of extremely long text, e.g. summarising a book (e.g. Wu et al. (2021) or a large corpus of documents (Haque et al., 2013)). Existing automatic summarisation systems, e.g. Parveen et al. (2016) struggle to maintain coherence throughout the whole of the generated text.

It is imperative to state the extrinsic value of a summary; summary is more than a summary: it connects with cognitive theory (Lemaire et al., 2005) and how people make sense (Mao et al., 2009). The expectation of an online discussion participant from a summary is to get a general sense of the content, that will help him to understand without having to read the long text in its entirety (Almahy and Salim, 2014). Furthermore, such a summary should provide a navigational *affordance*, i.e. what to give attention to and offer ways to examine each point in detail if needed, see previous Chapter 4.

According to information foraging theory ([Pirolli and Card, 1999](#)), people adjust their strategies or the structure of the environment to maximize their rate of gaining valuable information. In Pirolli and Card’s model of Sensemaking ([Pirolli and Card, 2005](#)), it is argued that information processing is happening in distinct steps and is overruled by an “information foraging” loop and a parallel “sensemaking” loop. In that sense, automatic text summarisation becomes an important apparatus for a user to forage relevant information in a long text in a short time with little effort. As people struggle to comprehend large or dynamic discussions, computational aids such as an automatic summary have the potential to enable participants to quickly, but systematically grasp the overall state of a debate before or while they are contributing. This may lead to better quality debate, for instance, by enabling people that are new to the discussion to inform where they could better contribute. Also, automatic summarisation can also be particularly useful in highly dynamic debates, which tend to evolve fast, for returning participants to quickly grasp the progress made since they last contributed, thus reducing idea repetition.

Therefore, a summary generated from an automatic summarisation system should be able to adhere to a series of quality characteristics: fluency, comprehensiveness ([Stunkel et al., 2010](#)), coherence ([Zhang et al., 2016b](#); [Alemany and Fuentes, 2003](#)), readability ([Abdi et al., 2015](#)) among others.

Leveraging previous results on comparison of automated reports from Chapter 4, we developed a long-form text summariser based on prompting a large language model. This type of summariser is best suited for the specific scenario context of an extremely long online discussion, as it accurately captures the essence of the discussion in a human-comprehensible and natural form. Different prompts and hyperparameters (top-p, temperature, length) have been evaluated by human annotators for accuracy, factuality and adequacy. In comparison, other state-of-the-art transformer architectures for long sequences (e.g. long-T5, BigBird) achieve

similar performance in ROUGE-x/BLEU metrics when compared to sample human-generated summaries, however, their final output suffers from readability and fluency, hence our choice to use a GPT-3 model (3rd generation Generative Pre-trained Transformer).

## 5.2 Methodology

Our main aim is to develop a summariser that is suited for enacting Sensemaking and address the challenges of automated reporting presented in the previous Chapter 4. Therefore our main focus is to produce coherent summaries with human-natural language. However, there is an inherent problem in assessing such summaries that pursue this quality characteristic. Automatic summarisation metrics are good indicators for the quality of a summary, they are not, however, on their own sufficient indicators for a summary that serves the purpose of enacting sensemaking. We hypothesise, that there is a significant discrepancy between computational and human evaluation metrics. To test this hypothesis we generated summaries using various state-of-the-art methods and conducted a hybrid evaluation comprising of human judgements and established automatic summarisation metrics.

We sourced long discussions from BCause<sup>1</sup> online deliberation platform (this platform is later presented in full in Chapter 7). The topics chosen were representative of highly debated issues with a plethora of different opinions or arguments on each (climate change, COVID policies, ethics of veganism, comment filtering, hate speech censorship). Moreover, the discussion in BCause platform is organised in an argumentative fashion following the IBIS (Kunz and Rittel, 1970b) approach, i.e. organising posts in positions and supporting or opposing arguments. For each discussion, we generated a summary using the following methods and models:

---

<sup>1</sup><https://bcause.kmi.open.ac.uk/>



- BART is an autoencoder seq2seq model with a generalized BERT bidirectional encoder and GPT decoder (Lewis et al., 2019). We use a BART-Large model that is finetuned on the SamSum Corpus (Gliwa et al., 2019) so it can handle large conversational text chunks.
- T5 is a transformer-based text-to-text language model with a unified framework that converts all text-based language problems into a text-to-text format. Though very capable, T5 is limited to tasks that require shorter input sequences (sequence length: 512 tokens).
- long-T5 is an extension of the original T5 encoder with global-local attention mechanisms that handle long sequences (sequence length up to 16k tokens).
- BigBird is a sparse-attention BERT-style transformer capable to handle much longer sequences (sequence length up to 4k tokens) with SOTA performance in tasks involving long sequences such as summarisation
- GPT-3 an autoregressive large language model with 175 billion parameters. We use it for summarisation by prompting with two different templates  $GPT3_1$  and  $GPT3_2$  (see 1,2). GPT-3 text-davinci-003 model's context window is 4000 tokens long.

At the initial stages of investigating the creation of a summariser of long discussions, we explored various prompts that each gives a different nuance in the generated output summary. We provide a comprehensive list in Table 5.1. We decided to use the somewhat simplistic, yet highly efficacious, prompts shown in Prompts 1 and 2 as those align with the *intended audience* and the desired *quick digestion* nuance we wanted to achieve for displaying those within an online discussion. In addition, we impose limits on the length of the generated discussion by adding the *be brief but complete* directive that resulted in a good compromise between adequacy (summary

---

covering as many parts of the discussion as possible) and limited hallucinations (the generation of text that is false or nonsensical - in the context of summary: the lack of meaningful connection with the original input text ([Ji et al., 2023](#))).

Prompt	Nuance	Output summary
<p>In 2 or 3 sentences, what are the main points one should remember from this discussion on {{discussion_topic}} {{discussion_content}}</p>	<p>highly compressed abstractive</p>	<p>The main points to remember from this discussion are that comment space should be regulated and respectful to others, and that moderation, re-education, and awareness of online actions can help reduce hate speech.</p>
<p>Extract key point from the discussion below based on which policy makers can make their decision more efficient:</p> <p>debate                      topic: {{discussion_topic}}</p> <p>debate                      description: {{discussion_description}}</p> <p>debate                      discussion: {{discussion_content}}</p>	<p>intended audience purpose extractive</p>	<p>Key points:</p> <ol style="list-style-type: none"> <li>1. Regulation of comments by setting standards to promote tolerance and open-mindedness.</li> <li>2. Moderation of comments by a team of humans or automatically.</li> <li>3. Rewiring perspectives to cultivate values such as open-mindedness and respect.</li> <li>4. Requiring personal data/information to be submitted to track down people who participate in hate speech and attempt to re-educate them.</li> </ol>
<p>Condense the discussion below to the essentials to present it in the form of short cards used in a presentation</p> <p>debate                      topic: {{discussion_topic}}</p> <p>debate                      description: {{discussion_description}}</p> <p>debate                      discussion: {{discussion_content}}</p>	<p>subjective fairly comprehensive</p>	<p>Card 1: Comment space should be regulated, not eliminated, to reduce hate speech.</p> <p>Card 2: Moderators should consistently monitor and regulate comments.</p> <p>Card 3: Education and rewiring perspectives can reduce hate speech.</p> <p>Card 4: Personal data submission may discourage hate speech.</p>

Generate a TLDR (too long didn't read) summary of the fol- lowing discussion on the topic of "{{discussion_topic}}"	fairly comprehensive abstractive quick digestion	TLDR: Comment space should be regulated to sup- port tolerance and open-mindedness, and hate speech should be removed. We should moderate comments automatically or by a team of humans, and users should be responsible for their attitude. Education and rewiring perspectives can help cultivate values like open-mindedness and respect. Another solution could be requiring personal data/information to be submitted before commenting, to make people aware of their actions online.
debate discussion:{{discussion_content}}	discus-	
Provide me a thorough synop- sis of this debate that is about: {{debate_topic}} described as {{debate_description}}		
Cover as many viewpoints as you can.	fairly comprehensive	
Focus on the main arguments used in the discussion.	highly abstractive wide coverage	The general consensus of the discussion seems to be that regulating comments – by removing hate speech and encouraging respectful discussion – is a good idea, but that it is difficult to enforce. There are a variety of opinions on whether or not removing the comment space from newspaper websites is a fair choice, but the majority seem to think that it is not.
Use a narrative style instead of simply outlining points on the dis- cussion		
Discussion content:{{discussion_content}}	con-	

Table 5.1: Different nuances on generated summaries on the same discussion using different prompts

### 5.2.1 Generation Strategies

The discussion was pre-processed to convert it to a long text document using two variations: (i) *simple*: each participant's post was given in a new line, the final document is simply a concatenation of all the posts (ii) Participant's name was preserved and prefixed before each statement, and discussion argumentative operations were explicitly stated (e.g. "User John *replied* to above position stating that ... "). Action verbs to attach the argumentative notion (as that was imposed by the design of the platform) such as *replied*, *opposed*, *supported*, *counter-argued* were used (see Generation Strategy 3).

During the initial phases of the experimental trials, it became apparent that simple strategy was inadequate; therefore it was omitted for the rest of the experiment. This is in accordance with the literature; as LLMs are widely acknowledged that are not proficient in understanding the presence of multiple, concurrent, and potentially conflicting meanings or interpretations within a given text [Mitchell and Krakauer \(2023\)](#).

---

#### **Prompt 1** Prompt template - prefixed synopsis

---

Provide me a thorough synopsis of this discussion:

{{ discussion\_content }}

Be brief but as complete as you can. Start your synopsis with: This discussion is about

---



---

#### **Prompt 2** Prompt template - prefixed synopsis follow on

---

This discussion is about {{ discussion\_previous\_summary }}

The discussion above continues:

{{ discussion\_content }}

Combine with the above discussion and provide a thorough synopsis of the whole discussion.

Start your synopsis with: This discussion is about

---

**Generation strategy 3** Discussion to text generation strategy

---

```

$long_text = ""
for $position in $positions
    long_text.append("User {{$position.userName}} {{$has/replied with}} the posi-
tion {{$position.text}}")
    if position.hasArguments:
        for $arg in $position.arguments:
            if $arg.isOpposing:
                $long_text.append("User {{$arg.user}} opposed it with {{$arg.text}}")
            else if $arg.isSupporting:
                $long_text.append("User {{$arg.user}} supported it with {{$arg.text}}")

```

---

Topic	# Pos	# <i>Arg</i> <sub>+</sub>	# <i>Arg</i> <sub>-</sub>	# words	<i>words</i>	# Users
How Can Manmade Climate Change Be Reversed?	140	98	106	10335	30.0	70
Should all humans go vegan?	5	5	5	259	17.3	6
How to make "Living With COVID" more than an empty political slogan?	5	4	3	248	20.7	3
Do you think removing the comment space from newspapers' website is a fair choice to reduce hate speech?	35	18	24	2054	26.7	23
What are the key factors influencing the fairness of filtering mechanisms to reduce hate speech?	11	7	13	688	22.2	18

Table 5.2: Descriptive stats of the debates used in the creation of the recommendation corpus

## How to make "Living With COVID" more than an empty political slogan?

All over the world, governments decide that "COVID is over" and that we need to live with it. Restrictions/protections are cancelled. However, there is a risk that especially the vulnerable are now forced to lock themselves up indefinitely. How can we find a middle ground that allows society to open up sensibly, so that life can resume a safe extent, while not excluding a large portion of the population from participating in it?

> (Position): *Roger Lane* : We need a minimal set of protection measures (masks, ventilation...) in public spaces and public transportation.

>>> (+ pro argument) *Roger Lane*: There is a growing body of evidence that everybody - not just the vulnerable - wearing masks significantly reduces the viral load in a confined space..

>>> (+ pro argument) *Roger Lane*: This goes beyond just technical issues, it is also a matter of civilization for the general public to show it cares about fellow citizens with disabilities. .

> (Position): *Anthony Harrison* : I believe that mandatory mask-wearing policies infringe upon personal liberty, and that mask mandates are politicized and are manipulative tactics by certain politicians and special interest groups as ways to increase state control

>>> (- con argument) *Anthony Harrison*: I disagree - these policies must be mandated otherwise when they are voluntary (like in the UK now) people don't wear them - as a result infection rates remain high .

>>> (- con argument) *Roger Lane*: But "personal liberty" can't be absolute: one person's liberty should not force another person into hiding. Also, the whole "state control" meme seems to be disingenuous: in times of great crisis the state not only may but MUST intervene for the safety and security of all. That is the core reason for the state to exist..

>>> (+ pro argument) *Roger Lane*: Exactly, if not mandated there's also a strong negative social pressure effect where many people who are actually in favour of wearing masks don't dare to do so because of fear of being the social outlier or even treated aggressively. .

> (Position): *Marissa Price* : position

> (Position): *Anthony Harrison* : make it so

>>> (+ pro argument) *Anthony Harrison*: indeed .

> (Position): *Anthony Harrison* : living with covid will need much new investment in primary care

>>> (- con argument) *Anthony Harrison*: no it wont.

Figure 5.1: Example discussion as it was rendered for the mechanical turk crowd-worker

### Machine generated summary

Below there is a machine generated summary of the discussion above:

. Primary care needs a comprehensive strategy that looks at the needs of the entire health system, including staffing, infrastructure, and services, to ensure it can continue to deliver quality care during the pandemic. This should include a focus on increasing access to primary care, improving access to mental health services, and strengthening the capacity of primary care providers to deliver care. Additionally, primary care should be supported in developing new models of care that are more efficient and cost-effective, such as telehealth and virtual visits, to ensure that care can be provided to more people in more locations.

Figure 5.2: Example summary as it was rendered for the mechanical turk crowdworker

### 5.2.2 Experimental design

For most of the models used we had to employ a recursive generation approach (apart of longT5 model that can handle 16k tokens long text which for 4 out of 5 debates it was enough). For that, we used the following two strategies:

- *Mix\_generations* (see Prompt 1): Split overall text into sequences of up to *model sequence length* (use sentence segmentation to keep sentences intact). Generate a summary for each segment. Concatenate segment summaries and if necessary recursively repeat until the desired length of output summary is reached.
- *Follow\_sequences* (see Prompt 2): Generate the summary of the first segment of *model sequence length* tokens - set the summary length to half of *model sequence length*. Use the generated summary as a prefix of the next segment of the text and generate another summary. Proceed until you have processed all the text. Note that after the first, sequences are up to half of the model's maximum length.

An example of using the above *Mix\_generations* and *Follow\_sequences* strategies on summarising the same debate about climate change is presented in Tables 5.3 and 5.4 respectively. The *Mix\_generations* offers a more comprehensive summary as it collects information from all parts of the discussion regardless of at what point they appear, while *Follow\_sequences* gives less importance to text that appeared early, as on the last iteration the information about the rest of the debate is given in compact form (and already filtered through the summarisation). This can be in some cases desirable, for example when you want to emphasise the latest chronological developments in a discussion; however, in most cases, a more comprehensive summary as in *Mix\_generations* is preferred. On the other hand *Mix\_generations* requires more invocations of the summarising process, namely  $n \lceil \log_2 n \rceil + 1$ , while *Follow\_sequences* only  $n$  number of invocations, where  $n$  is the number of sequences created after splitting text (in our example  $n = 2$ , as our sequence max length was 512 tokens).



**(i) GPT-3 (2nd var.)**

This discussion is about the potential positive environmental benefits of veganism and whether it is feasible for the majority of the population. Some users argued that veganism is the only sustainable lifestyle and should not be forced on people. Others argued that vegetarianism can achieve environmental benefits and that veganism is too extreme and expensive for many people. It was also mentioned that enforcing what people should eat violates the principle of freedom, and that choosing an unsustainable diet affects not only the individual but the entire planet.

**(ii) t5-large**

This discussion is about veganism. while more ethical, good for the environment, and healthier diet, is not feasible for a great proportion of the human population. some users oppose it. All the environment benefits can be achieved just by a vegetarian diet, you don't have to go fully vegan. others support it saying veganism is a choice, it should not be forced - especially to children Another user opposed it with "this is extreme".

Figure 5.3: Examples of synoptical summaries on the debate topic : “Should all humans go vegan?” using (i) GPT-3 (2nd var.) and (ii) t5-large

This is important in case invoking the summariser is expensive, e.g. paid API, or creation time is crucial for user experience, e.g. expecting a quick re-generation of a summary on every small edit of the input text. For example, in the case of a discussion spanning 5000 words ( $\sim 6.7k$  tokens), that would be split into  $n = 7$  sequences of 1024 tokens each. In the case of *Mix\_generations* it would take 22 summariser invocations; for *Follow\_sequences* it would only take 7, a great cost saving.

For each summary, we calculated the ROUGE-1, ROUGE-2, ROUGE-L, BLEU, METEOR and BERTScore scores against two reference summaries (created by two separate experts). The average scores are presented in Table 5.5

### 5.2.3 Task

We used Mechanical Turk<sup>2</sup> to recruit crowd workers for a reward of 8\$ per hour and carried out a cross-sectional between subjects study. Each crowd worker was presented with a simplified rendered output of the original discussion 5.1 among with the generated summary 5.2 and was tasked to rate the summary on the 9 metrics of table 5.6 using a 5-point Likert scale. Furthermore, they were asked to evaluate their understanding of the debate using the 9 sensemaking features of Table 5.7. Notice

<sup>2</sup><https://www.mturk.com/>

---

Summary of seq 1

---

The discussion is about the idea that it may be possible to reverse climate change, but it is getting harder and harder every year. The two main arguments in the discussion are that it may be technically possible to reverse climate change, but it is getting harder and harder every year, and that if governments around the world treat climate change collectively - the same way they treated World War 2 for instance - then is possible, no matter the cost. Other users also weigh in, with some supporting the idea, and others opposing it. One user argues that every year that passes we lose tools or opportunities that could work to counteract climate change, as well as more damage is added. Another user argues that even if it is getting harder we have to at least try. However, another user believes that it is not possible to technically change the climate problem.

---



---

Summary of seq 2

---

The discussion is about the idea that technology can't solve the climate change problem. Some users agree that technology can't solve the climate change problem, while others disagree. One user supports the idea that technology can't solve the climate change problem, while another user opposes the idea. However, the majority of users in this discussion believe that technology can help solve the climate change problem. One user supports the idea that technology can help solve the climate change problem, while another user opposes the idea. However, the majority of users in this discussion believe that technology can help solve the climate change problem.

---



---

After mixing sequences and resummarising

---

The discussion revolves around the possibility of reversing climate change, with two main arguments. Some believe it's technically possible but getting harder every year, while others argue that collective action by governments worldwide can make it possible regardless of cost. Some users support the idea, while others oppose it. Some argue that every year we lose tools and opportunities to counteract climate change, while others believe that even if it's getting harder, we should try. There's also a discussion about the role of technology in solving the climate change problem, with some users agreeing and others disagreeing. Overall, the majority believe that technology can help solve the problem.

---

Table 5.3: Example summary using the *Mix<sub>generations</sub>generationstrategy*

After sequence 1
<p>This discussion is about whether or not it is possible to reverse climate change. Some users believe that it may be possible to reverse climate change, while others believe that it is getting harder year by year and it is not feasible with current technology. The discussion also focuses on the economic impacts of climate change, and whether or not it is more important to prevent the problem or to focus on other priorities.</p>
After sequence 2
<p>This discussion is about whether or not it is possible to reverse climate change. Several users discuss the possibility of reversing climate change, with some believing that it is possible and others believing that it is becoming harder and harder to do so. It is also discussed whether or not it is more important to prevent the problem or to focus on other priorities. The discussion also focuses on the economic impacts of climate change, and whether or not it is more important to adapt to the changing climate or to try and fix it. Despite the challenges, many users believe that it is still possible to reverse climate change if enough effort is put into it.</p>

Table 5.4: Example summary using the *Follow\_sequences* generation strategy

Table 5.5: ROUGE-1, ROUGE-2, ROUGE-L, METEOR, BLEU and BERTScore scores on different models used for summary generation

model	R-1	R-2	R-L	METEOR	BLEU	BERTScore-F1
BigBird	14.56	1.73	9.63	13.36	11.77	0.690
BART-L	<b>25.33</b>	<b>4.21</b>	<b>13.91</b>	<b>23.47</b>	19.40	0.753
long-T5	17.15	2.85	12.00	16.08	16.21	<b>0.756</b>
t5-large	23.77	3.31	13.72	18.06	<b>27.27</b>	0.745
GPT-3 (1st var.)	21.52	4.00	12.83	19.59	25.63	0.724
GPT-3 (2nd var.)	22.43	3.70	13.63	19.44	24.94	0.772

Table 5.6: Human evaluation factors and corresponding given prompt to Mechanical Turk crowdworkers

Code	Factor	Crowdworker Prompt
qos1	Fluency	The machine generated summary is grammatically correct and easy to understand
qos2	Comprehensiveness	The machine generated summary covers the key points and main ideas of the original text
qos3	Coherence	The machine generated summary is logically organized and flows smoothly from one idea to the next
qos4	Relevance	The machine generated summary includes information that is relevant to the main topic of the original discussion
qos5	Concision	The machine generated summary is as concise as possible without losing important information
qos6	Readability	The machine generated summary is written at an appropriate level of difficulty
qos7	Accuracy	The machine generated summary accurately represents the content of the original text
qos8	Factuality	The machine generated summary accurately reflects the facts and information presented in the original discussion
qos9	Adequacy	The machine generated summary is sufficient to convey the main ideas and key points of the original discussion

that some crowdworker prompts are stated in a negative form - this was to avoid or reduce response bias. The scores for the negative questions were subsequently reversed to align consistently with all other questions. Each summary was annotated by 3 workers and the average rating was used. In total 600 (200 unique) annotations are used for our analysis.

## 5.3 Results

### 5.3.1 Computational Automatic Evaluation

We present the results of the computational evaluation of the summaries in Table 5.5. We observe that BART-L model demonstrates the highest score for ROUGE-1, ROUGE-2, ROUGE-L and METEOR metrics, while the highest BLEU score is achieved by T5-large model. At the same time, the lowest score is consistently

Table 5.7: Sensemaking evaluation factors and corresponding prompts to Mechanical Turk crowdworkers

Code	Factor	Crowdworker Prompt
SM1	Reflection	I was able to reflect on the debated question
SM2	Insight	I was provided with unexpected insights on what is the question and what are the main arguments for and against.
SM3	Focus	I was not able to focus on different aspects of the debate
SM4	Argumentation	I was able to find structure in the information provided in this debate and find a way to organise it
SM5	Explanation	I was not able to identify the main points raised in this debate
SM6	Evaluate facts and evidence	I was able to assess facts and evidence provided in this debate
SM7	Distinguish	I was able to distinguish between different people’s claims
SM8	Assess assumptions	I was not able to assess my initial assumptions about this debate
SM9	Change assumptions	Some initial assumptions I had about this question changed

achieved by BigBird. In general, we observe that the long-sequences models (long-T5 and BigBird) underperform compared to standard sequence length models that follow a custom split-sequence strategy. Both GPT-3 prompt methods (variation 1 and 2 of generating the discussion content) yield comparative results in relation to other models, they fail however to attain the highest performance score.

### 5.3.2 Human evaluation

We present the results of the human evaluation in summary quality metrics in Table 5.8 and scores in sensemaking features in Table 5.9. The main observation is that both GPT-3 prompting variations exhibit the highest level of performance score in Fluency, Comprehensiveness, Coherence, Relevance, Concision, Readability features while ranking closely in Accuracy, Factuality and Adequacy features (BART and T5-large perform best in these features). The least favourable performance is observed by BigBird.

We observe limited correlation between computational and human evaluation metrics, see Figure 5.4. There is as expected high intra-correlation between computational metrics, with the exception of ROUGE-1 and ROUGE-2 pair relatively low  $r=0.36$ .

Table 5.8: Human evaluation metrics scores

	Fluency	Compr.	Coherence	Relev.	Concision	Read/ity	Accu.	Fac/ity	Adeq.	Avg
BigBird	1.5	0.85	1.3	1.25	1.05	1.6	0.9	0.8	1.5	1.19
BART	2.65	2.5	2.525	2.425	2.35	2.6	2.425	2.4	2.125	2.44
GPT-3	3.2	2.825	3.05	2.925	2.575	<b>2.9</b>	2.3	1.925	1.775	2.61
(1st)										
GPT-3	3.25	2.5	2.9	2.7	2.45	2.875	1.925	2.125	1.9	2.51
(2nd)										
long-T5	2.3	2.05	2.1	2.5	1.7	2.25	1.6	1.3	1.45	1.92
T5-large	2.65	2.625	2.55	2.45	2.325	2.35	2.425	2.225	2.575	2.46
Average	2.73	2.38	2.545	2.475	2.215	2.53	2.065	1.945	1.97	

Similarly, there is a high correlation to similar concepts in human evaluation metrics (e.g. Fluency/Coherence:  $r=0.68$ , Compreheviness/Relevance:  $r=0.62$ ). Interestingly, there is a low correlation in the pairs of Readability and Factuality ( $r=0.27$ ), Fluency and Adequacy ( $r=0.21$ ), Fluency and Factuality ( $r=0.24$ ).

### Effect on Sensemaking

We present the score of each method in the 9 factors of sensemaking (SM1-SM9) in Table 5.9. Overall, we observe high scores in SM1-SM4 features (Reflection, Insights, Focus, Argumentation) and relatively low in the rest SM5-SM9 features (Explanation, Evaluate facts and evidence, Distinguish, Assess and Change Assumptions); that verify the findings of a quantitative comparative study of Chapter 4.

Upon conducting a column-wise comparison of the models, we observe that GPT3, whether the 1st or 2nd variation, outperforms the other models in 8 out of 9 features. However, it should be noted that SM3 (Focus) does not receive the highest score, in that particular feature long-T5 receives the highest score.

## 5.4 Discussion

In this chapter, we have tested the performance of various state-of-the-art summary models for the use case of summarising long online discussions. We performed a hybrid evaluation approach, to measure both the computational performance using

model	SM1	SM2	SM3	SM4	SM5	SM6	SM7	SM8	SM9	Average
BigBird	2.85	2.56	2.29	2.60	1.60	2.31	1.60	2.26	2.00	2.23
BART	2.80	2.54	2.63	2.78	1.88	2.14	1.63	2.24	1.91	2.28
long-T5	2.95	2.31	<b>3.03</b>	2.80	1.75	2.17	1.45	2.30	2.55	2.37
t5-large	2.75	2.41	2.73	2.93	1.75	2.25	1.65	2.22	1.88	2.29
GPT3 (1st var.)	<b>2.98</b>	2.54	2.67	<b>2.95</b>	<b>2.05*</b>	2.61	<b>1.95</b>	2.22	2.20	2.46
GPT3 (2nd var.)	2.83	<b>2.89*</b>	2.93	2.88	1.86	<b>2.75*</b>	1.84	<b>2.43</b>	<b>2.58*</b>	<b>2.55</b>
Average	<b>2.86</b>	2.54	2.71	2.82	1.82	2.37	1.69	2.28	2.19	2.36

Table 5.9: Summarisers models scores in Sensemaking dimensions

standardised metrics and the quality of summaries when judged by humans. Further, than the intrinsic evaluation of the quality of the output summary, we performed an extrinsic evaluation to measure the impact on the overarching task for which the summary is intended. We, therefore, compared the impact of Sensemaking of the overall discussion, when the summary is presented alongside the original discussion. We recognise that our study has limitations. Firstly, the dataset used for evaluation was limited only to a few (yet diverse) topics. A bigger sample originating from distinctly different thematic categories would have significantly reduced any bias arising from the choice of topic. Also, this multi-condition experimental study would be better to be conducted as a within-subjects study (instead of a between-subjects) with pairwise comparisons as it would have reduced variability (eventually requiring a smaller sample), however, within-subjects study may lead to order or carryover effects and render the research question transparent (subjects will be aware of the experiment purpose). Moreover, a qualitative analysis, even an error analysis (compare summaries with distant scores) would be more helpful to uncover the preferable characteristics of a summary that a human prefers. It is worth noting that the generation strategies do not fully leverage the inherent semantics of the conversation, although this limitation was addressed by explicitly incorporating them into the document preparation process (as explained in Section 5.2.1).

Considering the aforementioned limitations, we can still confirm the original hypoth-

esis that computational metrics do not directly correlate with human evaluation metrics. Whereas BART and T5 outperformed in computational metrics, their performance was average according to the human evaluation metrics. GPT-3 as a large language generative transformer model when prompted correctly can generate coherent, narrative and argumentative text that factually and adequately reflects the given text. Its rhetorical capability outperforms other models that have been fine-tuned or explicitly designed for summary applications.

But can we tolerate imperfect summaries for the benefit of readability or fluency? The process of Sensemaking is intrinsically connected to cognitive processes in humans. Specifically, when presented with a long text, moreover a complex long discussion of various actors and various exchanged arguments, a highly demanding cognitive effort is required to comprehend it, comprising of several information-seeking loops before reaching the point to construct knowledge ([Pontis and Blandford, 2015](#)). It is evident that users are willing to “forgive” loss (up to a certain point) of accuracy and adequacy (appearance of hallucinations) - a known issue of LLMs ([Bender et al., 2021](#))- if this is beneficial for their easier understanding. Contrary, users disapprove of highly accurate yet incomprehensible summaries. Recent research seeks to make generated text hallucination-free, for example, to post-hoc examine if the entities in the output summary correspond to the ones in the reference text ([Zhao et al., 2020](#)). All models examined are based on the transformer architecture which allows them to attend to the various segments of the given text and assign its importance. Consequently, they manage very well to reveal the underlying meaning in each segment they examine, however when addressing the long discussion as a whole they fail to convey a coherent narrative story - according to the study subjects it feels like a bullet list of factual statements. Exceptionally though, GPT-3’s performance is impressive compared to all other models examined, considering that we used a zero-shot approach in a model not fine-tuned for summarisation. Regardless, if GPT-3 is just a stochastic parrot ([Bender et al., 2021](#)) or has indeed the necessary knowledge and reasoning capacity, it can indeed produce a concise meaningful summary (with minimal “hallucinations”) that can empower human sensemaking.



## 5.5 Summary

Overall, this chapter tests the performance of various state-of-the-art summary models on online discussions using a hybrid evaluation approach. This approach measures both computational performances using standardised automatic summarised metrics and summary quality based on human judgements, with an extrinsic evaluation assessing the impact on Sensemaking of the discussion. Results suggest that prompting Large Language models (LLMs) is the best method for generating quality summaries, with LLM-based summaries having the highest positive effect on Sensemaking.

The subsequent chapter of this thesis will showcase the design and development of the second artefact, a scientific Argument Recommender System (*SciArgRecSys*). Following this, in Chapter 7, a comprehensive evaluation of both artefacts will be presented based on their deployment in a live online discussion platform (BCause).

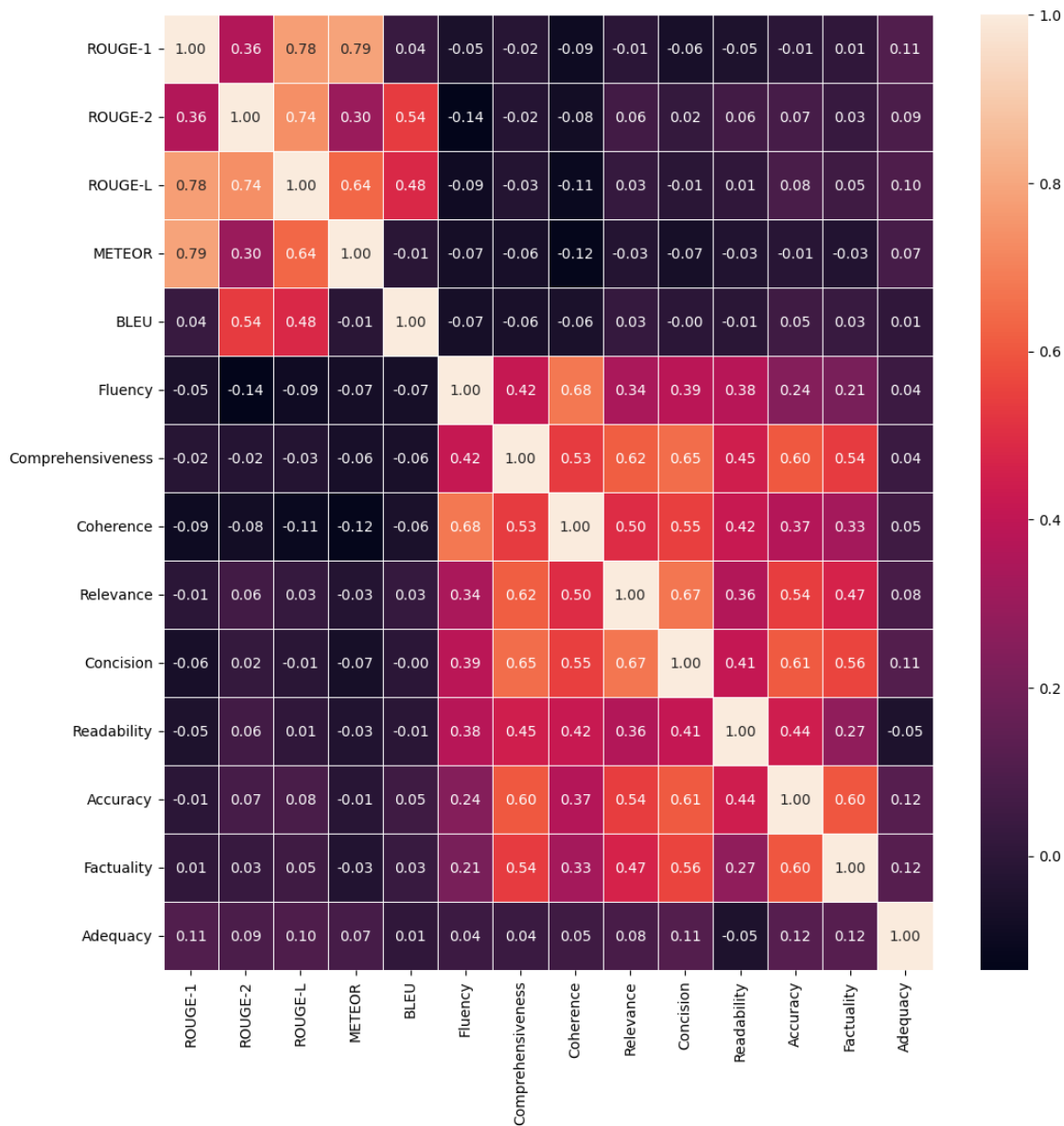


Figure 5.4: Pearson Correlation matrix of computational metrics (ROUGE-x, METEOR, BLEU, BERTScore) and human evaluation metrics

## Chapter 6

# Scientific Argument Recommender System

“The real voyage of discovery  
consists not in seeking new  
landscapes but in having new eyes.”

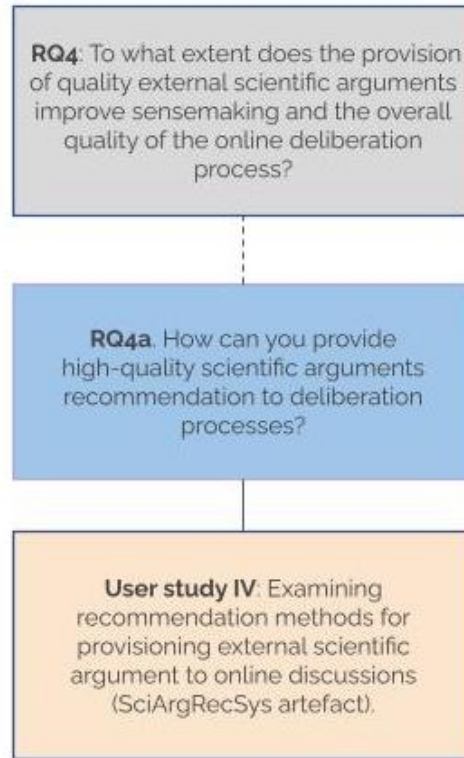
---

Marcel Proust.

Previous research has shown that online discussions often build on shallow content and evidence of unreliable quality. Integrating a recommender system for scientific arguments sourced from scientific literature posits a remarkable opportunity to foster Evidence-Based reasoning (EBR) in online discussions. In this chapter, we delineate the methodological trajectory undertaken in the development of a recommender system designed to provide scientific arguments in online discussions.

We investigate research question RQ4a [1.3.2](#): “*How can you provide high-quality scientific arguments recommendation to deliberation processes?*”. In order to accomplish this, our initial step involves a thorough investigation into whether extracting arguments from scientific literature can be executed accurately at scale (Section [6.1.1](#)).

We present a method of generating a summary of the research paper abstract that conveys the main argument as a proposed recommendation unit in Section 6.2.3. We then evaluate different methods of recommending scientific evidence taking into account distinct granular levels of argument: i. short quoted extracts (excerpts) from research papers, ii. research paper abstracts and iii. summarised abstracts depicting the main argument of the research paper (Section 6.3).



## 6.1 Extracting Arguments from Scientific Literature in Scale

### 6.1.1 Argument Mining for Scientific Discourse

Argument mining from a general-purpose corpora (discovery of argumentative units, e.g. propositions, conclusions, premises, warrants, etc.) differs from argument mining from scientific corpora (discovery of scientific claims, interpretation, strength of experiment evidence, etc). ArguminSci (Lauscher et al., 2018a) was one of the first publicly available systems that provided fine-grained argumentative analysis of scientific publications (earlier attempts would include Green (2014) that focused on the biomedical domain). It would also generate a publicly available corpus (Lauscher et al., 2018b) of argument-annotated scientific publications (in the domain of computer graphics). While there are various argument models - and their selection

affects significantly model design and performance - ArguminSci's argumentation scheme is rather simple consisting only of three types of argumentative components, namely (i) *background claim*: a statement of argumentative nature, which is about or closely related to the work of others or common practices in a research field or about background facts related to the topic of the publication. (ii) *own claim*: a statement of argumentative nature, which related to the authors' own work and contribution., and (iii) *data*: factual evidence for or against a claim. We deliberately avoided choosing any of the much richer argumentation schemes, e.g. Toulmin (Toulmin, 2003), Reed and Walton (Walton et al., 2008), schemes more specific to scientific literature such as Argumentative Zoning that assign roles to large spans of text or schemes that are domain-specific, e.g. Green (2018) for biomedical articles. Choosing a simple argumentation scheme in building an argument recommender system for online discussion has several advantages, justifying the selection over more complex and capable models. For example:

- ease of understanding: no special training is required for interpreting the class of argument by the end-users - especially if those are not academics
- computationally efficient and scalable: it is expected that a simpler model will require less computational resources and can scale up to a large number of users with no special customisation
- reduced overfitting: more complex models can lead to overfitting, i.e. perform well on the training data but poorly in unseen data. This is due to the added degrees of freedom that need large training datasets to generalise well. Also, the available datasets for complex argumentative schemes often consist of imbalanced classes that may overfit by focusing disproportionately on frequent classes.
- transparent and explainable: simpler models can be easier to interpret and explain, which helps to build trust to the system

We used the ArguminSci corpus to train a BERT-based model (Devlin et al., 2018)

for automatic argument unit identification. We did that by splitting the ArguminSci corpus into sequences up to 512 tokens and using those to fine-tune BERT for text classification (using the transfer learning paradigm (Torrey and Shavlik, 2010; Sun et al., 2019)). To start by inserting a special [CLS] token at the beginning of each sequence. The [CLS] token is used as the aggregate sequence representation for classification tasks. We then pass it through the fully connected layer of BERT and a softmax function to generate probabilities for each class, in our case 4 classes: own\_claim, background\_claim, data and NO\_LABEL. We trained BERT large uncased<sup>1</sup> (336M parameters) using HuggingFace API<sup>2</sup> using an NVIDIA Tesla V100 SXM2 16 GB GPU for 10 epochs, with a batch size of 4, learning rate=5e-05 and max sequence length of 512.

We also attempted to use a generative model for text classification. In this case, we used GPT-3 (Brown et al., 2020) two different engines: (i) davinci: the fully capable 175b parameters and (ii) curie: a small but still capable (6.3b parameters). For both engines, we used two prompts - a zero-shot (see Figure 4) and a few-examples (see Figure 5). We also used GPT-3 as is and a fine-tuned version with few samples (9000 for davinci and 19159 for curie - less for davinci due to cost considerations).

---

**Prompt 4** Prompt template - 0-shot argument extraction

---

Classify the following sentence in the following argumentative labels: own\_claim, background\_claim, data or NO\_LABEL:

Sentence: {{sentence}}

Label:

---

## Results

To evaluate the performance of three models for argument component identification: (i) ArguminSci, (ii) fine-tuned BERT (iii) GPT-3 using two engines (davinci, curie), two prompting templates (0-shot, few-shot) and two tuning level (normal and fine-tuned) - in total 2x2x2=8 different GPT configurations. We present the performance

---

<sup>1</sup><https://huggingface.co/bert-large-uncased>

<sup>2</sup>[https://huggingface.co/transformers/v3.0.2/main\\_classes/trainer.html](https://huggingface.co/transformers/v3.0.2/main_classes/trainer.html)

**(i) ArguminSci**

Promoting urban greenery through tree planting strategies has been considered as a measure to mitigate climate change. While it is essential to understand the temporal dynamics of urban forest structure as well as its services and contribution to human well-being in cities, it has hardly ever been examined whether the future contributions of these services after different possible planting strategies can comply with climate change policy goals; [...] We conclude that urban tree planting has a small impact on carbon mitigation in the study area, most likely because of the young age of trees in Tabriz as well as the fact that the planted trees cannot deliver all their benefits over a 20-year period and need more time. Thus, the use of urban trees serves only as a complementary solution rather than an alternative climate mitigation strategy. Our quantitative approach helps urban environmental policymakers to evaluate how much they can rely on urban forest strategies to achieve climate change mitigation targets.

**(ii) BERT based**

Promoting urban greenery through tree planting strategies has been considered as a measure to mitigate climate change. While it is essential to understand the temporal dynamics of urban forest structure as well as its services and contribution to human well-being in cities, it has hardly ever been examined whether the future contributions of these services after different possible planting strategies can comply with climate change policy goals; [...] We conclude that urban tree planting has a small impact on carbon mitigation in the study area, most likely because of the young age of trees in Tabriz as well as the fact that the planted trees cannot deliver all their benefits over a 20-year period and need more time. Thus, the use of urban trees serve only as a complementary solution rather than an alternative climate mitigation strategy. Our quantitative approach helps urban environmental policymakers to evaluate how much they can rely on urban forest strategies to achieve climate change mitigation targets.

**(iii) GPT-3 based**

Promoting urban greenery through tree planting strategies has been considered as a measure to mitigate climate change. While it is essential to understand the temporal dynamics of urban forest structure as well as its services and contribution to human well-being in cities, it has hardly ever been examined whether the future contributions of these services after different possible planting strategies can comply with climate change policy goals; [...] We conclude that urban tree planting has a small impact on carbon mitigation in the study area, most likely because of the young age of trees in Tabriz as well as the fact that the planted trees cannot deliver all their benefits over a 20-years period and need more time. Thus, the use of urban trees serves only as a complementary solution rather than an alternative climate mitigation strategy. Our quantitative approach helps urban environmental policymakers to evaluate how much they can rely on urban forest strategies to achieve climate change mitigation targets.

Figure 6.1: Examples of automatic argument annotation using (i) ArguminSci model, (ii) BERT based model and (iii) GPT-3 based model. Argumentation component classes are own claim, background claim and data

**Prompt 5** Prompt template - few shot argument extraction

Here are some examples of sentences classified as own\_claim, background\_claim, data or NO\_LABEL:

Sentence: {{sentence\_1}}

Label: {{label\_1}}

Sentence: {{sentence\_2}}

Label: {{label\_2}}

...

Sentence: {{sentence\_n}}

Label: {{label\_n}}

Decide whether the following sentence is classified as own\_claim, background\_claim, data or NO\_LABEL:

Sentence: {{sentence\_c}}

Label:

of each model in Table 6.1, showing the overall accuracy and the F1-score in each class (own claim/oc, background claim/bc, data/d and non-argument/no label). Most of the models -compared to ArguminSci as a baseline- have greater macro-F1 scores and accuracy. It is clear that fine-tuned BERT outperforms all other models, even the fine-tuned GPT-3 most powerful engine (davinci). Comparing fine-tuning and plain GPT-3 models, we observe the rather counter-intuitive drop in performance when prompting with few examples and the increase in performance in the case of a 0-shot example. Furthermore, the davinci engine (~175 billion parameters) significantly outperforms as expected the smaller curie engine. Similarly, when comparing the same model, using few example prompt performs better -as expected- than 0-shot (with a noticeable exception of the fine-tuned GPT-3 davinci).

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (6.1)$$

$$precision = \frac{TP}{TP + FP} \quad (6.2)$$

$$recall = \frac{TP}{TP + FN} \quad (6.3)$$

$$F_1 = 2 \frac{precision \times recall}{precision + recall} = \frac{2TP}{2TP + FP + FN} \quad (6.4)$$



model	prompt	accuracy	oc	bc	d	no label	$F_1$
ArguminSci	-	0.25	0.27	0.11	0.28	0.19	0.21
BERT finetuned	-	<b>0.84</b>	<b>1</b>	<b>0.21</b>	<b>0.75</b>	<b>0.74</b>	<b>0.67</b>
GPT-3 curie	0-shot	0.15	0.12	0.19	0.12	0.18	0.15
GPT-3 curie	few-shot	0.23	0.04	0	0.41	0.29	0.19
GPT-3 davinci	0-shot	0.36	0.56	0.07	0.15	0.12	0.23
GPT-3 davinci	few-shot	0.49	<b>0.64</b>	<b>0.21</b>	0.51	0.4	0.44
GPT-3 curie fine	0-shot	0.23	0.17	0.06	0.14	0.34	0.18
GPT-3 curie fine	few-shot	0.27	0.23	0.19	0.2	0.35	0.24
GPT-3 davinci fine	0-shot	0.28	0.37	0.17	0.16	0.26	0.24
GPT-3 davinci fine	few-shot	0.25	0.13	0.2	0.16	0.38	0.22

Table 6.1: Experiment results - accuracy, individual classes F1 and macro-F1

where,

$TP$  = Number of true positives, class identified correctly

$TN$  = Number of true negatives, no class identified correctly

$FP$  = Number of false positives, class identified incorrectly

$FN$  = Number of false negatives, no class identified incorrectly

## Discussion

This study presents a small investigation of whether a large language model (LLM) - in this case, GPT-3 - can be utilised in a few-shot learning scenario for extracting argumentative components from scientific literature. The model's performance is compared with (i) ArguminSci - an older neural model for the same task and (ii) a fine-tuned BERT model. The evaluation was based on ArguminSci corpus.

The results show that while GPT-3 improves significantly the performance of this task when compared to the initial baseline model, it still struggles when compared with a standard transformer model. This could be due to several reasons, for instance, the randomness introduced by temperature and diversity of sampling while generating could lead to non-deterministic results, which is not well suited for classification tasks. Also, due to cost considerations, fine-tuning was limited to a small sample (9000 sentences). We also observed in early trials to deduce the prompt template a huge

variation of results depending on the prompt. Admittedly, if we engaged in exhaustive prompt engineering, the results would be significantly better but deliberately avoided obtaining fair and non-overfitted results.

Furthermore, we present in Figure 6.1 an example of annotating using the best of those three models. The text used is the abstract of a paper on climate change (Amini Parsa et al., 2019), a topic completely different from the topic of ArguminSci corpus (computer graphics). This comparison is shown to highlight the quite evident discrepancies between the three models when compared on the same document. We observe that there is little inter-model agreement on the class of each segment - with an exception on the opening statement that all three models classify as *background claim*. Also, the concluding statement of this abstract is also identified as an argumentation component (only with different class: own claim or background claim). For the rest of the document, however, there is a mismatch not only in the predicted class of each segment but also in the segment boundaries.

The results of this small-scale investigation allude to the comparative inferiority of generative models for text classification tasks when contrasted to discriminative models, like BERT. Consequently, we elect to use the fine-tuned BERT model as our argument extraction unit for the rest of the design and development of our argument recommender.

In contrast, GPT models exhibit superior capabilities in tasks centred around text generation, thus providing the rationale to pursue utilising these models for reforming articles' text to a novel representation that encapsulates the main argument of the paper and is more comprehensible to the end-user. Indeed, in the next section, we propose such a transformation process, purposed to be served as the recommendation unit to the end user, harnessing the power of a Large Language Model - specifically in this case GPT-3.

## 6.2 Recommending Scientific Evidence

During discussions on subjects of high societal impact, argumentation plays a crucial role in shaping opinions and in making personal or collective choices. The web offers a vast amount of argumentative content, however, well-structured argued information is intermixed with content of poor quality (e.g. highly biased, populist, or even fake information).

Scientific literature serves as an excellent source of unbiased and well-reasoned arguments due to its foundation in rigorous research methodologies and quality assurance through peer-review process. It is the distilled output of the scientific method that emphasizes empirical evidence, replication, and reported findings are supported with observations and hard data. As a result, scientific literature fosters a culture of intellectual integrity and critical thinking, making it a reliable source for robust, evidence-based arguments that can withstand scrutiny and contribute meaningfully to ongoing debates and discussions.

However, navigating the vast corpus of scientific literature to find specific arguments can be a daunting task. The sheer volume of research papers, coupled with the specialised terminology often used and complexity of the issues discussed, can make it difficult -if not impossible- to retrieve, assess and reuse relevant arguments.

There have been attempts to develop and evaluate technologies capable of effectively retrieving high-quality and relevant arguments from large-scale corpora. For example, IBM's Project Debater ([Gretz et al., 2020](#)) or Touché shared task ([Bondarenko et al., 2021, 2022](#)) have created large corpora of arguments sourced from the web. Furthermore, specialised search engines have been created for argument search, notably [Wachsmuth et al. \(2017\)](#) created <https://www.args.me/> and [Stab et al. \(2018\)](#) <https://www.argumentsearch.com/>. Those attempts though rely on generic data from public crawls of the web, scraped debate portals or from curated online sources. We present below the scientific argument recommendation task - which can be seen as an extension of argument retrieval task.

### 6.2.1 Scientific Argument Recommendation Task

At a glance, the scientific argument recommendation task can be defined as the task of: given a debate participant position on a given topic, to provide a list of recommended evidence sourced from scientific literature that supports this given position. The ultimate aim of this task is to facilitate a more informed and substantiated exchange of ideas within the debate context, promoting a deeper engagement with empirical evidence and enhancing the overall quality of argumentation.

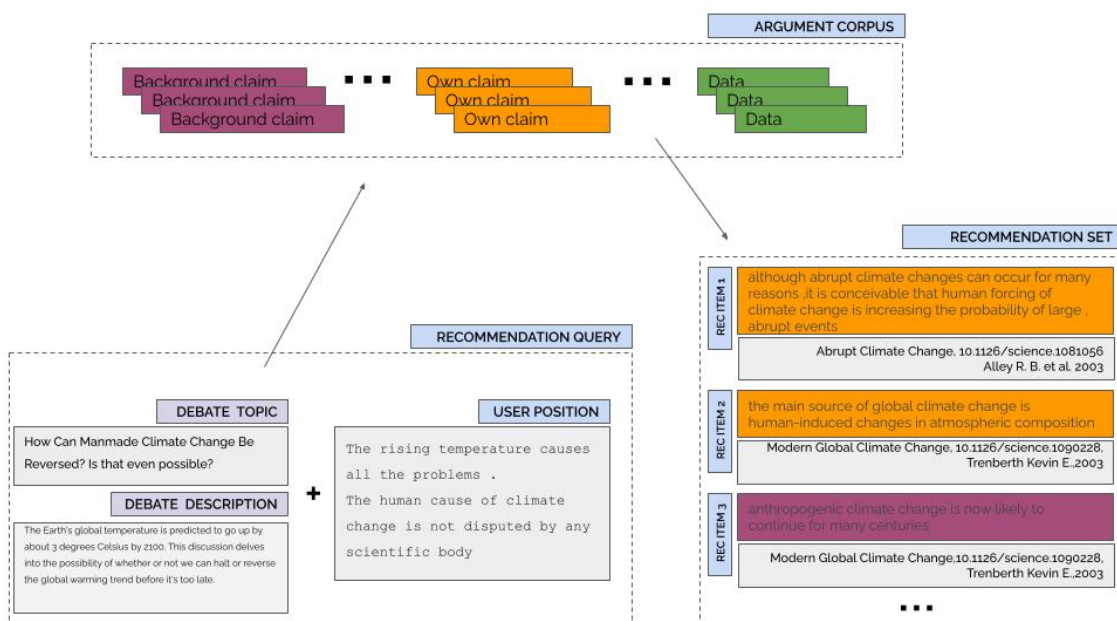


Figure 6.2: Example of recommendation query

### 6.2.2 Content-based Filtering with Embeddings

Content-based filtering with embeddings is a popular recommendation approach that utilizes neural network-based language models to create embeddings that capture semantic information about content items (Zhang et al., 2016a). These embeddings are then used to recommend items with similar content to the user. This approach is especially useful in cases where there is a lack of user data or when there are new items to be recommended. By using embeddings to capture the meaning and similarity of items, content-based filtering can recommend personalized items to



Figure 6.3: Recommendation API response

users based on their preferences.

Semantic search is an information retrieval technique that involves the use of semantic analysis to understand the meaning and context of the search query, and to provide results that match the intent of the user. One approach to semantic search involves the use of embeddings, which are representations of words or phrases in a lower-dimensional space that capture their semantic meaning (Baeza-Yates et al., 2008; Lashkari et al., 2019). By using embeddings, semantic search algorithms can identify documents or information that are semantically similar to the user's query, even if they do not contain the exact same words. This approach allows for more accurate and relevant search results and has been shown to be effective in various applications, including e-commerce, digital libraries, and information retrieval systems.

- Sentence-BERT (sbert) is a popular embedding approach that generates sentence embeddings by fine-tuning a pre-trained BERT model. It has shown promising results in various natural language processing (NLP) tasks, including semantic similarity and text classification. From various models trained using the sentence transformer paradigm we choose *all-mpnet-base-v2* that is an

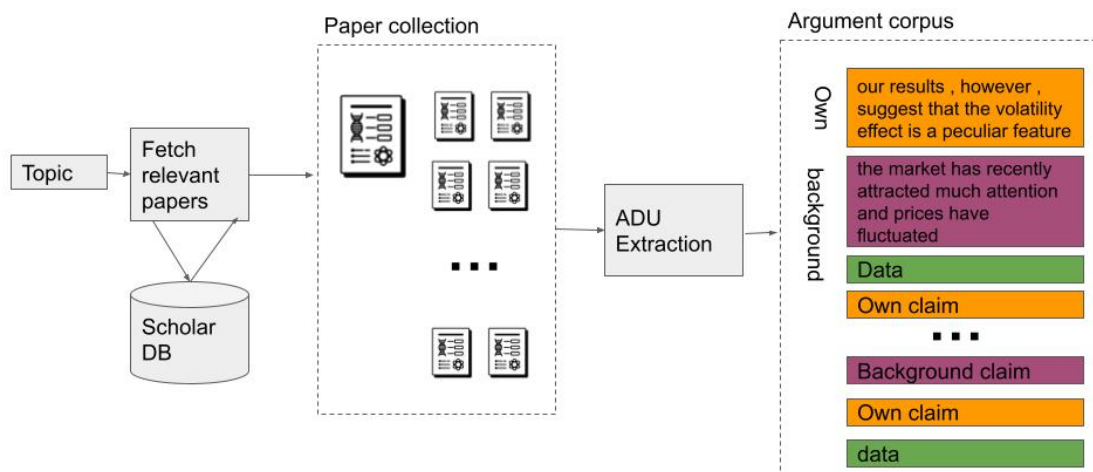


Figure 6.4: High level view of argument corpus creation

all-round model trained on a large and diverse dataset of over 1 billion training pairs.

- SPECTER is another sentence-transformer model that has been trained on a large collection of academic papers (a subset of Semantic Scholar corpus (Ammar et al., 2018)). It has been shown that it outperforms other state-of-the-art embeddings models in scientific domain specific tasks<sup>3</sup>, e.g. citation prediction, document classification and recommendation (Cohan et al., 2020)
- GPT-3 language model can also generate high-quality embeddings for various NLP tasks, including content-based filtering.

We show a brief summary of the embeddings used in Table 6.2.

### 6.2.3 Recommending an Argumentative Summary

We show in Table 6.3, the transformation of the abstract of a random paper (De Lin et al., 2021) to its main argument summary representation.

<sup>3</sup>see SCIDOCs evaluation task - <https://github.com/allenai/scidocs>

Embedding	Max Sequence Length	Dimensions	Size	Training data
all-mpnet-base-v2	384	768	420 MB	1B+ training pairs
SPECTER	512	768	440 MB	146K query papers
GPT3	2048	1024	N/A	45TB text data

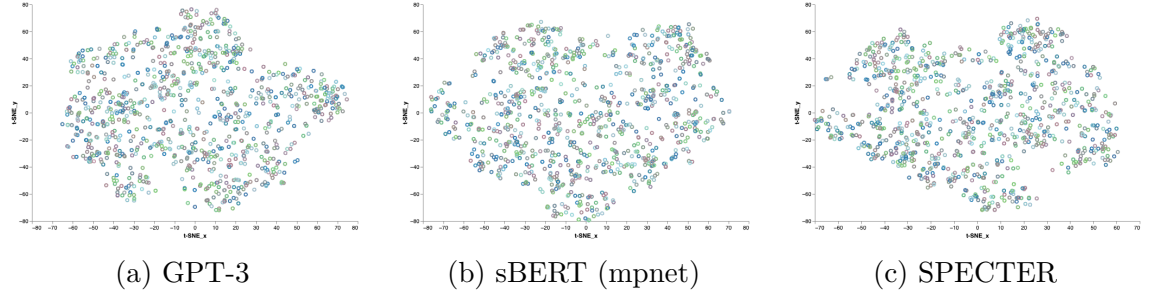
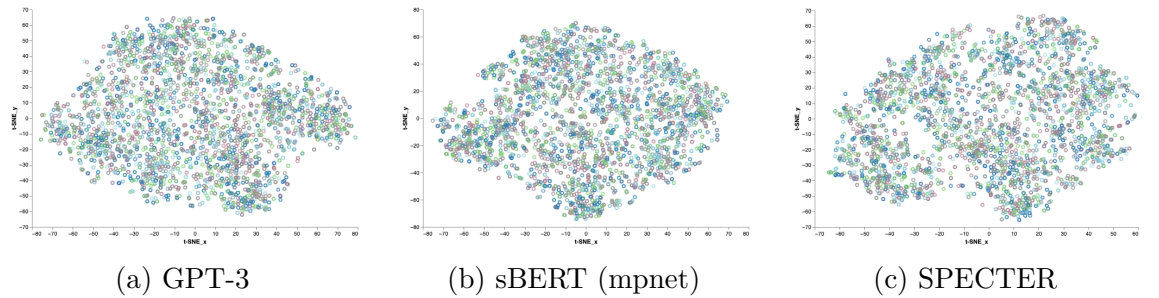
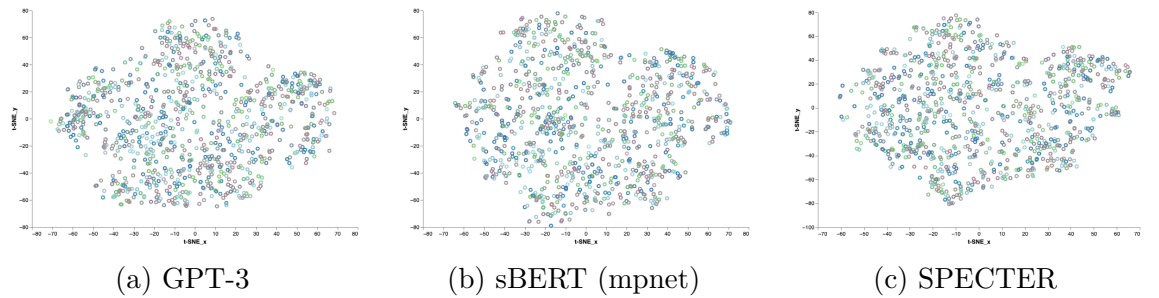
Table 6.2: Details of embeddings used in SciArgRecSys

**Prompt 6** Prompt template - transforming to main argument summary

What is the main argument of the following:

{{ paper\_abstract }}

Include everything that the paper excerpt has to say about the answer. Make sure everything you say is supported by the extract.

Figure 6.5: t-SNE visualisation of embeddings from a 5000 paper sample from  $C_{abs}$ Figure 6.6: t-SNE visualisation of embeddings from a 5000 paper sample from  $C_{arg}$ Figure 6.7: t-SNE visualisation of embeddings from a 5000 paper sample from  $C_{sum}$

Full abstract	Main argument summary
<p>Having students write short self-reflections at the end of each weekly session enables them to reflect on what they have learnt in the session and topics they find challenging. Analysing these self-reflections provides instructors with insights on how to address the missing conceptions and misconceptions of the students and appropriately plan and deliver the next session. Currently, manual methods adopted to analyse these student reflections are time-consuming and tedious. This paper proposes a solution model that uses content mining and NLP techniques to automate the analysis of short self-reflections. We evaluate the solution model by studying its implementation in an undergraduate Information Systems course through a comparison of three different content mining techniques namely LDA-bigrams, GSDMM-bigrams, and Word2Vec based Clustering models. The evaluation involves both qualitative and quantitative methods. The results show that the proposed techniques are useful in discovering insights from the self-reflections, though the performance varied across the three methods. We provide insights into comparisons of the perspectives, which are useful to instructors.</p>	<p>The main argument of the paper is that using content mining and NLP techniques can automate the analysis of short self-reflections, which is currently a manual and time-consuming process. The paper evaluates the effectiveness of three different content mining techniques (LDA-bigrams, GSDMM-bigrams, and Word2Vec based Clustering models) through both qualitative and quantitative methods. The results show that the proposed techniques are useful in discovering insights from the self-reflections, though the performance varied across the three methods.</p>

Table 6.3: Example showing the transformation of an abstract to its main argument summary representation



We list below some metrics to measure the quality of clusters formed by text embeddings, i.e. how well the embedding process has captured the inherent structure and relationship between the items in the corpus. In our case, we used the following metrics to check how well each embedding method preserves the cluster label (in our case label was the id of the debate) after executing K-Means clustering with  $k=5$  (number of debates).

- Purity: introduced by [Zhao and Karypis \(2001\)](#) is a measure of the extent to which clusters contain only data points which are members of a single class. It is calculated by assigning each cluster to the class which is most frequent in the cluster, and then the accuracy of this assignment is measured. A high purity score indicates a good clustering, though it is sensitive to the number of clusters (more clusters typically lead to higher purity).
- Also introduced by [Zhao and Karypis \(2001\)](#), entropy is a measure of the disorder or randomness in the clusters. Lower entropy is better because it means there is less disorder and the data points within each cluster are more similar to each other. In the context of clustering, entropy is calculated for each cluster individually and then a weighted sum is taken to compute the overall entropy of the clustering.
- Homogeneity ([Rosenberg and Hirschberg, 2007](#)) measures whether each cluster contains only members of a single class. A clustering result satisfies homogeneity if all of its clusters contain only data points which are members of a single class. A perfect homogeneity score is 1, which means that the clusters are perfectly homogeneous, i.e., each cluster contains data points from only a single class.
- Completeness ([Rosenberg and Hirschberg, 2007](#)) is a measure that each class contains only members of a single cluster. A clustering result satisfies completeness if all the data points that are members of a given class are elements of the same cluster. Similar to homogeneity, a perfect completeness score is 1, meaning that all data points from the same class belong to the same cluster.

	Corpus	GPT-3	sBERT	SPECTER
homogeneity	abstract	0.15621	0.12676	0.13580
	argument excerpt	0.11261	0.10814	0.09366
	arg. summary	0.14677	0.12628	0.12444
completeness	abstract	0.11049	0.08878	0.09533
	argument excerpt	0.07883	0.07563	0.06563
	arg. summary	0.10343	0.08867	0.08754
entropy	abstract	3.63353	3.70333	3.68362
	argument excerpt	3.72630	3.73572	3.75437
	arg. summary	3.65661	3.69791	3.69656
purity	abstract	0.43640	0.40730	0.41750
	argument excerpt	0.38617	0.39544	0.37626
	arg. summary	0.42878	0.41401	0.39424

Table 6.4: Quality of embeddings used in the three corpora for clustering

We present the comparison of these metrics against sBERT (all-mpnet-base-v2), SPECTER and GPT-3 embeddings in Table 6.4. It is evident that GPT-3 performs the best across most metrics and corpus aspects, demonstrating higher Purity and lower Entropy, indicating well-defined clusters with lower disorder. The homogeneity and completeness scores also suggest that GPT-3 effectively captures the clustering of data points within the same class. In contrast, sBERT and SPECTER generally exhibit slightly lower Purity, higher Entropy, and somewhat lower homogeneity and completeness scores. These results suggest that GPT-3 excels in the context of this specific corpus and clustering task, offering more coherent and accurate representations of the data compared to sBERT and SPECTER. It's essential to consider the trade-offs in sensitivity to the number of clusters when interpreting the Purity results, as more clusters may yield higher Purity scores.

## 6.3 Study IV - Human-centric Evaluation of SciA-rgRecSys

### 6.3.1 Recommender System Human Evaluation Factors

Recommender systems are typically evaluated using one of the following three methods ([Ricci et al., 2010](#)):

- User studies: In this setup, users receive recommendations generated by the evaluated recommendation approach and are asked to explicitly rate the quality of the recommended item in a range of dimensions; therefore deducing the most effective system.
- Offline evaluation measures the performance of a recommender system against a ground truth dataset. Metrics used are typically accuracy, precision, precision at position  $n$  ( $p@n$ ), recall, F-measure, normalized discounted cumulative gain (nDCG) and others.
- Online evaluation method offers a more objective measure as they measure the acceptance rates of recommendations in real-world recommender systems. Typically acceptance rates are measured by click-through rate (CTR), i.e., the ratio of clicked recommendations to displayed recommendations, but can also be the number of downloads, number of buys, etc.

Often, user studies are considered the optimal evaluation method ([Shani and Gunawardana, 2011](#)). However, due to the cost of creating and executing a user study, is not usually considered; but rather offline setups are preferred in development, regardless of the high initial cost of creating a dataset since it allows continuous iteration of proposed methods until you reach considerable levels of performance. Additionally, in [Pu et al. \(2011\)](#), the authors suggest an evaluation framework, named *ResQue* (Recommender systems' quality of user experience), which introduces an evaluation questionnaire comprising sets of questions in four dimensions: (i)

perceived system qualities, (ii) users' beliefs, (iii) users' perceived attitude, and (iv) users' behavioral intentions. For the evaluation of the proposed scientific argument recommender (*SciArgRecSys*) we use an adapted version of *ResQue* that focuses more on the qualitative features of our case and disregarding features that are more appropriate for irrelevant case, e.g. e-commerce. We present below the various dimensions of the ResQue framework:

- *Perceived System Qualities*: These include the *Recommendation Quality* which is the perceived accuracy of the degree to which users feel the recommendations match their interests and preferences. Another factor would be *Novelty* (or discovery) is the extent to which users receive new and interesting recommendations. Also *Attractiveness* refers to whether or not the recommended items are capable of stimulating users' imagination and evoking a positive emotion of interest or desire. Further, *Diversity* measures the diversity level of items in the recommendation list. Moreover, *Context compatibility* evaluates whether or not the recommendations consider general or personal context requirements.
- *Interface Adequacy* is concerned with how to optimize the recommender page layout to achieve the maximum visibility of the recommendation. *Interaction Adequacy*: whether the system allows for user feedback and adapts to specific user's preferences. *Information sufficiency and explicability*: whether the recommendation item sufficiently displays all information needed to help users with making a decision. This in the context of e-commerce site would be price, picture, stock quantity, user reviews and others.
- *Beliefs Perceived Usefulness*: Perceived usefulness of a recommender is the extent to which a user finds that using a recommender system would improve his/her performance, compared with their experiences without the help of a recommender. *Decision support*: thus measures the extent to which users feel assisted by the recommender. In addition to the efficiency of decision-making, the quality of the decision (decision quality) also matters. The quality of a system facilitated decision can be assessed by the confidence criterion, which

is the level of a user's certainty in believing that he/she has made a correct choice with the assistance of a recommender. *Perceived Ease of Use*: measures users' ability to quickly and correctly accomplish tasks with ease and without frustration. *Control and Transparency*: measures if users felt in control of their interaction with the recommender. *Transparency* determines whether or not a system allows users to understand its inner logic, i.e. why a particular item is recommended to them

- *Attitudes*: is a user's overall feeling towards a recommender, which is most likely to be derived from her/his experience as s/he interacts with a recommender. *Overall satisfaction*: determines what users think and feel while using a recommender system. It gives users an opportunity to express their preferences and opinions about a system in a direct way. *Confidence inspiring*: refers to the recommender's ability to convince users of the information or products recommended to them. Finally, *Trust*: indicates whether or not users find the whole system trustworthy.
- *Behavioral Intentions*: whether or not the system is able to influence users' decision to use the system and purchase some of the recommended results.

The above framework includes several irrelevant to scientific argument retrieval dimensions of evaluation, therefore we choose to use a minimal version of it. As in the task of scientific argument retrieval, we need to emphasise in *accuracy*, *novelty*, *diversity*, and *adequacy* dimensions as some of the key objectives of the task, while we drop *attractiveness*, *Interaction Adequacy* and *Control and Transparency* as dimensions that may be important in e-commerce, but add unnecessary complexity and diversion in this specific task. While we aim for a minimal and concise framework, we still consider user experience, therefore the minimal version of the framework still considers dimensions like *satisfaction* and *trust*. From the available pool of possible dimensions listed above, we distil to use the factors shown in Tables 6.5 and 6.6 that best fit our use case.

Table 6.5: Evaluation factors per position item

Code	Factor	Crowdworker prompt
i_f1	Relevance	This argument is relevant to the position given
i_f2	Argument grounding	This argument helps to ground the given position
i_f3	Polarity	The argument takes the same stance on the examined topic as the given position

Table 6.6: Overall evaluation factors of the recommendation set

Code	Factor	Crowdworker prompt
o1	Accuracy	The arguments recommended match the interest of the given position
o2	Novelty	The arguments recommended are novel to me - they helped me discover new knowledge
o3	Diversity	The arguments recommended are diverse (they collect various approaches to the given position)
o4	Adequacy	The arguments are clear and adequate (suitable)
o5	Perceived Usefulness	The recommended arguments are good suggestions to support the given position
o6	Satisfaction	Overall, I am satisfied with this set of recommended arguments
o7	Trust	These recommender arguments can be trusted for their validity

### 6.3.2 Methodology

This study was designed to examine the quality of argument recommendation when evaluated from the human perspective. Specifically, we compare two main methods of querying (TF-IDF and kNN) against a random baseline method, three different types of embeddings for the kNN method and also three different types of recommended unit (arguments as text excerpts, full paper abstract, main argument summary).

We used 5 different datasets from real debates to randomly select a user’s position and create the recommendation query input as the concatenation of the position itself and the debate topic’s description. We used a between-subjects design, i.e. different groups of participants were exposed to different treatments. In our case, different groups were shown different recommendation sets generated by different methods and post-hoc asked to assess the quality of the recommendations according to the selected evaluation factors. In the next subsection, we present the pipeline used to create the corpora and the experimental data used in this study.

#### Pipeline

The pipeline for creating the recommendation set later used for the Mechanical Turk rating study involves several steps:

- $P_1$  - Paper corpus creation: For each debate  $d_i$ , crossref API is queried with the following criteria:
  - query: the debate title  $d_{i,title}$  removing any redundant words
  - filter: has-abstract and is published within a time range  $t_{from} - t_{until}$
  - sort by “is-referenced-by-count” ( $cit - count$ )
  - limit responses to 1000 items

We used two distinct time ranges  $t_{from} - t_{until} = \{1970 - 2010, 2010 - 2022\}$  and then merged the two resulted corpora to a unified (up to 2000 items) big corpus ( $C_{abs}$ )

- $P_2$  - Generate auxiliary corpora: For each item in a given corpus extract arguments using the artefact described in 6.1 ( $C_{arg}$ ) and transform the abstract of the given paper to its main argument summary( $C_{sum}$ ). An illustrative example of this transformation for a single paper is shown in Figure 6.8.
- $P_{3a}$  - Calculate embeddings for each item in corpora  $C_{abs}$ ,  $C_{arg}$ ,  $C_{sum}$ . Embeddings can be either mpnet, SPECTER or GPT-3.
- $P_{3b}$  - Calculate tf-idf features for each item in corpora.
- $P_4$  - Store the auxiliary corpora in a vector database - a database that supports ANN (approximate nearest neighbour) index. Popular options for vector storage would be Pinecone<sup>4</sup>, Spotify's ANNOY<sup>5</sup>, Facebook's FAISS<sup>6</sup> and others; but we chose OpenSearch<sup>7</sup> (AWS version of ElasticSearch<sup>8</sup>) for convenience as we could store both tf-idf features and embeddings in the same database engine.
- $P_5$  - Generate queries pool: Select randomly from each debate 5 positions or arguments and create a pool of queries as the concatenation of debate title and position ( $q_{i,j} = d_{i,title} + p_{i,j}$ ).
- $P_6$  - Apply each method in the query-corpus pair to get candidate recommendation items.
  - Random: Fetch randomly 5 items for the given corpus. The query item in this case plays no role.
  - tf-idf: Transform  $q_{i,j}$  into search terms and execute a boolean term query combining with AND operator
  - kNN: Calculate embedding of  $q_{i,j}$  and execute a similarity search query to fetch the k nearest neighbours of this item to the given corpus.

---

<sup>4</sup><https://www.pinecone.io/>

<sup>5</sup><https://github.com/spotify/annoy>

<sup>6</sup><https://github.com/facebookresearch/faiss>

<sup>7</sup><https://opensearch.org/>

<sup>8</sup><https://www.elastic.co/>



By the end of the pipeline we have tuples of:

$$\langle method_m, corpus_c, d_i, p_{i,j}, rec_1, rec_2, rec_3, rec_4, rec_5 \rangle \quad (6.5)$$

where,

$method_m$  = the method used for querying the corpus, {random, td-idf,  $kNN_e$ }

$kNN_e$  = kNN using embeddings  $e=\{\text{mpnet}, \text{SPECTER}, \text{GPT-3}\}$

$c$  = the type of corpus item, {abs, arg, sum}

$d_i$  = the debate queried

$p_{i,j}$  = the position coming from debate i

$rec_k$  = the recommended item

In total, there were 5 debates x 5 methods (2 + 3 different kNN) x 3 corpus x 5 positions = 375 such tuples. Each tuple was assigned to 3 different annotators, resulting in a total of 1125 ratings.

## Task

Users were presented with a simple user interface where a user's position on a given debated topic was presented. The topic title and short background info were given. A list of recommended items was presented and users were asked to rate (by the use of Likert scales) individually on the three evaluation factors presented in Table 6.5 and overall on the seven evaluation factors presented in Table 6.6. Users were recruited via Amazon Mechanical Turk and offered a compensation of 10\$ per hour (50p for 3 min task average time duration task). Several failsafe checks were employed to ensure faithful responses (as explained in Section 6.3.2) - and failed HIT were resubmitted to new users until a valid response was received. A screenshot of the simple interface of this HIT is shown in Figure 6.9.

**(i) Plain abstract**

Global food demand is increasing rapidly, as are the environmental impacts of agricultural expansion. Here, we project global demand for crop production in 2050 and evaluate the environmental impacts of alternative ways that this demand might be met. We find that per capita demand for crops, when measured as caloric or protein content of all crops combined, has been a similarly increasing function of per capita real income since 1960. This relationship forecasts a 100–110% increase in global crop demand from 2005 to 2050. Quantitative assessments show that the environmental impacts of meeting this demand depend on how global agriculture expands. [...] In contrast, if 2050 crop demand was met by moderate intensification focused on existing croplands of underyielding nations, adaptation and transfer of high-yielding technologies to these croplands, and global technological improvements, our analyses forecast land clearing of only 0.2 billion ha, greenhouse gas emissions of 1 Gt y<sup>-1</sup>, and global N use of 225 Mt y<sup>-1</sup>. Efficient management practices could substantially lower nitrogen use. Attainment of high yields on existing croplands of underyielding nations is of great importance if global crop demand is to be met with minimal environmental impacts.

**(ii) Summarised main argument**

The main argument of the paper is that global food demand is increasing rapidly, as are the environmental impacts of agricultural expansion. The paper forecasts that per capita demand for crops, when measured as caloric or protein content of all crops combined, has been a similarly increasing function of per capita real income since 1960. This relationship forecasts a 100–110% crop demand from 2005 to 2050. Quantitative assessments show that the environmental impacts of meeting this demand depend on how global agriculture expands.

**(iii) Argument snippets**

we find that per capita demand for crops , when measured as caloric or protein content of all crops combined , has been a similarly increasing function of per capita real income since 1960

quantitative assessments show that the environmental impacts of meeting this demand depend on how global agriculture expands

efficient management practices could substantially lower nitrogen use attainment of high yields on existing croplands of underyielding nations is of great importance if global crop demand is to be met with minimal environmental impacts

Figure 6.8: Examples of recommendation units (i) Plain abstract with no process, (ii) Automatic argument-focused summary and (iii) Automatically extracted argument snippets by ArguminSci (argumentation component classes are own claim, background claim and data)

### Evaluating scientific arguments/claims recommendations

This survey examines the quality of scientific claims (arguments) recommended given a position on a debated topic.

#### Task

The **topic** debated is: *How Can Manmade Climate Change Be Reversed? Is that even possible?*

It is described as : CO2 reaches historic levels everyday. Earth's global temperatures are record-breaking each year for the past decade. People react to these shocking statistics by being eco-friendly and going meatless. But is it enough? Have we gone past the point of no return? The Earth's global temperature is predicted to go up by about 3 degrees Celsius by 2100. This discussion delves into the possibility of whether or not we can halt or reverse the global warming trend before it's too late.

A user presents the position: *If we identify the negative feedback loops and break them, then we have a chance of slowing or even reversing climate change*

Below are a set of recommendations of scientific claims sourced from scientific articles for the user to use to make his argument stronger

1. [... we find that per capita demand for crops , when measured as caloric or protein content of all crops combined , has been a similarly increasing function of per capita real income since 1960...]

Sourced as a **own claim**, from Global food demand and the sustainable intensification of agriculture, Tilman David, Balzer Christian, Befort Belinda L., 2011 [View](#)

Please indicate your level of agreement with the following statements:

This argument is relevant to the position given (required)

☐ Strongly disagree ☐ Disagree ☐ Neutral ☐ Agree ☐ Strongly agree

This argument helps to ground the given position (required)

☐ Strongly disagree ☐ Disagree ☐ Neutral ☐ Agree ☐ Strongly agree

The argument takes the same stance on the examined topic as the given position (required)

Figure 6.9: Recommendation rating HIT interface

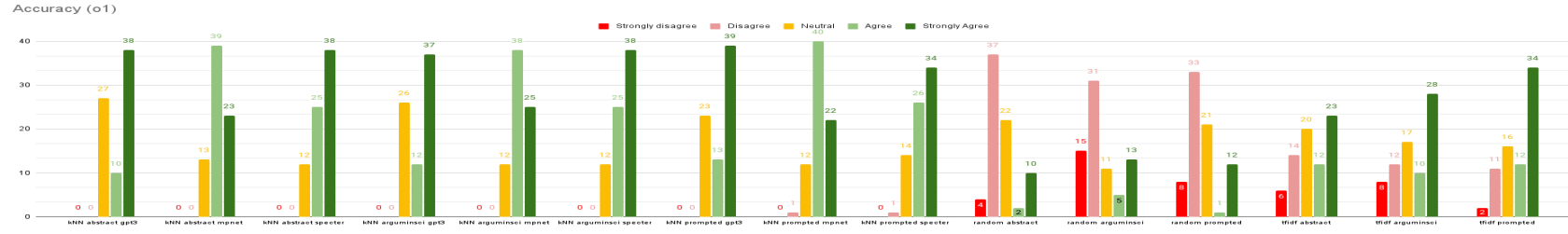
### Validity of Responses

To check the reliability of the questionnaire responses from the MTurk workers we employed several methods:

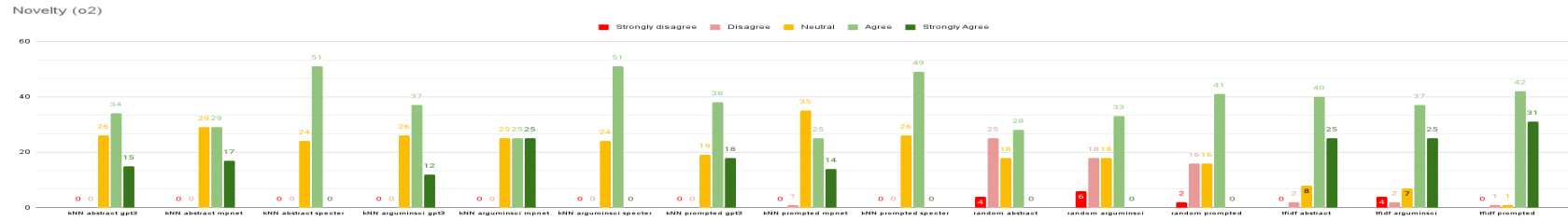
- *Inclusion of attention check items:* We included questions designed to catch respondents who are not paying attention or providing random responses (e.g., “Select ‘strongly agree’ for this item”).
- *Response patterns:* We post-hoc looked for unusual response patterns, such as straight-lining (selecting the same response option for all items) or alternating responses, which may indicate a lack of attention or engagement with the questionnaire.
- *Check internal consistency* reliability which indicates how well the items in a questionnaire measure the same underlying construct. A popular reliability test would be Cronbach’s alpha reliability coefficient, this would be useful if multiple questions were used to measure the same construct.
- We included quick *training* to our participants, providing a quick intro and showcasing an example rating before proceeding to display the actual task.
- Check *inter-rater reliability*. The average Krippendorff’s alpha statistic among the three raters of each tuple (debate\_topic, user\_position, query\_method) of 22 categories (5x3 individual recommendation rating + 7 overall recommendation set rating) was 0.29. This is considered a low level of reliability but could be attributed to various factors, with the most prominent being the high level of subjectivity of each rating and the complexity of the task.
- *Test-retest:* The between-subjects design of the study did not permit the use of other means of examining validity, such as test-retest reliability. Test-retest reliability entails administering the same questionnaire to the same group of respondents at various time points to gauge the level of correlation between their responses.

### 6.3.3 Results

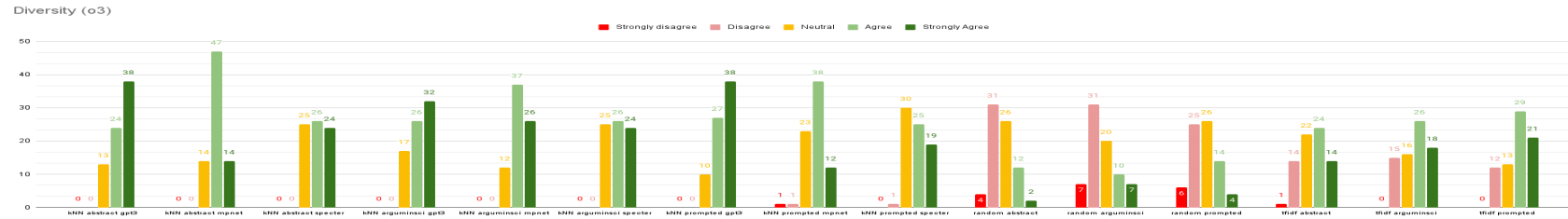
We transformed the 5-scale Likert scale to numerical values by assigning a value from 1 to 5 (Strongly Disagree: 1, Disagree: 2, Neutral: 3, Agree: 4, Strongly Agree: 5). The results of the human evaluation are shown in Tables [6.7](#), [6.8](#) and [6.9](#), where in each cell the mean of 5 debates x 5 positions x 3 annotators is shown. A breakdown of each factor rating for each method is shown in Figures [6.10](#) and [6.10](#).



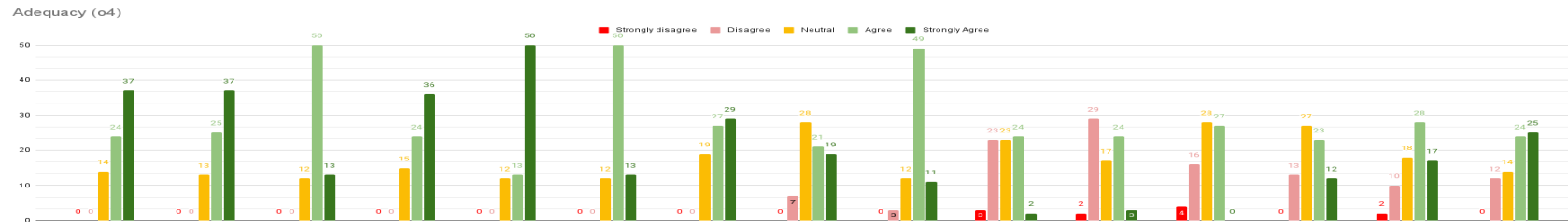
(a) Counts of participants' responses on Accuracy (o1)



(b) Counts of participants' responses on Novelty (o2)

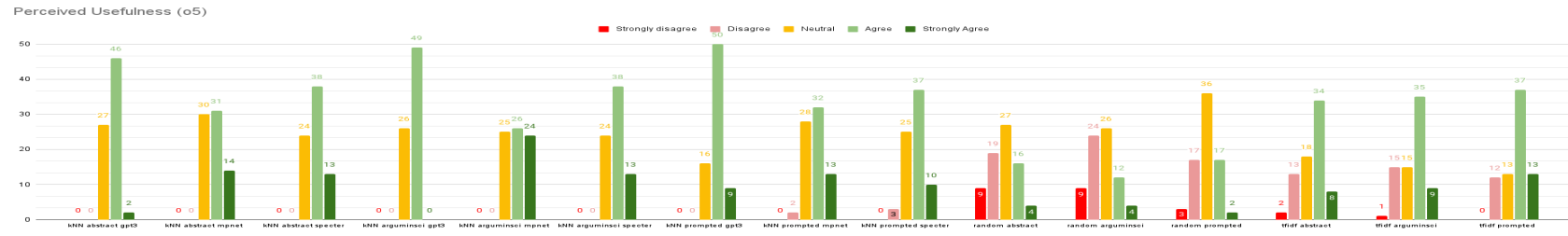


(c) Counts of participants' responses on Diversity (o3)

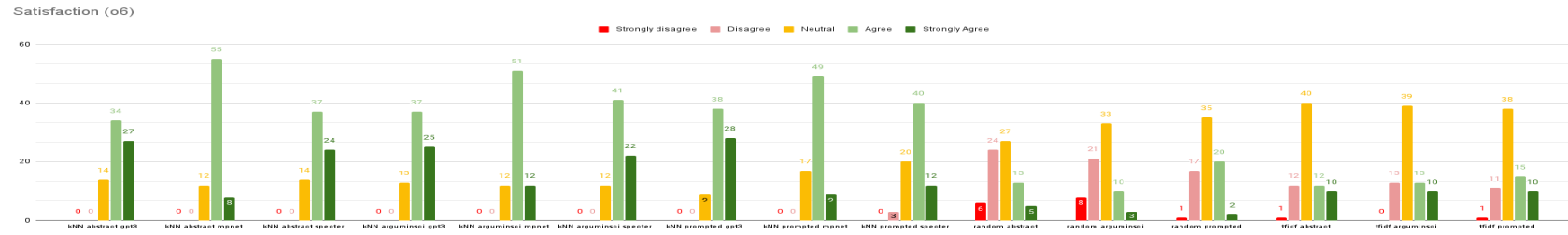


(d) Counts of participants' responses on Adequacy (o4)

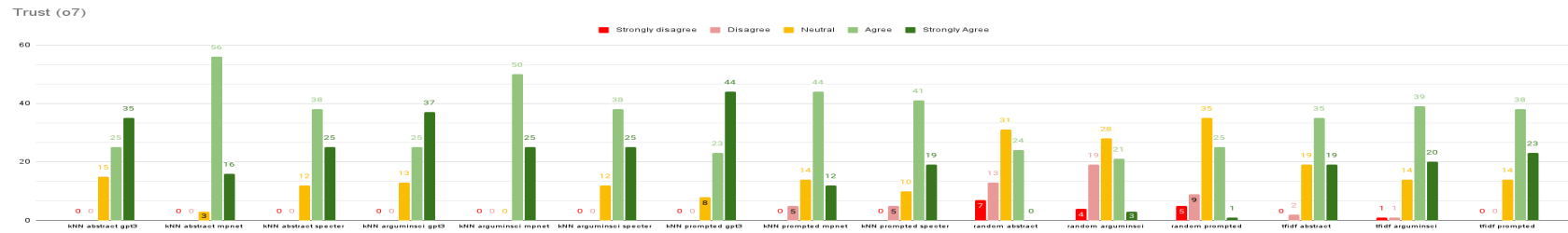
Figure 6.10: Counts of participant responses in Accuracy, Novelty, Diversity and Adequacy



(e) Counts of participants' responses on Perceived Usefulness (o5)



(f) Counts of participants' responses on Satisfaction (o6)



(g) Counts of participants' responses on Trust (o7)

Figure 6.10: Counts of participant responses in Perceived Usefulness, Satisfaction and Trust

Comparing any of the two examined methods to the random baseline, we observe a significant improvement in all seven evaluation variables used; which provides a sanity check that at least the methods used are better than random chance. It is still though interesting to observe that even for the random method - that potentially includes irrelevant recommended items - the preferred recommendation unit was the argument summary (rated higher in all seven variables) and the least preferred method alternates between abstract and argument excerpt (abstract is perceived slightly better for Accuracy (o1), Adequacy (o4), Perceived Usefulness (o5) and Satisfaction (o6) factors).

The same pattern (argument summary to be rated significantly higher and argument excerpts and abstract be approximately the same) is repeated in the other two methods (with very few exceptions). If we cross-compare the tf-idf and kNN methods, we observe that kNN performs significantly better (for any embedding used) in all evaluation factors, with the notable exception of Novelty (o2) factor where tf-idf appears to get rated slightly better (average in recommendation units/embeddings of 3.191 compared to 2.8). This could be explained as tf-idf might be more effective at detecting novel information in text documents, as it prioritises terms that distinguish a document from the overall corpus; while kNN classifies new examples based on the similarity to the given example - it is therefore expected the results not to differ much to each other.

Cross-comparing only the kNN results based on their embedding, we observe that GPT-3 embeddings are performing better in Novelty (o2), Diversity (o3), Adequacy (o4), Perceived Usefulness (o5), Satisfaction (o6), Trust(o7) when it comes to argument summary but only in Novelty (o2), Diversity(o3), Satisfaction(o6), Trust(o7) for abstract and only in Diversity (o3), Satisfaction (o6) when it comes to argument excerpt. We conjecture that GPT-3 embeddings may be more effective at representing text generated by GPT-3 itself compared to other embeddings, like Sentence-BERT (sBERT) because inherently it follows the same internal structure, semantics and syntactic representation of text. In comparison, using embeddings trained on scientific text justifies why SPECTER performs better when used to



	Accuracy	Novelty	Diversity	Adequacy	Perceived Usefulness	Satisfaction	Trust
	o1	o2	o3	o4	o5	o6	o7
Abstract	1.693	1.933	1.693	1.987	1.827	1.827	1.960
Argument excerpt	1.600	2.040	1.720	1.960	1.707	1.720	2.000
Arg/ sum- mary	1.680	2.280	1.800	2.040	1.973	2.067	2.107

Table 6.7: Evaluation of the overall recommendation set using random method

	Accuracy	Novelty	Diversity	Adequacy	Perceived Usefulness	Satisfaction	Trust
	o1	o2	o3	o4	o5	o6	o7
Abstract	2.427	3.173	2.480	2.453	2.440	2.240	2.947
Argument excerpt	2.507	3.027	2.627	2.640	2.480	2.267	3.013
Arg/ sum- mary	2.867	3.373	2.787	2.827	2.680	2.293	3.120

Table 6.8: Evaluation of the overall recommendation set using tf-idf method

represent randomly selected scientific text.

		Accuracy	Novelty	Diversity	Adequacy	Perceived Usefulness	Satisfaction	Trust
Embedding		o1	o2	o3	o4	o5	o6	o7
Abstract	mpnet	3.133	2.840	3.000	3.320	2.787	2.947	3.173
	SPECTER	3.347	2.680	2.987	3.013	2.853	3.133	3.173
	GPT-3	3.147	2.853	3.333	3.307	2.667	3.173	3.267
Argument excerpt	mpnet	3.173	3.000	3.187	3.507	2.987	3.000	3.333
	SPECTER	3.347	2.680	2.987	3.013	2.853	3.133	3.173
	GPT-3	3.147	2.813	3.200	3.280	2.653	3.160	3.320
Argument summary	mpnet	3.107	2.693	2.787	2.693	2.747	2.893	2.840
	SPECTER	3.240	2.653	2.827	2.907	2.720	2.813	2.987
	GPT-3	3.213	2.987	3.373	3.133	2.907	3.253	3.480

Table 6.9: Evaluation of the overall recommendation set using kNN method

<b>Factor</b>		<b>H statistic</b>	<b>p-value</b>
Accuracy	o1	243.15	7.50E-44
Novelty	o2	257.38	8.54E-47
Diversity	o3	277.66	5.30E-51
Adequacy	o4	259.25	3.50E-47
Perceived Usefulness	o5	175.04	6.57E-30
Satisfaction	o6	320.16	7.31E-60
Trust	o7	291.62	6.59E-54

Table 6.10: Kruskal-Wallis statistical test of human evaluation factors

The results of Kruskal-Wallis statistical test are shown in Table 6.10. We omit to present Dunn's test post-hoc examination results as it would be excessive (there are  $13 \times 12 / 2 = 78$  different condition pair combinations for each of the 7 overall evaluation factors).

### 6.3.4 Discussion

We show the comparison between the two predominant techniques used, the term frequency-inverse document frequency (TF-IDF) method and the k-Nearest Neighbors with GPT-3 Embeddings (kNN-GPT3 approach). Although the GPT3 embeddings are arbitrarily selected for this comparison, it is anticipated that this choice has no significant impact on the results, as our primary objective is to qualitatively assess the relative performance of these two methods. Further to this comparison, we test what unit of recommendation (Abstract, Argument excerpt or Argument summary) is more appreciated by humans for the cognitive dimensions discussed in Section 6.3.1. We deduce that recommending abstractive summaries of the main arguments provides better-perceived Accuracy, Adequacy, Diversity, Perceived usefulness, Satisfaction and Trust rather than recommending whole abstracts or argument excerpts.

At the time this thesis is written, a substantial surge in highly prolific and capable Large Language Models (LLMs) is observed, that according to many researchers is indicative of a paradigm shift (a tectonic one!) in search, retrieval and recommendation. Consequently, these advanced models enable recommendation systems developers to shift their attention from the retrieval phase of recommendation to the filtering and reranking phase. For example, we can now include a large number of candidates of documents with their summary (or other generated representations) within a prompt and in a second phase to prompt an LLM (with a very large context window) to get extremely relevant results back. This renewed paradigm looks like in Figure 6.11. This template allows us to define different retrieval nuances, for example, `{{retrieval_nuance}}` could be *relevance*, but also other nuances that are critical for high-quality argument retrieval, such as *quality*, *perspective*, *novelty* and others. Differentiating for such appropriate nuance according to the use case can be easily switched by small alterations to the prompt template.

The prompt for refinement could be something like the prompt template 7. Note that the second stage which is centered around re-ranking, requires a sufficiently long context window to include a substantial number of candidate items. Interestingly

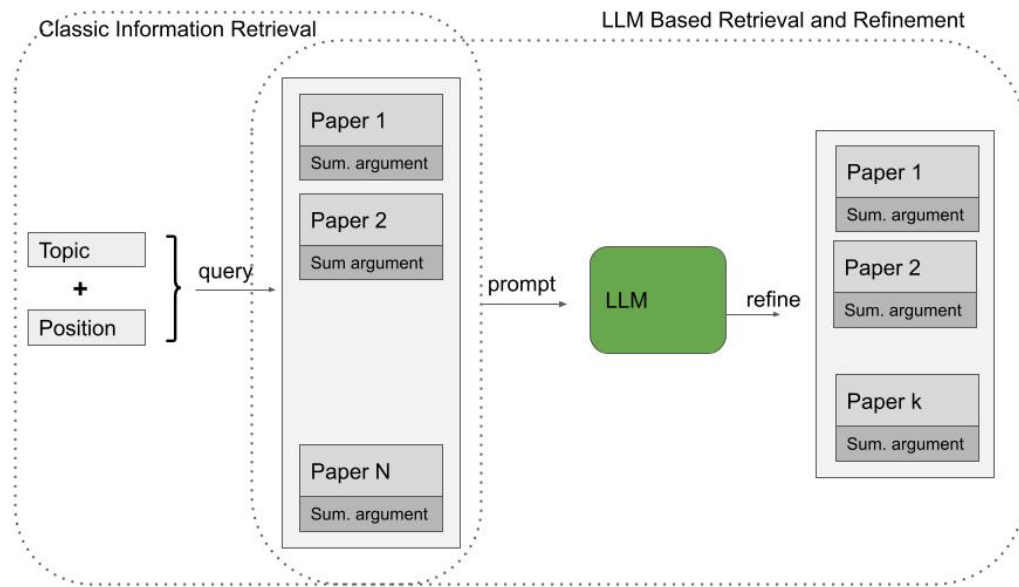


Figure 6.11: LLM based retrieval

though, this task can still be achieved even with a less advanced large language model, with comparable performance.

---

**Prompt 7** Prompt template - LLM-based refinement
 

---

Below there is a list of 100 papers about:

{{ debate\_topic }}

related about the position of:

{{ user\_position }}

Each item in the list contains a document identifier and its main argument summary.

{{ document\_id\_1 }}

{{ document\_summarised\_argument\_1 }}

{{ document\_id\_2 }}

{{ document\_summarised\_argument\_2 }}

...

{{ document\_id\_100 }}

{{ document\_summarised\_argument\_100 }}

Refine the list, keeping only the 5 most {{retrieval\_nuance}} papers that {{ supports/opposes }} the user position

---

## 6.4 Summary

In this chapter, we undertook a comparative analysis of diverse methods pertaining to recommending scientific arguments, with a specific focus on identifying those that demonstrate the greatest efficiency at scale - as the intention is to integrate them into a large-scale online discussion system. We confirmed that discriminative models are better than generative models for argument component identification, though generative models have a significant potential in constructing abstracted summaries of academic paper abstract that convey the article's main argument. We showed that recommending abstracted arguments (transformed summarised via LLM arguments) from the scientific literature is better than recommending argument excerpts or paper abstracts, as they provide better-perceived usefulness, relevance, argumentation, and polarity identification of the argument recommendation. In the next Chapter, we present a holistic approach to integrating this identified artefact and the summariser artefact presented in Chapter 5 into a novel online discussion platform called BCause. The evaluation extends from just the intrinsic quality of the integrated methods but also examines extrinsic factors, such as participants' sensemaking and the dynamics of the discussions. This holistic approach would enable a thorough understanding of the system's overall efficacy and impact on discourse.

# Chapter 7

## Integrating and Assessing Computational Argumentation Artefacts in a Real Online Discussion System

Providing socio-technical support for public interest debates requires a thorough consideration of various principles to ensure healthy and fruitful deliberation. In previous chapters, we demonstrated the potential that abstractive summarisers and recommender systems of scientific arguments have to improve participants' sensemaking and the overall Quality of the Deliberation.

In this chapter, we examine the effect of integrating those two artefacts in a novel online discussion platform, called BCause,

**RQ5:** To what extent does the automated reporting and provision of scientific arguments in combination improve Sensemaking and the quality of the online deliberation process?

**User study V:** Examining effect of automatic synopsis and SciArgRecSys in SM, Eng, MU, Aesthetics, and Social Dynamics in online discussions.

which we present in the following section. We describe an exhaustive evaluation study undertaken to examine an array of variables of deliberation quality such as participants' *Sensemaking*, *Mutual Understanding*, *Aesthetics* and *Engagement*. In addition, beyond the self-reported metrics, we conducted a social network analysis of the interactions observed in our experimental trials. This analysis revealed improved discussion dynamics (e.g., a reduction in subcommunity creation). Finally, we conducted a topic modelling analysis of the content of the deliberations and measured improvements in the consistency and quality of the deliberation content.

## 7.1 Description of BCause Deliberation Platform

### 7.1.1 Design and Rationale

In this dissertation to this point, we have elucidated the primary constraints inherent in group deliberation systems and argued for the necessity of developing platforms promoting evidence-based reasoning and decision-making. This research has been conducted within the context of a larger research and development project at The Open University ([bcause.kmi.open.ac.uk](http://bcause.kmi.open.ac.uk)), which is tasked to develop such a novel platform.

The BCause "Reasoning for Change" platform is a structured and decentralised online discussion system for distributed decision-making. The platform was developed by an interdisciplinary group of designers and technology developers (of which I was part) with the goal of providing structured online discussions for groups to make decisions that are consulted, reflected and critically assessed by all discussion participants. BCause aims to overcome three fundamental limitations of discussion systems when applied to decision-making contexts: i. the lack of overall quality of discussion, particularly in terms of data structure and evidence-based reasoning; ii. the lack of functions that support sensemaking and situational awareness to enable people to participate meaningfully in discussion; and iii. the lack of data ethics, in terms of data centralization, which often implies that organizations share even sensitive



data with centralised proprietary companies in order to gain access to discussion technologies.

With BCause, we have adopted an approach that combines three main innovations:

- i. low-cost argument structuring: with an accessible user interface for users to contribute and analyze arguments in an online discussion process;
  - ii. the distinctive use of automated discourse analysis and advanced visualizations with visual analysis and automated report/summary features to support sensemaking by discussion participants;
  - iii. decentralization: with a data infrastructure that enables secure decentralization of discussion data and user identity, and gives users autonomy of choice and full control over the ownership of personal data.
- The core functionality of BCause is organising the discussion in positions and pro/con arguments (Fig. 7.1).

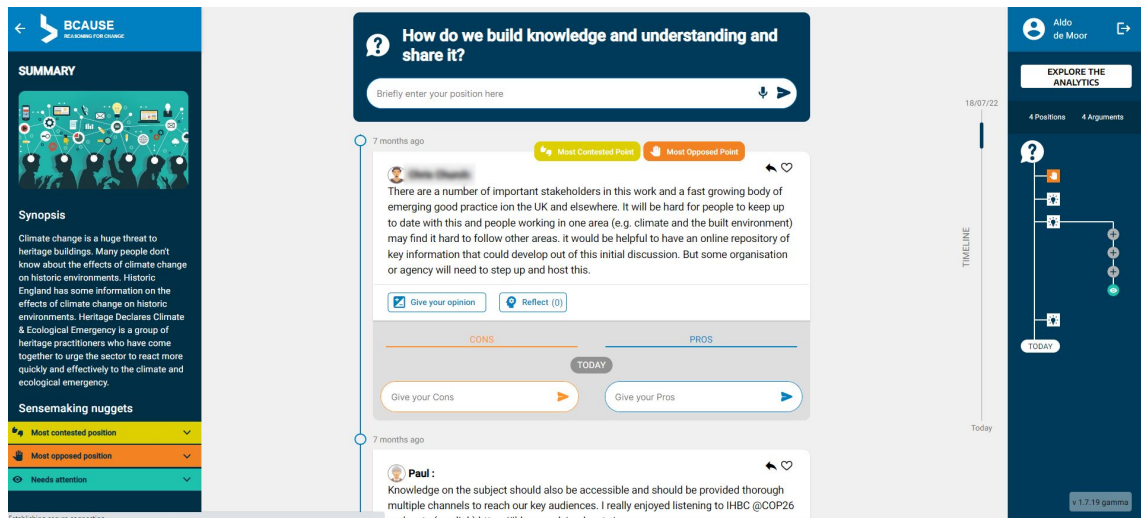


Figure 7.1: BCause discussion interface

In addition to this core discussion functionality, we have incorporated the two computational artefacts described in Chapters 5 and 6 and used this platform as the hosting environment to carry out a semi-naturalistic live evaluation of the two artefacts.

### 7.1.2 Creation of a Low-fidelity Prototype

Our approach was to design a tool that considers its impact on society and individuals, upfront and all along the design process, striving to mitigate the problematic phenomena observed in current deliberation systems. This is in alignment with Value Sensitive Design (VSD) (Friedman et al., 2002) approaches to support human values and promote social justice through technological innovation.

Following a kickoff meeting where we used Q-Methodology adapted for HCI (O’Leary et al., 2013) we set a list of aspirations and fears of our designers and engineers. Having theoretically grounded the values and principles governing the design of our platform, we began the development using iterative design sprints (Banfield et al., 2015). Whereas some of the design aspirations identified, required systemic organizational actions (for example, allowing for different modes of online discussion, e.g. informal, goal-based, adapting Collective decision-making techniques to business/enterprise workflows and procedures, agile and adaptable system to community needs, and others) that goes way beyond system design, other guidelines could be followed by making design choices in terms of UX/UI (for example enabling design processes that allow users to inspect, confirm, dispute and correct past conversations, facilitate transparency, especially in key pieces of information processes, avoid pure argument-centric solutions, employ hybrid interfaces that retain time order and loosely visualize argument structures, and others). This helped to elicit users’ perspectives and finally deduce the following design interventions:

- *Argument-centric structure* of discussion. In BCause, we organise the deliberation data as a tree structure made up of debate topics (issue to be discussed), positions (opinions or possible solutions to the topic imposed), and arguments (statements that support (pro) or oppose (con) the parent position), see Figure 7.2. This follows the well-known paradigm of IBIS systems (see 3.4), and it has many advantages, such as better signal-to-noise ratio, logical structure, implicit encouragement to support with hard evidence, and others, but is not widely adopted as is considered difficult to integrate in scale, is thought to

require skilful information mappers, and enables limited participation. In the design of BCause we therefore chose to combine a linear/time-line dialogical interaction with a structured tree exploration.



Figure 7.2: Argument-centric structure in BCause

- *Agreement* slider: Before entering a pro or con argument, a user is asked to enter his level of support or disagreement to the given position (ranging from “Strongly disagree” to “Strongly agree”, see Figure 7.3. This is a gentle implicit “nudge” to reflect and state his agreement before supporting/refuting it with a concrete argument. In the end, he is shown the collective distribution of the group agreements on this position.
- *Reflection* card: We identified four important reflection dimensions, see Figure 7.4, as features to promote critical thinking in argumentation of online discussion ((Jiménez-Aleixandre and Puig, 2012; Lewiński et al., 2010)):

- *trustworthiness* of the information given in the position (as trust is important requirement for self-reflection as highlighted by [Talboom and Pierson \(2013\)](#) and [Thornton et al. \(2022\)](#))
- whether the position is *polarized*. Identifying polarised positions (or intriguing participants to reflect whether this position is polarized) is important as it can be used as a measure to prevent further division and rigidity in personal beliefs
- whether should be *prioritized*: beyond being useful for the individual sensemaking, this reflection dimension can later be used within a holistic deliberation system (i.e. not limited only to discussion but also include for example ideation, voting, decision-making stages) to rank positions
- *group agreement*: Similarly to the prioritization dimension, a “prediction” of the group agreement can be deemed useful for later stages of deliberation by identifying points of consensus or points of conflict

. In the end, these reflections are visualised in a radial chart along with the community’s average - to provoke a comparison to the “crowd” means. Together with *agreement* slider, the reflection cards are considered nuanced reflective feedback elements (which aim to go beyond appreciative-only feedback as “like”/“thumb up”).

- *Reply* functionality: a reply button enables to directly address a position or argument - without entering an additional position, see Figure 7.5 This helps to incorporate additional semantic information and scope user’s action context.

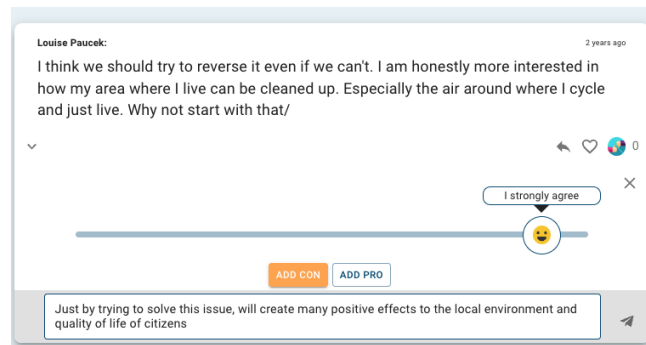


Figure 7.3: Argument input prologued by agreement slider



Figure 7.4: Reflection card two stage interaction

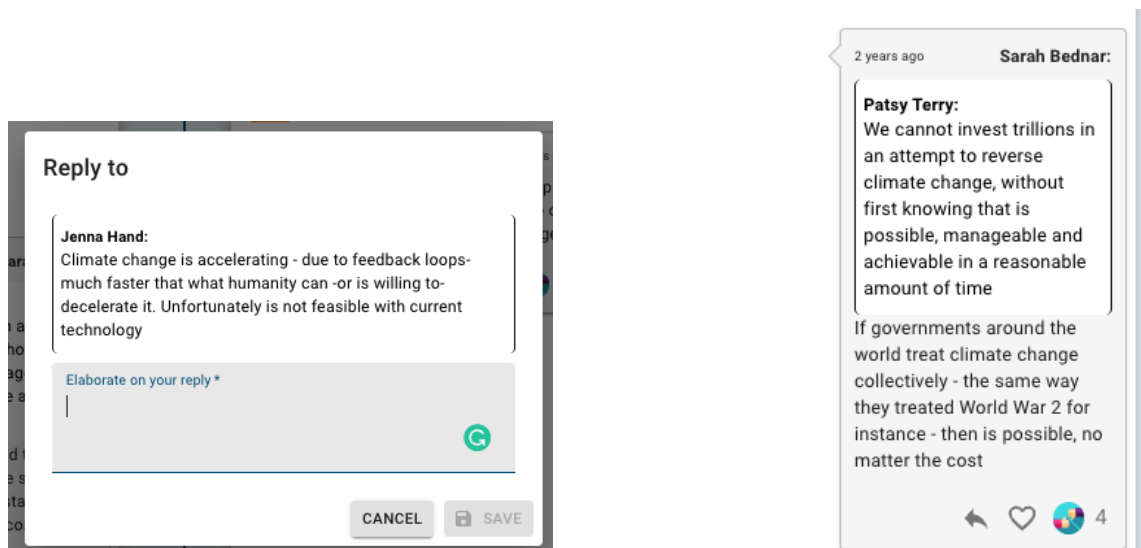


Figure 7.5: Reply dialog box and rendered “quoted” text within argument

## Early Prototypes

In Figures 7.6 and 7.7, we provide preliminary prototypes of the BCause platform, including proposed interventions. Our initial objective was to integrate the agreement slider with the reflective button; however, we concluded that this combination could potentially create user confusion. As such, we ultimately omitted this particular integration in later iterations. In Figure 7.7, Also in the 2nd figure, we introduce the concept of employing mined arguments from other sections of the discussion to be reused or auto-generating arguments that align with the user's current position. However, this idea was later abandoned. At the time, the argument mining components in our disposal exhibited insufficient efficiency, rendering them incapable of supplying accurate or intriguing proposals. Instead, their usage risked to create user confusion and potentially misleading suggestions.

## Sensemaking nuggets

In addition to the elements of reflective feedback (presented in the previous Section 7.1.2), we aimed to incorporate certain subtle prompts, or 'nudges', to intrigue the interest of discussion participants and provide a variety of entry points to approach and explore the discussion. This was considered particularly important for newcomers or people returning to an online debate after a while and when the scale of the debate has grown considerably. We designed 3 nudges with the objective of directing attention towards specific aspects of the conversation that had been overlooked or making transparent what positions or arguments are most contested.


**Most opposed position** The most opposed position is simply the position with the most opposing arguments attached to it, i.e.

$$P_{max-opp} = \arg \max_i (\#args_-(P_i)) \quad (7.1)$$



where,

**The coronavirus is bringing down ill-prepared national economies** 12/08/2013



Think global, act local. Coronavirus might be colonising every corner of the globe, but its economic impact is limited to a select few nations.

The Chinese economy has been affected as it is the epicentre. Economic forecasts suggest that at the very least, the Chinese economy will be hit until Q4 2020. Other nations can use its example to safeguard [.]


In an ideal world perhaps we'd all do the best for everyone else. Unfortunately, that isn't how politics works, nor world leaders operate. The President's only loyalty is to his citizens, and making sure they're protected is his only job.

The US economy was affected by a climate of fear. Fearmongering and uncertainty amongst US investors is causing share prices to plummet at record speeds.

In an ideal world perhaps we'd all do the best for everyone else. Unfortunately, that isn't how politics works, nor world leaders operate. The President's only loyalty is to his citizens, and making sure they're protected is his only job.

+

**YOU ARE 60% SUPPORTING OF THIS OPINION:**

OPPOSES  NEUTRAL  60 SUPPORTS

**YOUR DECLARED CONDITIONS:**

I agree with this only if

- it does affect the human aspect of economy
- it gives the opportunity for green growth and does not contribute towards climate change

**DECLARE YOUR LEVELS OF SUPPORT:**

**CONFIDENCE**

I am fully confident on the validity of the argument

**CONSENSUS**

I believe this argument will be supported by the majority of people

**EVIDENCE-BASED**

The argument is based on hard evidence

**BALANCE**

I have considered a diversity of opinions in this argument

CANCEL ACCEPT

Figure 7.6: Early prototype of BCause agreement slider and reflection card

The interface is a web-based form for the BCause platform. At the top, there is a header bar with a speech bubble icon and a thumbs up icon. Below this is a large orange circle with a white plus sign. The main content area is a light beige background. It features a progress bar at the top indicating 'YOU ARE 60% SUPPORTING OF THIS OPINION:' with a purple circle at 60% and the word 'NEUTRAL' in the center. Below the progress bar, there is a section titled 'YOUR DECLARED CONDITIONS:' with a text input field containing 'I agree with this only if' and a list of conditions: 'it does affect the human aspect of economy', 'it gives the opportunity for green growth', and 'and does not contribute towards climate change'. Below this, there are two columns of argument cards. The left column is titled 'REUSE SOME OF ARGUMENTS' and the right column is titled 'AUTO-GENERATED ARGUMENTS'. Both columns contain several cards with text like 'In an ideal world perhaps we'd all do the best for everyone else. Unfortunately, that isn't how politics works, nor world leaders operate. The President's only loyalty is to his citizens, and making sure they're protected is his only job.' Below the argument cards is a section titled 'ELABORATE ON YOUR ARGUMENT' with a large text input field and 'CANCEL' and 'ACCEPT' buttons. At the bottom, there is a section titled 'YOUR LEVELS OF SUPPORT:' with a central graphic of a head with four arrows pointing outwards to four categories: 'CONFIDENCE' (I am fully confident on the validity of the argument), 'BALANCE' (I have considered a diversity of opinions in this argument), 'EVIDENCE-BASED' (The argument is based on hard evidence), and 'CONSENSUS' (I believe this argument will be supported by the majority of people).

YOU ARE 60% SUPPORTING OF THIS OPINION:

OPPOSES NEUTRAL 60 SUPPORTS

YOUR DECLARED CONDITIONS:

I agree with this only if

- it does affect the human aspect of economy
- it gives the opportunity for green growth
- and does not contribute towards climate change

REUSE SOME OF ARGUMENTS

AUTO-GENERATED ARGUMENTS

ELABORATE ON YOUR ARGUMENT

Input here ...

CANCEL ACCEPT

YOUR LEVELS OF SUPPORT:

CONFIDENCE

I am fully confident on the validity of the argument

CONSENSUS

I believe this argument will be supported by the majority of people

BALANCE

I have considered a diversity of opinions in this argument

EVIDENCE-BASED

The argument is based on hard evidence

Figure 7.7: Early prototype of BCause auto-generated arguments with reflection card



$p_i$  = is the  $i_{th}$  position

$\#args_-$  = is the number of con arguments

Despite its simplicity, this metric is significant, as it swiftly pinpoints that most contentious standpoint. This with conjunction with the next most contested position metric is crucial for enabling participation in contested positions.

**Most controversial position** For finding the most controversial position (which delves into the contested nature of collective intelligence (De Liddo et al., 2012)), we use a mix of the number of opposing arguments, the number of pro arguments and the list of agreements to that position. This intermix provides a richer representation of what is the contested score.

$$Contested - score = [(1 + \#args_-)(1 + \#args_+)]^{1+\sigma(z)/200} \quad (7.2)$$

where,

$\#args_-$  = Number of opposing arguments

$\#args_+$  = Number of supporting arguments

$z = \{z_1, z_2, \dots, z_n\}$  = agreement score list,  $z_i \in (-100, 100)$

$\sigma(z)$  = the standard deviation of the agreement scores

Therefore the most controversial position would simply be the position with the highest contested score, i.e.

$$P_{most-controversial} = \arg \max_i (contested - score(P_i)) \quad (7.3)$$

We provide a list of typical synthetic examples to demonstrate the rationale of the above formulas, in Table 7.1:

$\#args_+$	$\#args_-$	$z$	$\sigma(z)$	$Contested - score$
0	0		0	1
5	5		0	36
5	5	10,-50,50,-10,75,-75	57.36	100.61
5	5	70,80,90,100	12.91	45.37
5	5	-70, -80, -90, -100	12.91	45.37
9	1		0	20
9	1	10,-50,20,-65	42.5	37.80
9	1	-80,-75,-95,-85	8.54	22.73

Table 7.1: Contested score example values

We observe that if for example, we have an equal number of pro and con arguments, a higher variance of the agreements provided by the users will result in a much higher contested score. Also, as in the Table’s last row, even if there is a big number of pro arguments (which would indicate a small contested score) but lots of agreement disparity at the same time, as this disparity is a signal of contestation, there is a high contested score.

**Needs attention** Social score is simply the count of social interactions (given thanks to author, number of agreements, number of reflections) of a position. Note that we count all social interactions positively, for example, agreement levels of -100% (pure disagreement) would still count as +1 in social score.

$$social - score = \#thanks + \#agreements + \#reflections \quad (7.4)$$

We compare this synthetic score to a predicted score of argumentation, inferred by an argumentation transformer model <sup>1</sup> Therefore the *needs attention* score is simply the difference of social score ( $social - score$ ) and the argumentation score ( $arg - score$ ) normalised, since  $social - score \in [0 - \infty)$  and  $arg - score \in [0.0 - 1.0]$

$$needs - attention - score = 100 \times \frac{arg - score}{1 + social - score} \quad (7.5)$$

<sup>1</sup><https://huggingface.co/chkla/roberta-argument>, finetuned RoBERTa model for classifying sentences to ARGUMENT (1) or NON-ARGUMENT (0)

## 7.2 UX Prototype Testing

We present below two small-scale pilot studies that were used for evaluation of the user interface (UI) and user experience (UX) of the Synoptical summariser and SciArgRecSys artefacts when deployed to BCause platform. The findings from both investigations helped us to make essential modifications to the artefacts to enhance their user experience (UX).

### 7.2.1 Assessing Synopsis Creation

Within the context of a lecture on media literacy at the University of Milan (POLIMI), an early prototype of BCause platform was used to debate on the topic “Do you think removing the comment space from newspapers websites is a fair choice to reduce hate speech?”. In this early prototype, we had only the summariser artefact deployed. 22 students contributed 76 posts in a course of  $\sim 30$  min of allocated time. In a short post-hoc examination, students provided feedback on the user experience of debating an issue with the use of the BCause deliberation platform. Also, a short questionnaire was given after the completion of the lecture (with a low response rate - 5 out of 22, 23%) containing open-ended questions regarding the experience. This small-scale experiment helped to identify issues and establish a workflow for the integration of the summariser artefact. An issue reported by participants was frequent regeneration and update of the summary presented, as exemplified by U1:

The automated summary kept on glitching and then after a few minutes when it actually stuck to one answer it did bring out a quick summary of the whole discussion.

In response to the identified issue, a protocol was established wherein the generation of a summary occurs upon the addition of considerable new content, or the lapse of a specified time duration. As such, the invocation of the summarisation function was not happening on the creation of any new post, but rather following the submission of three (3) new posts or five minutes since the last entry - whatever occurred first.

Also, the length of the summary was identified as an issue, proposing a shortened length or a different format. As expressed by U3:

Maybe the summary part can be sectioned for a better understanding.

In response to this issue, we altered the desired length of the summary to 128 words or only 10% of the discussion (whatever is bigger) and changed the prompt for the summary to be more comprehensive.

Moreover, we observed an early confirmation of the value of such artefact deployed in a debate:

It was interesting, especially because of the possibility of seeing the sum-up of all the opinions and how you can engage in a conversation.

Further to design alterations, the debate data was used in the major evaluation study (Section 7.3) as a starting point for each trial and in the summariser evaluation study reported in Section 5.2.2.

### 7.2.2 Assessing SciArgRecSys User Experience

Two different configurations of BCause were used in the context of teaching mechanical engineering courses at University of Napoli (UNINA). One configuration had the *SciArgRecSys* enabled and the other did not. Students were split into two even groups assigned randomly. Students were given an initial training to familiarise themselves with the platform by posting on a debate about “What measures and actions should be taken to try to counter climate change and its increasingly evident and recurring consequences?”. They were then split into two groups (A with recommender, B without) and discussed the topic of “What are the economic impacts of the war in Ukraine on our economic-productive fabric?”. The task took place during a lecture with total time allocated for interaction to the platform approximately 20 minutes. The SciArgRecSys was configured to recommend argument excerpts, see Figure 7.8. This resulted in long discussions of 119 and 117 positions and arguments, see table 7.2. A special configuration of the recommender was used to handle the non-English

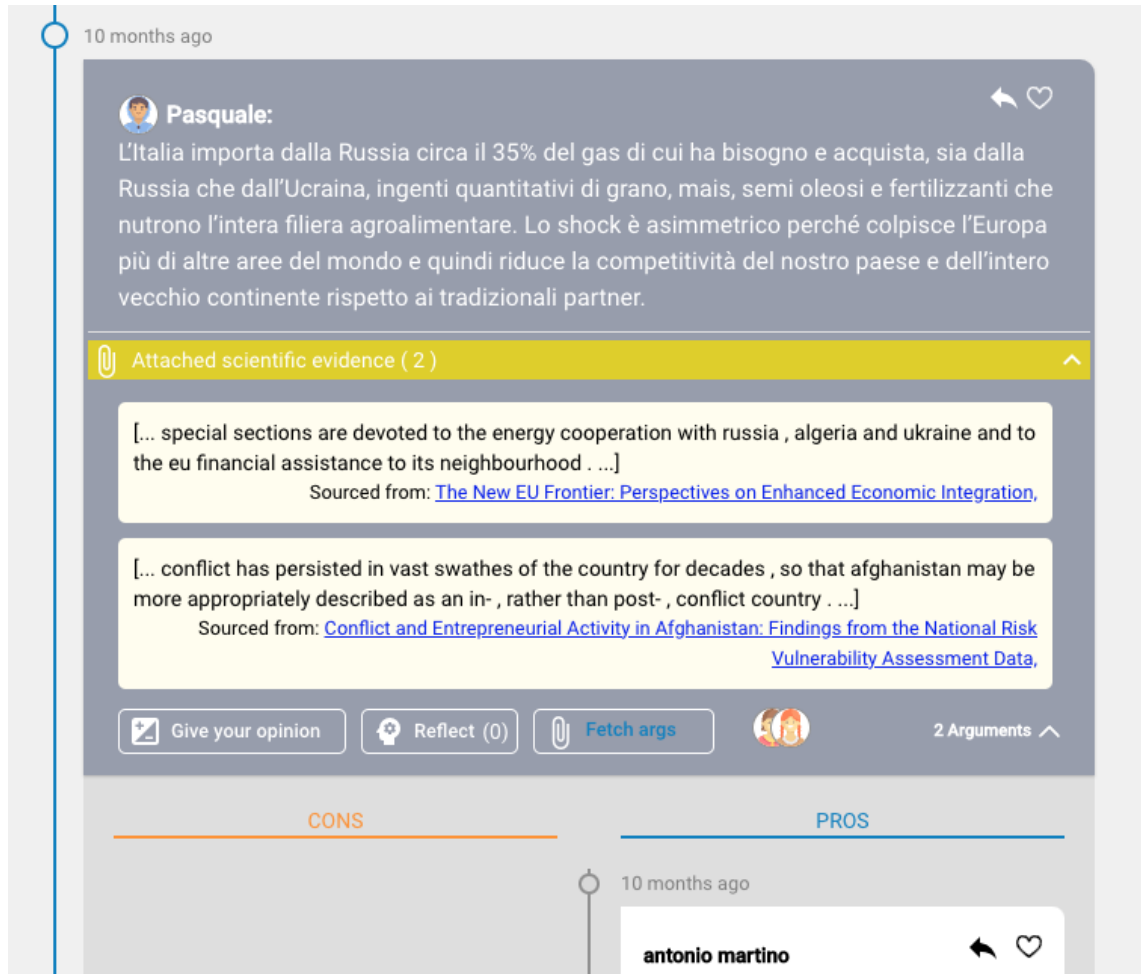


Figure 7.8: SciArgRecSys instantiation in UNINA use case

language (Italian) used in the debate. As such recommender underlying corpora or embeddings used are not multilingual, so an intermediate translator of the user's position to English was used.

Group	# Pos	# $Arg_+$	# $Arg_-$	# $Ev_{pos}$	$Ev_{arg}$	words	$\overline{words}$	# Participants
A	65	25	31	5	13	13133	108.5	16
B	50	71	20	-	-	11191	79.4	12

Table 7.2: Descriptive statistics of the debates formed in UNINA use case

We observed a relatively low use of attaching evidence function (18 pieces of evidence attached in a total of 121 posts, approximately one piece of evidence attached per 7 posts). An extensive follow-up questionnaire comprising of quality of discussion

(QoD), sensemaking (SM), aesthetics (A), mutual understanding (MU), quality of collaboration (QoC), engagement and usability (U) questions (see table 7.3) was given to broadly assess users' experience with the platform. A reduced version of this extended questionnaire was used later for the major evaluation study reported in Section 7.3.

Factor	$\mu_A$	$\sigma_A$	$\mu_B$	$\sigma_B$	stat	p
QoD1	4.667	0.356	4.467	0.249	138.5	0.217
QoD2	4.667	0.222	4.200	0.293	160	0.027
QoD3	4.067	0.596	3.533	0.516	151.5	0.088
QoD4	4.733	0.196	4.467	0.249	142.5	0.150
QoD5	4.533	0.382	4.333	0.222	137.5	0.248
QoD6	4.800	0.160	4.267	0.729	153	0.050
QoD7	4.800	0.160	4.400	0.373	151.5	0.058
QoD8	4.333	0.622	3.800	0.427	155.5	0.060
QoD9	4.200	0.427	3.067	0.862	184	<b>0.002</b> *
QoD10	4.267	0.596	3.733	0.862	148	0.125
SM1	4.733	0.196	4.600	0.240	127.5	0.462
SM2	3.733	1.129	3.800	0.827	114.5	0.948
SM3	1.733	0.596	2.267	0.862	75	0.095
SM4	3.933	0.729	3.467	0.516	149	0.111
SM5	1.733	1.396	1.867	0.649	88	0.276
SM6	4.467	0.382	4.133	0.382	144	0.153
SM7	4.600	0.240	4.133	0.382	156	<b>0.045</b> *
SM8	1.400	0.240	1.733	0.462	84	0.193
SM9	2.733	0.996	3.000	0.400	102	0.644
QoC1	4.600	0.373	4.200	0.827	139	0.225
QoC2	4.933	0.062	4.533	0.249	157.5	<b>0.016</b> *
QoC3	4.800	0.160	4.200	0.427	168	0.009
QoC4	1.733	0.729	2.267	0.596	74.5	0.098
QoC5	4.533	0.516	4.467	0.516	119	0.771
QoC6	4.733	0.196	4.467	0.382	137	0.237
QoC7	4.733	0.196	4.267	0.462	154	0.053
QoC8	4.933	0.062	4.800	0.160	127.5	0.307
MU1	4.333	0.222	3.933	0.996	135	0.294
MU2	4.533	0.382	4.133	0.249	154.5	0.052
MU3	3.867	0.516	3.800	0.560	118.5	0.806
MU4	4.533	0.382	3.800	0.427	173.5	<b>0.007</b> *
MU5	4.533	0.382	4.000	0.400	160.5	<b>0.031</b> *
MU6	2.133	0.782	2.067	0.729	117.5	0.843
A1	3.867	0.916	3.533	0.516	146	0.132

A2	3.867	0.649	3.333	0.622	153	0.077
A3	3.933	0.596	3.600	0.507	141	0.195
A4	2.800	0.693	2.933	1.129	101	0.633
A5	3.933	0.196	3.667	0.222	140	0.146
A6	2.600	0.773	2.400	0.907	126.5	0.556
A7	3.533	1.049	3.667	0.889	103	0.691
A8	3.467	1.049	3.467	1.316	111.5	0.983
A9	4.067	0.329	3.733	0.729	133	0.335
A10	3.467	0.916	3.533	0.649	108	0.859
E1	3.867	0.649	3.333	0.489	153	0.071
E2	2.867	0.782	2.333	0.889	147.5	0.135
E3	2.067	0.729	2.267	0.996	103	0.687
E4	1.200	0.160	1.533	0.649	88.5	0.223
E5	1.467	0.649	1.667	0.756	96.5	0.460
U1	4.133	0.249	3.800	0.427	143.5	0.140
U2	2.067	0.729	2.267	1.662	109	0.897
U3	4.000	0.533	3.733	0.596	131	0.418
U4	1.867	1.049	1.533	0.516	129	0.459
U5	3.533	0.782	3.533	1.049	112	1.000
U6	1.867	0.649	2.133	1.316	100.5	0.614
U7	3.800	1.360	4.000	0.800	105.5	0.777
U8	2.467	1.316	2.733	1.262	94.5	0.453
U9	3.800	1.227	3.733	0.462	128.5	0.491
U10	2.533	1.449	2.067	0.729	134.5	0.347

Table 7.3: Mann-Whitney U test between group A (with recommender) and group B (without recommender) of UNINA use case. Quality of Discussion (QoD), Sensemaking (SM), Quality of Collaboration (QoC), Mutual Understanding (MU), Aesthetics (A), Engagement (E) and Usability (U) factors are examined

We observe statistically significant results ( $p < \alpha = 0.05$ ) in 7 out of 58 evaluation factors. In particular, we observe SciArgRecSys group demonstrate better results in QoD2 (Reasoning), QoD9 (Polarisation), SM7 (Distinguish facts), QoC2 (Interesting discussion), QoC3 (Problem solving), MU4 (Mutual understanding), MU5 (Shared understanding). This can be attributed to the introduction of the argumentative aspect of this artefact. It should be noted that the above factors are intrinsically tightly coupled with argumentation, thereby providing a plausible explanation of the observed improvement in these dimensions. The absence of any degradation in Aesthetics or Engagement levels (and in any of the examined variables) serves as a

### 7.3. Study V - Assessing the Combined Impact of Automated Reporting and Scientific Arguments Recommendation

promising indication that integrating the SciArgRecSys artefact did not overwhelm or burden the users. In contrast, the noticeable enhancement in some variables, suggests a more enriched user experience. Contrary to our initial expectations, we did not observe significant improvements in Sensemaking, with the exception of SM7, or in the quality of discussion, apart from QoD9. This outcome prompted us to modify the argument recommender to deliver a different recommendation unit (argument summary instead of argument excerpt as it was configured for this pilot study). This adjustment was influenced by the results obtained in Chapter 6, with the intention of enhancing the aforementioned factors.

## **7.3 Study V - Assessing the Combined Impact of Automated Reporting and Scientific Arguments Recommendation with a Semi-Naturalistic Online Deliberation Experiment**

### **7.3.1 Research Question**

The research question investigated in this study is RQ5 (1.3.2): *To what extent does the automated reporting and provision of scientific arguments in combination improve Sensemaking and the Quality of the Online Deliberation process?*

The quality of the online deliberation process consists of a complex array of variables which encompass things such as participants' perception, their actions and the results of these actions on the deliberation output (the online deliberation corpus). To strengthen the robustness of our investigation we looked at all these three aspects. In the study we focus on *Mutual Understanding*, *Engagement*, *Aesthetic*, as it will be explained in Section 7.3.2, to assess participants' perceptions of deliberation quality. We then looked at the social network created by participants' communicative interactions on the platform to assess Social Dynamics. Finally, we carried out



topic modelling to assess the online deliberation corpus generated by the online deliberating process. We, therefore, followed three distinct data analysis methods for addressing our RQ: (i) group comparison and statistical test of the user-reported perception of Sensemaking (SM), Engagement (E), Mutual Understanding (MU) and Aesthetics (A), (ii) Social Network Analysis for uncovering social dynamics resulting from the deliberation process and (iii) Topic modelling for exploring quality of the generated discussion corpus.

### 7.3.2 Experiment Design

To address the research question, we chose a combination of a Randomized Controlled Trial (RCT) with a 2x2 factorial design. A 2x2 factorial design is a type of experimental design that allows researchers to study the effects of two independent variables simultaneously, along with any interaction effects between them (Telford, 2007). In our case, the two independent variables are the presence (or not) of the two artefacts (*Summariser*, described in Chapter 5 and *SciArgRecSys*, described in Chapter 6). The target variables are Sensemaking, Engagement, Aesthetics, Mutual Understanding.

		SciArgRecSys	
		No	Present
Summariser	No	Condition A	Condition C
	Present	Condition B	Condition D

Table 7.4: 2x2 Factorial experimental design

#### Conditions

Our controlled experiment had the following four conditions:

- *Condition A (alpha)*: this design variation contains a stripped-down version of the BCause interface (baseline), which does not contain any of the synopsis, sensemaking nuggets or SciArgRecSys artefacts. It does, however, contain the

### 7.3. Study V - Assessing the Combined Impact of Automated Reporting and Scientific Arguments Recon

---

feedback elements described in Section 7.1.2, i.e. agreement slider, reflection button, along with a “thank you” button). Discussion is organised in an argumentative fashion following a light IBIS approach (see 3.4), i.e. posts are organised in positions and supporting or opposing arguments. The interface of this condition is shown in Figure 7.9a.

- *Condition B (beta)*: This condition is an extension of *Condition A*, with the Synopsis artefact and Sensemaking nuggets appearing on the left column of the platform. The synopsis artefact is configured to generate a fresh summary every 3 newly added posts or after 5 minutes of the last inserted post (whatever occurs first). The sensemaking nuggets selected are: (i) the most controversial position, (ii) the most opposed position and (iii) needs attention. The interface of this condition is shown in Figure 7.9b with the Synopsis and Sensemaking nuggets annotated.
- *Condition C (gamma)* This condition is an extension of *Condition A* with the SciArgRecSys artefact enabled in the following fashion. A “fetch args” button appears in every position and argument. Upon clicking, the relevant scientific evidence list is shown in a dialogue box, where the user is prompted to select which they wish to attach to the given node as evidence. Upon selecting those, the attached evidence appears in a toggle panel below the position or argument. In Figure 7.9c we show the interface of Condition C along with explicit annotation of where the design elements appear.
- *Condition D (delta)* This condition is the fully-fledged BCause application: an extension of *Condition A* containing all artefacts (Synopsis, sensemaking nuggets and SciArgRecSys) in the same configuration as conditions B and C. The interface of this Condition is show in Figure 7.9d.

**BCAUSE**  
REASONING FOR CHANGE

**ABOUT**

**Description**

Some online newspapers decided to remove the comment space from their platform to reduce hate speech phenomenon. Do you think this is a fair choice in a democratic society? What are the pros and cons of removing the comment space from newspapers websites? (Is it a fair choice? Does it reduce polarisation? Does it infringe free-speech?)

**How can we fairly reduce hate speech from newspapers websites without eliminating space for readers comments?**

Briefly enter your position here

mindfulness, hate speech could be removed

Give your opinion Reflect (3)

9 months ago

**AndreeP:**

In order to reduce hate speeches there should not be any comment section. Newspapers should be one sided providers.

Give your opinion Reflect (2)

5 Arguments

**CONS**

9 months ago

**FrancescoT**

If that was the case, how would you go about allowing commenting on newspaper and journalist points (which are often really problematic) but very hard to counteract?

Reflect

9 months ago

**AndreeP**

If that was the case, how would you go about allowing commenting on newspaper and journalist points (which are often really problematic) but very hard to counteract?

**AndreeP**

Maybe it would trigger a movement to investigate

**EXPLORE THE ANALYTICS**

38 Positions 56 Arguments

**SUMMARY**

**Synopsis**

This discussion is about whether or not reducing freedom of speech on public articles is an effective way of avoiding hate speech. Based on the article, it is essential for humans to have the freedom to express their opinions. Free speech gives a voice to many groups of people to promote positive change. It is difficult to find enough moderators to review all of the comments on a website, and that some comments would inevitably slip through the cracks. Additionally, some people might abuse the flagging system by flagging comments they simply don't agree with, rather than those that are actually hateful. Any attempts to reduce hate speech from newspaper websites will be ineffective, and that the only way to eliminate it is to eliminate space for readers comments altogether.

**Sensemaking nuggets**

**Most contested position**

Reducing the freedom of speech on public articles is not the solution for avoiding hate speech. The way of using an algorithm which can observe and in case it is necessary remove / mark hate speech is a better choice

**Pro/Con argument summary**

The article argues that reducing the freedom of speech on public articles is not the solution for avoiding hate speech, and that a better solution is to use an

**BCAUSE**  
REASONING FOR CHANGE

**How can we fairly reduce hate speech from newspapers websites without eliminating space for readers comments?**

Briefly enter your position here

6 months ago

**PatriziaB:**

Reducing the freedom of speech on public articles is not the solution for avoiding hate speech. The way of using an algorithm which can observe and in case it is necessary remove / mark hate speech is a better choice.

Give your opinion Reflect (0)

3 Arguments

**CONS**

6 months ago

**AlanU**

I think that even people can flow in determining if something is really hate speech or if it is a actually only against their beliefs

Reflect (0)

6 months ago

**Bob**

Newspapers in general typically should not be open for comment sections, it is the opinion of that set newspaper. Also its easy to throw the word "Algorithm" around with not knowing the technology behind it.

Reflect (0)

6 months ago

**LawrenceM**

I agree that it is the current solution for hate speech as not all the comments are appropriate

**EXPLORE THE ANALYTICS**

37 Positions 54 Arguments

(a) Condition A - Baseline, no artefacts deployed

(b) Condition B - Synoptical summariser is present



(c) Condition C - SciArgRecSys is present

(d) Condition D - Both synoptical summariser and SciArgRecSys are present

Figure 7.9: BCause UI in four different experimental conditions

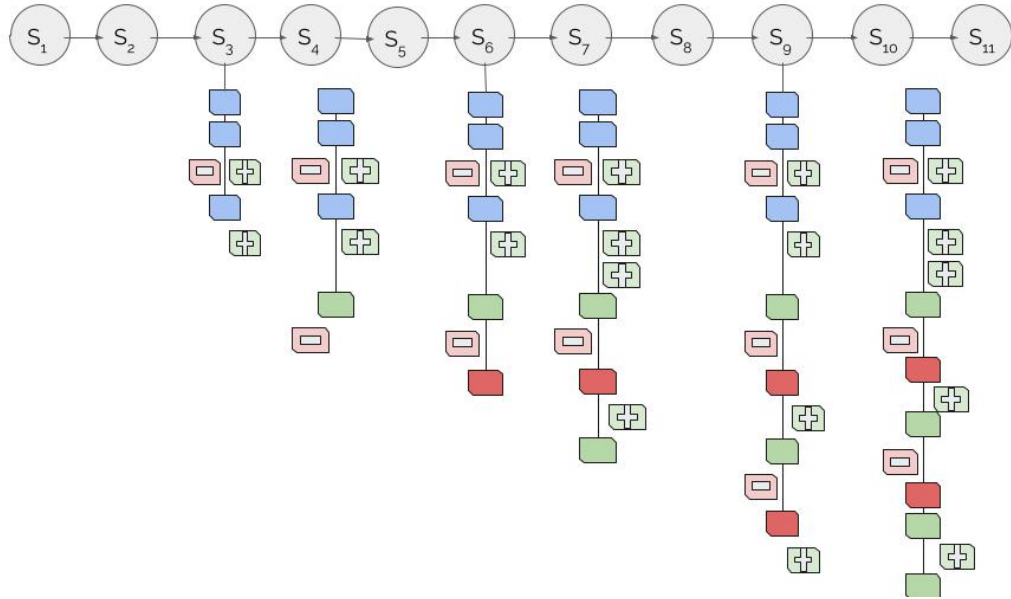


Figure 7.10: User task steps of interaction and the corresponding state of the discussion. Artificially injected posts are shown as red rectangles, user’s contributions as green rectangles

### Task

The task consisted of several stages of interacting with one of the interfaces of each condition A,B,C,D (described above) and answering questions in between. As depicted in the diagram of Figure 7.10, the experiment had the following steps:

- Step 1: A welcoming message including information about the study and a consent form for the participant to approve before proceeding to the next steps
- Step 2: Demographic data (Age, Gender, Education) were asked and instructions were given for the next step
- Step 3 (1st Interaction): Users were taken to BCause platform where they were already logged in and redirected to the discussion page with topic: “How can we fairly reduce hate speech from newspapers websites without eliminating space for readers comments?”. The discussion was pre-populated with 3 positions and 3 arguments as its initial state. These arguments were randomly selected by a real discussion on the topic runs in the pilot UX studies reported above.

### 7.3. Study V - Assessing the Combined Impact of Automated Reporting and Scientific Arguments Reco

Participants were tasked to contribute at least 1 (one) position or argument before proceeding to the next step.

- Step 4: The users were asked questions regarding the *Aesthetics* of the platform.
- Step 5: Instructions were given for the next interaction
- Step 6 (2nd Interaction): At this step, users are returning back to the same discussion as step 3. Between these two steps, up to two posts (either up to two positions or one position and one argument or one argument) were artificially injected by virtual users along with other contributions of the rest of the 9 users simultaneously undertaking the same study. Users were encouraged to reflect on contributions by other users and proceed to contribute at least one position or argument.
- Step 7: Users were asked questions regarding their *Mutual Understanding* of the topic with other users and provide their own short summary (this was as an attention check to ensure reliable responses from participants).
- Step 8: Instructions for the next (last) interaction step were given
- Step 9 (3rd Interaction): Users return in the same discussion as Step 3 and 6 where again up to 2 posts (a mixture of position and arguments) were artificially posted. They were asked to contribute at least 2 posts (new position or argument on a given position).
- 10: Questions regarding Engagement and Sensemaking
- 11: Open feedback questions regarding the Synopsis (for conditions B,D) and SciArgRecSys (for conditions C,D) and their overall experience with the platform.

We recruited participants using the prolific<sup>2</sup> platform. Our selection criteria included only the knowledge of English language (as the discussion was undertaken in English).

---

<sup>2</sup><https://www.prolific.co/>

A reward of the equivalent of £10 per hour was awarded: £5 for the anticipated finish time of 30 minutes. In case participants exceeded the 30 min time mark, an additional reward analogous to their extra time was rewarded, and an upper threshold of maximum 50 minutes to complete the task was used. Users were redirected from prolific in a custom survey platform developed by us, where we implemented the task steps described above. Each trial was executed with 10 users and repeated a total of 20 (twenty) times. In total, we had 4 conditions x 10 users x 20 trials = 800 participants in this survey and a total of £4870 was awarded. The selection of a substantial number of repetitions (20 trials) was influenced by several factors:

- *Anticipated Minimized Effect Size*: Given the expectation of a small effect size, it was deemed necessary to increase the number of repetitions to enhance the reliability of effect detection.
- *High Variability*: The inherent subjectivity associated with the evaluation questions of any human-centric assessment further underscores the need for a larger number of iterations. This high level of variability was evident, for instance, in the human-centric evaluations of SciArgRecSys, as discussed in Section 6.3.
- *Minimal Incremental Cost*: After the experimental framework had been established, the resource implications—in terms of researcher time—for additional repetitions were minimal. Consequently, increasing the number of repetitions was a feasible strategy for enhancing the robustness of our findings.

## Evaluation Factors

Our experiment targeted the following variables:

- Sensemaking, we reuse the same evaluation factors as introduced in Section 4.3.
- Mutual Understanding: adapted from existing validated scales of the literature on Grounding Cost theory (Clark and Brennan, 1991) and studies of common

### 7.3. Study V - Assessing the Combined Impact of Automated Reporting and Scientific Arguments Recon

ground building in mediated conversations ([Monk and Watts, 2000](#); [Convertino et al., 2007, 2009](#); [Whittaker et al., 1998](#)).

- Engagement: we adapted the scale developed by O'Brien ([O'Brien and Toms, 2010](#)) for use in the online discussion domain.
- Aesthetics: is well connected with Engagement and Usability, see for example [Ashby et al. \(1999\)](#). Furthermore is crucial for the creation of effective knowledge representations ([Iandoli et al., 2020](#)) therefore impacts sensemaking as “aesthetic reasoning in design drives problem solving even when the problem does not entail aesthetic preoccupations”.

To avoid user fatigue, we used a minimised questionnaire than the one of UNINA use case ([7.2.2](#)), shown in Table [7.5](#).



Variable	Factor	Code	Question
<i>Sensemaking</i>	Reflection	SM1	I was able to reflect on the debated question
	Insights	SM2	I was provided with unexpected insights on what the question is and what are the main arguments for and against
	Focus	SM3	I was not able to focus on different aspects of the debate
	Argumentation	SM4	I was able to find structure in the information provided in this debate and find a way to organise it
	Explanation	SM5	I was not able to identify the main points raised in this debate
	Assess facts and evidence	SM6	I was able to assess facts and evidence provided in this debate
	Distinguish	SM7	I was able to distinguish between different people's claims
	Assess assumptions	SM8	I was not able to assess my initial assumptions about this debate
	Change assumptions	SM9	Some initial assumptions I had about this question changed
<i>Mutual Understanding</i>		MU1	In general, I have not had problems to understand the meaning of other team members' posts
		MU2	In general, I think that the other team members have understood my contributions without difficulty
		MU3	I could easily understand who has done what
		MU4	My teammates and I developed better understanding about each other over the time
		MU5	My teammates and I developed shared understanding about the task over the time
		MU6	I found there are many irrelevant posts respect to the assigned task
<i>Aesthetics</i>	Aesthetics	A1	The interface is pleasing to see
		A2	The interface is attractive
		A3	The interface is beautiful
		A4	The interface is too crowded
	Perceived Complexity	A5	The interface has a lot of variety
		A6	The interface is complicated
		A7	The information available on the interface is easy to view
	Fluency	A8	It is easy to identify the controls and indicators on the interface
		A9	it is easy to understand the information available on the interface
		A10	I think this tool is easy to operate without instructions
<i>Engagement</i>	Focused Attention	E1	For a moment, I forgot about my immediate surroundings while discussing on this website
	Perceived usability	E2	I felt frustrated while visiting this discussion website
	Endurability	E3	Discussing on this website was not worthwhile
	Novelty	E4	The experience with the discussion website incited my curiosity
	Felt involvement	E5	I was not really drawn into the discussion

Table 7.5: Evaluation factors for inclusion of artefacts study

### 7.3.3 Results

We show the demographic data of the participants who overtaken the task in Figures 7.11 (gender), 7.12 (education level) and 7.13 (age). Regarding gender we had a split of 58% females, 41% males, 1% other). Regarding education level, we had 36% of high school graduates, 55% with higher education (Bachelor, Masters or Doctorate) and 8% other (no diploma or professional). Regarding the age of participants, we had 58% between 18-34, 31% between 35 and 54, and 10% over 55. Overall we had no significant discrepancies between the four conditions in any of the demographic categories.

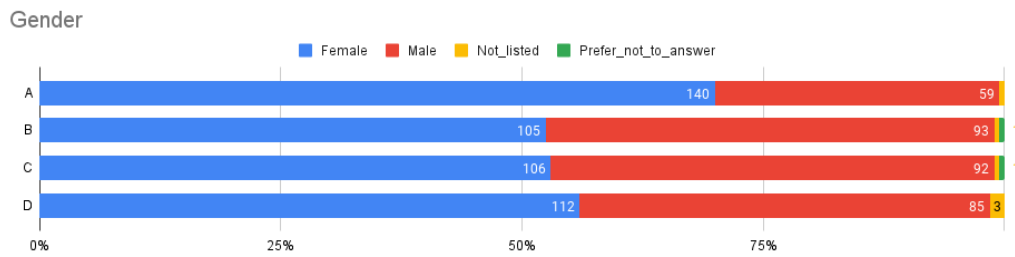


Figure 7.11: Participants' gender

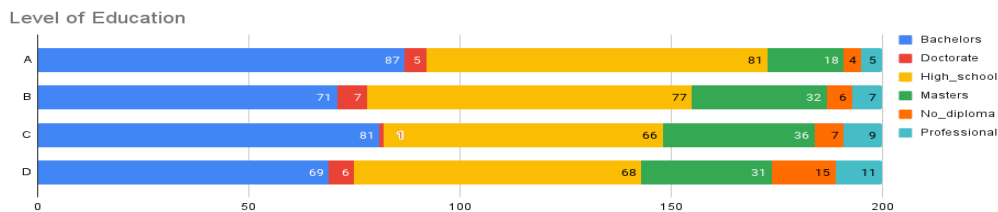


Figure 7.12: Participants' level of education

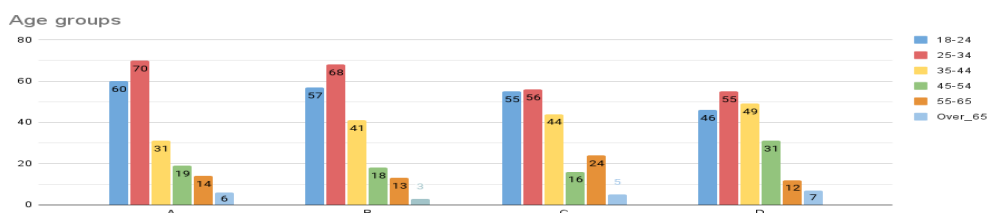


Figure 7.13: Participants' age groups

### Statistical Analysis

We show the total number of posts (positions or arguments) contributed by users averaged for each trial (and filtered from the initial posts or artificially injected between interaction steps) in Table 7.6. The average length -across all conditions- of a position was 225 characters long, while the average length of an argument was 148.1 characters long. Going from condition A to condition D, i.e. from the simpler solution to the full solution, we observe a slight reduction in the total number and length of posts. This decrease can be interpreted as an indication of users' inclination towards confining their contributions to concise yet meaningful inputs; particularly when aided by artefacts that support their sensemaking process.

			Per trial average			
			Number of posts		Avg. length of post (chars)	
	trials	N	positions	arguments	position	argument
A	20	200	46.7	56.0	241.2	150.9
B	20	200	45.0	55.2	222.0	153.5
C	20	200	42.5	52.1	216.4	140.0
D	20	200	39.7	45.9	220.4	147.8

Table 7.6: User contributions statistics over conditions A-D

In Table 7.7, we depict how many times the two artefacts were used when they were deployed. On average, the SciArgRecSys artefact was used 8.55 times in each trial. In condition C it was used in total 8.4 times (7.4 for positions and 0.8 for arguments), while for condition D it was used in total 8.9 times (7.8 for positions and 1.1 for arguments). The invocation of the SciArgRegSys artefact resulted in 6.3 posts having at least one external argument attached (in condition C 5.6 positions and 0.7 arguments, and in condition D 5.3 positions and 1.0 argument). We observe, that positions tend to have more external evidence attached (on average for conditions C and D, 5.55 positions with external evidence) compared to arguments (only 0.85 arguments with attached external evidence on average for conditions C and D). The higher use of the recommender for positions rather than for arguments

### 7.3. Study V - Assessing the Combined Impact of Automated Reporting and Scientific Arguments Recon

was a foreseeable outcome, given the greater demand for supporting evidence to substantiate claims made within a position rather than an argument. Summariser was invoked (generate a fresh summary of the current state of discussion) on average (10.7 for condition B and 14.4 for condition D) 12.55 times during each trial, i.e. the summary displayed on the side was updated 12.5 during the course of the discussion.

			Per trial average				
			External evidence				Summaries
	trials	N	in positions	positions with	in arguments	arguments with	generations
A	20	200	0.0	0.0	0.0	0.0	0.0
B	20	200	0.0	0.0	0.0	0.0	10.7
C	20	200	7.4	5.6	0.8	0.7	0.0
D	20	200	7.8	5.3	1.1	1.0	14.4

Table 7.7: Invocations of computational artefacts per trial

In Table 7.8, we show the number of times appreciative feedback elements was used (“thank you”/“like”, position agreement, reflection on position or argument). We observe a significant drop in the number of reflections given from condition A (baseline) (12.4) compared to other conditions (-27% drop on average). This may be explained that the introduction of the two computational artefacts introduced a distraction (information overload) to the user that neglected the use of this feature. In other features, however, we observe in some cases an increase in their use. For example, agreements are used 20% more in condition B compared to condition A. This may be due to a better comprehension of the posts in the presence of the summariser and sensemaking nuggets which induces users to voluntarily submit their agreement to positions. However, for this aspect of Sensemaking, we can make a safer conclusion with the analysis of the Sensemaking questions where users explicitly denote their level of Sensemaking in various dimensions.

			Appreciative elements (Per trial average)								
	trials	N	thank you/- like	pos. with agr.	agr. per pos.	args with agr.	agr. per arg.	pos. with refl.	refl. per pos.	args with refl.	re pe ar
A	20	200	1.0	10.5	1.2	0.6	1.5	12.4	1.3	1.5	0.
B	20	200	0.9	12.6	1.1	0.6	0.8	10.1	1.3	0.9	0.
C	20	200	1.0	9.6	1.1	0.4	0.7	7.6	1.4	0.7	0.
D	20	200	0.8	8.2	1.2	0.6	1.1	9.1	1.3	1.1	0.

Table 7.8: Appreciative and reflective feedback statistics

We present the mean score and variance of each of the evaluation factors 7.5 in Table 7.9, followed by statistical significance test (with the use of Kruskal-Wallis and post-hoc comparisons using Dunn test) in Table 7.10.

	<b>A</b>		<b>B</b>		<b>C</b>		<b>D</b>	
	$\mu$	$\sigma^2$	$\mu$	$\sigma^2$	$\mu$	$\sigma^2$	$\mu$	$\sigma^2$
A1	3.104	1.109	3.122	1.046	3.084	1.282	3.054	1.181
A2	2.955	1.237	3.022	1.117	2.976	1.180	2.986	1.156
A3	2.515	1.077	2.583	1.060	2.467	0.913	2.419	0.762
A4	3.465	1.404	3.583	1.406	3.629	1.138	3.581	1.089
A5	3.450	0.945	3.611	0.708	3.473	0.697	3.554	0.589
A6	3.104	1.546	3.200	1.491	3.395	1.289	3.284	1.429
A7	2.931	1.388	3.022	1.262	2.844	1.217	2.905	1.447
A8	3.099	1.323	3.150	1.335	2.916	1.150	3.027	1.006
A9	3.347	1.004	3.317	1.022	3.132	0.983	3.122	1.114
A10	2.955	1.297	2.900	1.275	2.814	1.309	2.770	1.307
E1	3.471	1.450	3.417	1.481	3.420	1.411	3.401	1.500
E2	2.752	1.338	2.904	1.808	3.375	1.721	3.220	1.742
E3	2.267	1.123	2.364	1.405	2.670	1.456	2.503	1.188
E4	3.275	1.030	3.620	0.893	3.403	1.191	3.340	0.947
E5	2.359	1.422	2.337	1.289	2.602	1.464	2.447	1.185
MU1	3.878	0.634	4.012	0.579	3.982	0.588	4.216	0.453
MU2	3.784	0.712	3.880	0.568	3.914	0.453	3.973	0.609
MU3	3.327	1.023	3.536	1.085	3.545	1.141	3.404	1.094
MU4	3.099	0.802	3.273	1.057	3.293	0.907	3.405	0.758
MU5	3.378	0.704	3.536	0.865	3.473	0.781	3.397	0.680
MU6	2.168	1.110	2.268	1.120	2.479	1.275	2.327	0.957
SM1	3.180	0.616	4.102	0.619	4.000	0.674	3.925	0.665
SM2	3.024	0.914	3.337	1.106	3.261	1.166	3.233	0.939
SM3	2.354	1.118	2.652	1.303	2.670	1.251	2.566	1.184
SM4	3.217	1.122	3.000	1.233	3.364	1.257	3.220	0.932
SM5	2.403	1.315	2.481	1.477	2.563	1.242	2.635	1.436
SM6	3.573	0.695	3.422	1.127	3.466	0.765	3.509	0.910
SM7	3.403	0.644	3.754	1.047	3.625	0.784	3.786	0.701
SM8	2.038	1.146	2.444	1.076	2.670	0.999	2.585	1.054
SM9	2.874	1.184	2.663	1.085	2.756	1.237	2.591	0.977

Table 7.9: Evaluation factors descriptive statistics

Variable	fl-Statistic	p-value	A-B	A-C	A-D	B-C	B-D	C-D
A1	0.121	0.948	-	-	-	-	-	-
A2	0.932	0.425	-	-	-	-	-	-
A3	0.125	0.946	-	-	-	-	-	-
A4	0.844	0.470	-	-	-	-	-	-
A5	0.723	0.539	-	-	-	-	-	-
A6	1.340	0.260	-	-	-	-	-	-
A7	1.920	0.125	-	-	-	-	-	-
A8	0.715	0.543	-	-	-	-	-	-
A9	1.464	0.223	-	-	-	-	-	-
A10	2.396	0.067	-	-	-	-	-	-
E1	1.753	0.155	-	-	-	-	-	-
E2	9.230	0.000	*	***	**	**	*	-
E3	4.453	0.004	-	**	*	*	*	-
E4	4.719	0.003	**	*	*	-	*	-
E5	1.961	0.119	-	*	-	-	-	-
MU1	3.291	0.020	*	-	**	-	*	*
MU2	2.093	0.100	-	*	*	-	-	-
MU3	1.368	0.251	**	**	*	-	-	-
MU4	0.348	0.791	*	*	*	-	-	-
MU5	0.914	0.434	*	-	-	-	-	-
MU6	1.224	0.300	-	**	-	*	-	-
SM1	3.544	0.014	**	*	*	-	-	-
SM2	3.191	0.023	*	*	*	-	-	-
SM3	3.419	0.017	*	*	*	-	-	-
SM4	5.625	0.001	-	**	-	**	-	-
SM5	1.341	0.260	-	-	-	-	-	-
SM6	0.925	0.428	-	-	-	-	-	-
SM7	3.118	0.026	**	*	**	-	-	-
SM8	2.884	0.035	*	***	*	-	-	-
SM9	2.449	0.062	-	-	-	-	-	-

Table 7.10: Evalaution factors Kruskal-Wallis statistical test and post-hoc comparisons

We observe no statistically significant difference in all *Aesthetics* factors. There is however an average neutral score across all conditions - an average of 3.097 can be interpreted that on average users are neutral on whether the interface was beautiful and fluent. The introduction of any of the two artefacts on their own or in combination did not disrupt this neutrality. Regarding engagement, there is a

### 7.3. Study V - Assessing the Combined Impact of Automated Reporting and Scientific Arguments Recon

statistically significant improvement in the usability (E2), endurability (E3) and novelty (E4) for all conditions compared to baseline condition A. The conditions where SciArgRecSys is deployed on its own perform better in E2 and E3 than Summariser (B-C comparison), whereas in summariser on its own performs better in E4. In general, users report better level of understanding of other team members (MU1-MU4) when any artefacts is deployed. Only in MU1 factor their understanding level is higher when both artefacts are deployed together. Regarding Sensemaking, there was a statistically significant improvement in Reflection (SM1), Insights(SM2), Focus(SM3), Argumentation (SM4), Distinguish(SM7), Assess assumptions (SM8) when any of the artefacts was deployed (compared to the baseline condition A). In Argumentation (SM4) factor, the SciArgRecSys did better when compared to Summariser only (B-C comparison).

### **Social network analysis**

Social Network Analysis (SNA) is the process of investigating social structures through the use of networks and graph theory ([Brandes and Wagner, 2004](#)). It involves mapping relationships and flows between people, groups, organizations, or other information/knowledge processing entities. We employ SNA method due to its inherent characteristic as a non self-reporting metric (compared for example with previous users answering a questionnaire). This methodology provides more robust and objective evidence regarding the interactions and dynamics among the participants in the discussion, therefore offering a more reliable understanding of the underlying phenomenon.

To carry out a social network analysis in a discussion platform, you need to construct an interaction graph.



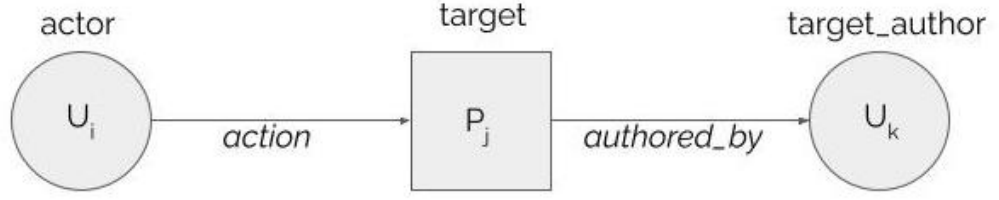


Figure 7.14: Interaction model

In the context of BCause, a user is interacting to posts authored by other users. In the interaction graph, nodes are the users (actors) and edges represent interactions between these users.

$$\langle U_i, A_{i,j}, P_j, U_k \rangle \quad (7.6)$$

So we model as interaction edge the following actions:

- given-heart
- given-reflection
- given-agreement
- posted-position-as-reply
- posted-opposing-argument and posted-supporting-argument
- attached-evidence

A visualisation of the graph  $G = V, E$ , where  $V = \{U_i, U_k\}$  and  $E = \{A_{i,j}\}$  for a typical use case is shown in Figure 7.15.

A more informative model occurs if, instead of only the users as nodes we also include in the node-set positions or arguments keeping the edges set the same, i.e. the graph  $G = V, E$ , where  $V = \{U_i, P_j\}$  and  $E = \{A_{i,j}\}$ . In this way, we can visualise the orphan positions - posts that have received no attention by any user except their author, see for example, the case as before in Figure 7.16.

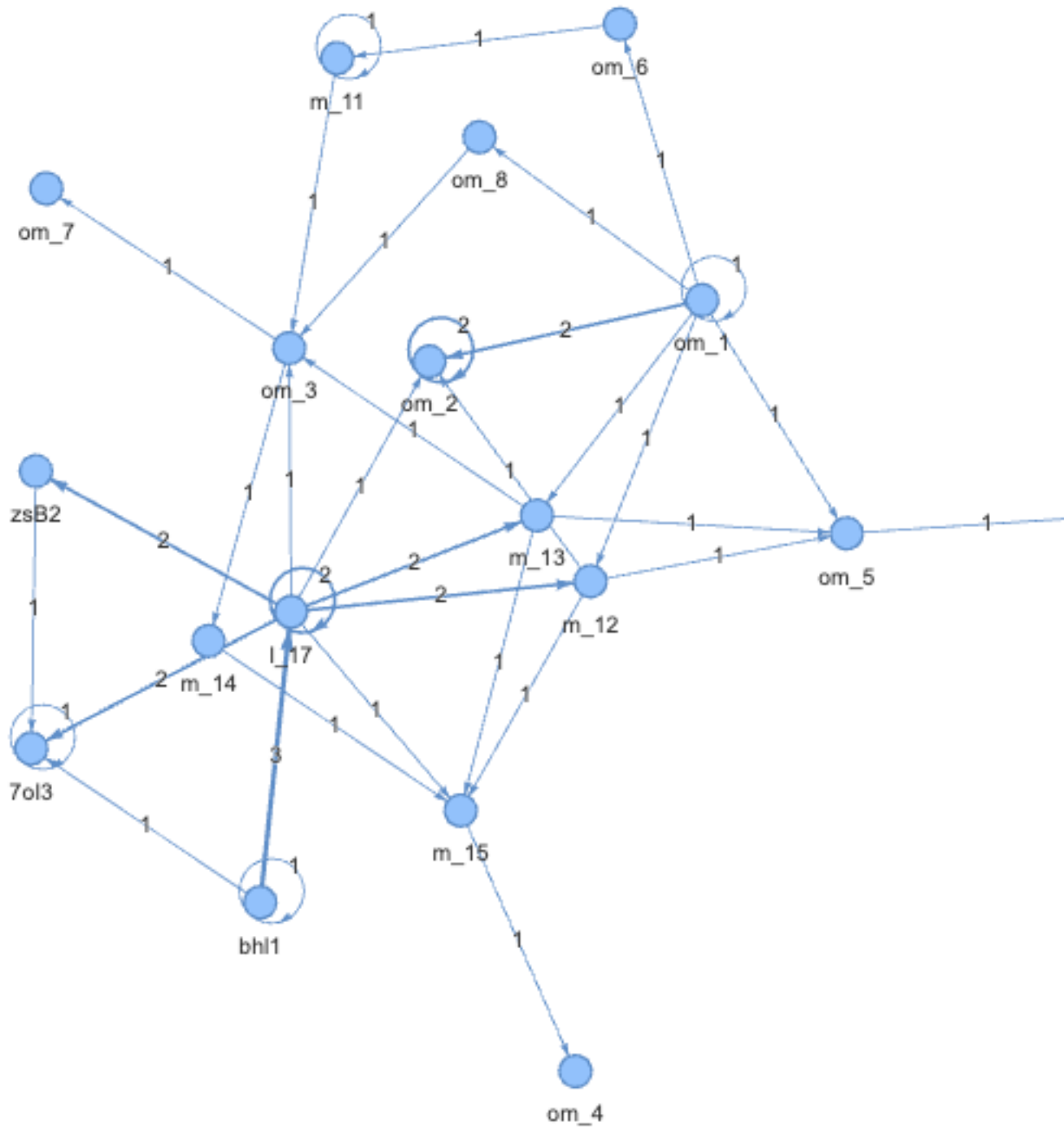


Figure 7.15: User interaction graph



### 7.3. Study V - Assessing the Combined Impact of Automated Reporting and Scientific Arguments Recon

	A	B	C	D
#nodes	<b>27.400</b>	27.150	27.000	23.950
#edges	<b>61.150</b>	54.300	54.000	50.300
average degree	<b>4.474</b>	3.983	3.972	4.119
density	0.156	0.143	0.143	<b>0.181</b>
diameter	5.000	5.000	n/a	<b>5.250</b>
transitivity	<b>0.223</b>	0.202	0.213	0.213
Occurances of connected graph	1	3	0	<b>12</b>
number of components	3.750	4.400	4.850	<b>2.300</b>
largest_component_size	<b>24.797</b>	24.650	23.850	22.650
largest component diameter	5.450	5.600	<b>6.000</b>	5.600
% largest component to total graph	0.905	0.908	0.883	<b>0.946</b>

Table 7.11: Social network analysis results

**Interpretation of SNA metrics.** In the following, we describe the metrics we calculated for each testing group and that will be used to inform insights on changes in social dynamics across the 4 experimental conditions (Table 7.11).

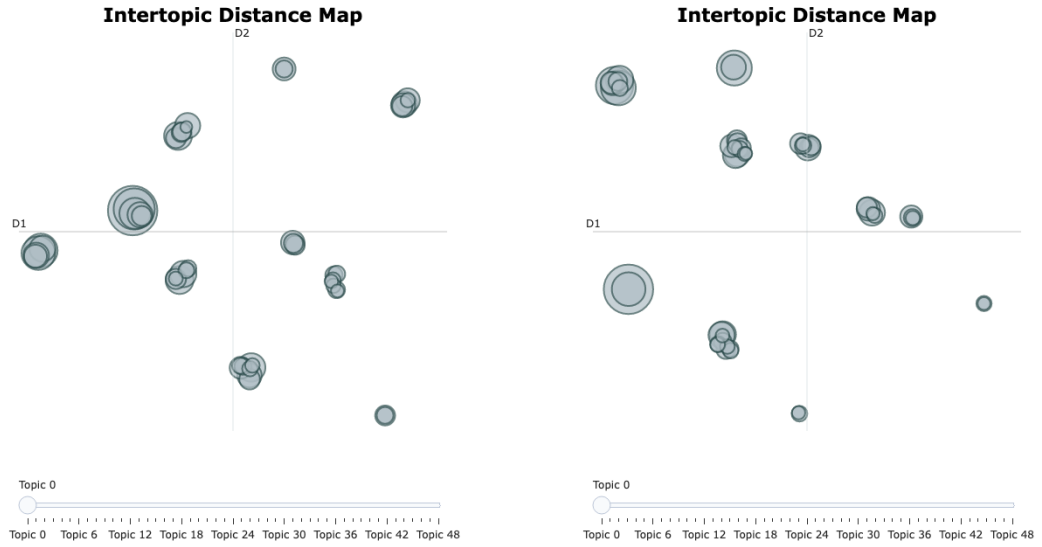
- *Number of nodes* is the average (over 20 trials of each condition) of nodes appearing, i.e. the total number of users participating in the experiment.
- *Number of edges* is the average (over 20 trials of each condition) of user interactions with a position or argument (by one of the aforementioned actions). It denotes the social interaction size, i.e. if people basically connect more to other people's ideas rather than posting isolated ideas.
- *Average degree* is simply the average number of edges per node in the graph. It can be interpreted as the social interaction coverage, basically how well the social interactions are distributed across the group.
- *Density* is a measure of the connectedness of the network in terms of the total number of connections divided by the maximum possible number (of the perfectly interconnected graph). So higher density means more interconnectedness.

- *Occurrences of connected graph*: the number of trials (out of 20 for each condition) that had a connected graph, i.e. there were no isolated “islands” of nodes.
- *Diameter* denotes the shortest distance between the two most distant nodes in the graph, or in other words, the maximum distance between any pair of nodes in the graph. This can be calculated if we have a connected graph, and can be interpreted as how “big” is the useful graph the derive results from.
- *Transitivity* is the overall probability for the network to have adjacent nodes interconnected, thus revealing the existence of tightly connected communities (or clusters, subgroups, cliques). This reveals the existence of more tightly connected communities.
- *Number of components*: a connected component (or just component) is a subgraph in which any two vertices are connected to each other by paths, and which is connected to no additional vertices in the supergraph. It can be interpreted as the number of subgroups or tribes forming within this small community during each trial of the experiment.
- *Largest component size*: the size (number of nodes) of the largest sub-graph. This denotes how big the centric group of users is. We observe no significant fluctuations on this metric.
- *Largest component diameter* similarly to Largest component size, this metric denotes how “big” is the “useful” part of the graph
- *% Largest component to total graph* Combined with the largest component size, we can say that participants in our experiment organised themselves around a large centric group rather than scatter to small isolated groups.

The social network analysis reveals several salient attributes of the behaviour of the network of interactions of the study participants we examined. From the results in Table 7.11, we observe that there is a small drop in the number of nodes, the number

### 7.3. Study V - Assessing the Combined Impact of Automated Reporting and Scientific Arguments Recon

of edges and the average degree with the introduction of the artefacts (Conditions B, C, D). This small drop from the baseline case A could be explained by information overload (attention diffusion) as more artefacts are introduced, a speculation that is enforced by the fact that condition D (both artefacts enabled) has the biggest drop from the baseline. The number of nodes fluctuates between 24 to 27 as the number of synthetic users injecting data between experiment stages is stochastically deduced (varies from 1 to 3 random injections on each round). However, in other metrics such as the Density, the fully-fledged BCause version (Condition D) where both artefacts were deployed, has an improved 16% value. It is worth noted that density representing the proportion of potential connections that are actual connections is relatively low (ranges from 0.156 to 0.181) meaning only a small portion of the possible ties in the network were realised. In terms of the network's diameter, which represents the longest distance between any two nodes, in Condition D we had the smallest value on average (5.255). This means that users interacting on the platform were not very distant from each other (compared to other conditions). Transitivity was approximately the same for all conditions - fluctuating from 0.202 to 0.223. This low transitivity score suggests that this network is characterized by a low degree of clustering. Moreover, in the metrics that directly reflect the creation of separate sub-communities, Condition D had emphatically better metrics. For example, it had a higher chance of a fully-connected graph (it occurred 12 out of 20 trials, whereas in other conditions A, B, and C only 1,3,0 respectively). Also, the average number of components (the number of unconnected sub-communities formed) was significantly lower (only 2.3 compared to 3.75 of baseline condition A, a 38% reduction). Moreover, the largest component diameter and percentage of the largest component to total graph, were significantly better for conditions B, C, D. Interpreting these two metrics together, we can say that participants in our experiment organised themselves around a large centric group rather than scatter to small isolated groups. This was more prominent in Condition D (+5% improvement from baseline) where both artefacts are present. This is a strong indication of more reciprocal engagement happening in conditions where any or both artefacts are deployed compared to the baseline

(a) Condition **A** topics distance map(b) Condition **B** topics distance map

condition.

### Topic analysis

We use topic analysis to automatically extract topics from the debates used in our experiment, and reflect on topic diversity and coherence within discussion groups. Similarly to SNA, this method does not rely on self-reported evidence and offers a less subjective indicator of discussion dynamics. Topic analysis, also known as topic modelling, is a machine-learning technique that automatically analyzes text data to determine cluster words for a set of documents. This is done in “unsupervised” fashion because it does not require a predefined list of tags or training data that’s been previously classified by humans. Specifically, we use BERTopic ([Grootendorst, 2022](#)), a topic modelling technique that leverages BERT embeddings and a class-based variant of TF-IDF to create dense clusters of similar documents.

The purpose of using topic-modelling in this study is to compare topic diversity and topic coherence of the discussion across the four experimental conditions. These two metrics are crucial aspects for the quality of the text data generated by the discussion and whether they are comprehensive and interpretable.

*Topic diversity* in the context of topic modelling refers to the variety or dissimilarity

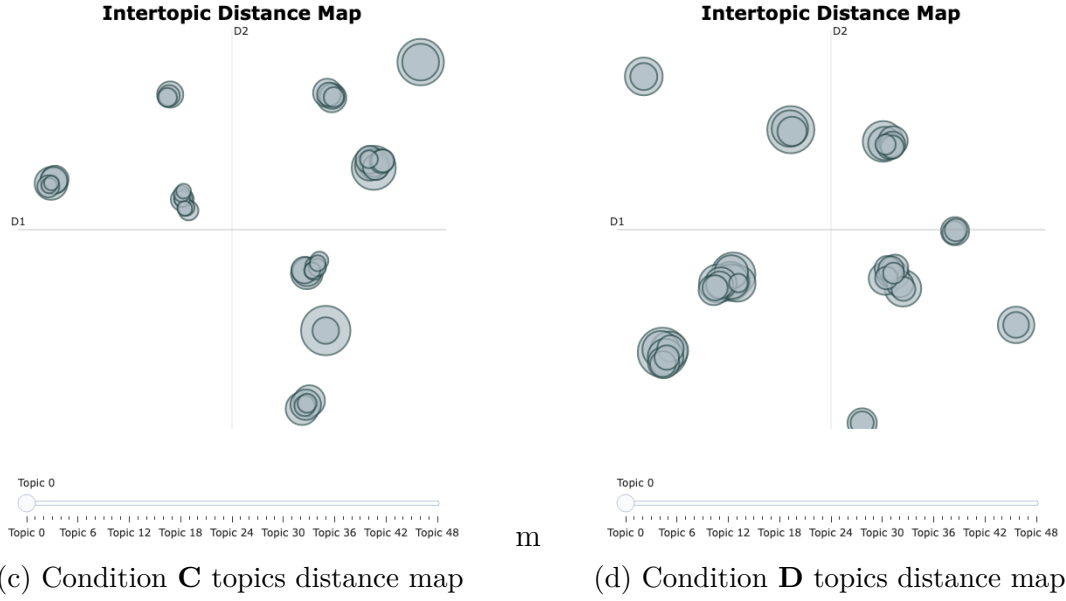


Figure 7.17: Intertopic distance map in four different experimental conditions

of topics that the model can generate. A model with high topic diversity will produce a wide range of distinct topics, covering a broad spectrum of themes or subjects. On the other hand, a model with low topic diversity may produce many similar topics or repeatedly focus on a narrow set of themes.

*Topic coherence* quantifies the semantic similarity of the high-scoring words within each topic. In other words, it is a measure of how much the words in a topic “go together” from the perspective of a human reader. A high coherence score suggests that the words within a topic are semantically related and thus the topic is likely to be interpretable and meaningful. On the other hand, low coherence scores may suggest that a topic is not very interpretable neither meaningful.

	Diversity	Coherence
A	0.8	0.1673042285
B	0.818	0.1723180599
C	0.812	0.1700532887
D	0.846	0.1863352518

Table 7.12: Topic diversity and coherence over conditions A-D



	A	B	C	D
1	speech, comments, hate, freedom, sections, allowed, commenting, commenters, ban, they	comments, hate, section, speech, be, report, will, sections, removing, it	none, feels, perfect, makes, sense, right, may, this, it,	negative, consequences, effective, curtailing, cism, it, freedom, sp balance, hate
2	freedom, free, speech, say, shouldn, others, their, opinions, expense, views	opinions, peoples, open, opinion, agree, say, allowed, idea, discussion, others	consequences, negative, curtailing, effective, criticism, freedom, speech, it, balance, hate	responsibility, curta tolerated, censoring, quences, freedom, ab subjected, punish, er ment
3	moderators, bias, moderation, biases, moderator, speech, checkers, complaints, hate, themselves	target, groups, vulnerable, damage, violence, protect, bigotry, lead, hate, consequences	participation, regulated, tolerance, disrespectful, standards, space, mindedness, social, should, support	newspapers, spe sided, order, sl reduce, platform, pr sections, comments
4	comments, moderated, moderating, moderation, moderator, inappropriate, flagged, moderate, commentators, allow	freedom, speech, hate, cases, trouble, subjectivity, specifically, violation, regulating, regulations	agree, sane, agreed, alison, galliah, ian, touch, honest, communication, discussion	hate, moderators, ments, speech, v monitoring, flag, ha delete, moderator
5	target, vulnerable, groups, hate, protect, reduce, lead, bigotry, believe, violence	its, profile, sided, newspapers, public, opinions, would, publicity, editorial, rounded	algorithm, mark, articles, public, avoiding, case, remove, reducing, which, speech	comments, monitore message, should, ina priate, monitor, repc user, media

Table 7.13: Top topic models of experiment corpus

We present topic diversity and topic coherence scores of topic modelling of each of the four conditions in Table 7.12. We observe a 5.7% improvement in topic diversity in the case of condition D compared to condition A (both artefacts deployed compared to none). Similarly, an 11.3% increase in topic coherence is observed between condition D and A.

It is noteworthy that both metrics, topic diversity and topic coherence, serve as indicators of topic modelling effectiveness to represent the generated corpus. Inherently, they can serve as quality indicators of the underlying corpus. Consequently, it can be inferred from the augmented diversity and coherence score of conditions B, C and -emphatically in- D, where one or both artefacts were deployed, that the discourse is more diverse, multifaceted and poly-thematic. We show the top 5 topics of each condition in Table 7.13.

#### 7.3.4 Discussion

We presented an evaluation study of the two artefacts that we developed in Chapters 5 and 6. The purpose of this evaluation was to examine the effect of those two artefacts when deployed together and in a real live discussion platform. We looked at participants' Sensemaking, various metrics of the quality of the deliberation process, Social Dynamics of the discussion, and Content of the Discussion as evaluation dimensions. Results suggest that, firstly, when any of the two artefacts is deployed in a discussion platform, there is an increase in the reported Engagement, Mutual Understanding, and Sensemaking. Specifically, there is an improvement in Perceived Usability (E2), Endurability (E3), Novelty (E4), in four out of 6 of the Mutual Understanding factors (MU1-MU4), in Reflection (SM1), Insights (SM2), Focus (SM3), Argumentation (SM4), Distinguish claims (SM7) and Assess assumptions (SM8). We observe a greater effect in Perceived Usability (E2), Endurability (E3), Post relevance (MU6), Argumentation (SM4) in case of the SciArgRecSys deployed (Condition C) when compared to only Summariser deployed (condition B). Interestingly, when both artefacts are deployed together there is no further improvement when compared with any of the conditions of only one artefact deployed (B-D and C-D comparison) - with the exception of one Mutual Understanding factor (MU1) and 3 Engagement factors, E2, E3, E4, where improvement is observed only in comparison to the Summariser condition (B-D comparison). Further, we employed Social Network Analysis to unveil the dynamics of the interaction of users when participating in a discussion with or without the presence of any of the two artefacts (or both together). We observe again a small drop in the size of interaction (number of edges) in the presence of artefacts; however, simultaneously we observe significantly better metrics that denote the quality of the discussion. Specifically, we observe better density, transitivity, diameter, the chance of a fully-connected graph to occur, a lower number of distinct components and the largest percentage of the largest component to the total graph. As these metrics reflect users' actions and behaviour, i.e. how many times and in what way they reply to others, have a higher value than what is self reported in

the user questionnaire though both report the same finding. The low number of components is a strong indicator of less chance of tribes formation which leads to polarization and echo chamber effects. Lastly, to explore the discussion content quality we performed topic modelling with the use of BERTopic. We observe a higher diversity and coherence of the discussion topic models when the two (any or both together) artefacts are present. This means that discussion evolves to be more polythematic, pluralistic and avoids redundancy (do not reiterate over the same topic). This is a good indication of the richness of the discussion, and that participants are engaging in more thoughtful, reflective discussion rather than simply repeating pre-existing viewpoints that could indicate to effective deliberation. We can also speculate that the improved diversity was due counter-acting participants' biases; for example in-group bias/partisan bias ([Tarrant et al., 2012](#)), false consensus bias ([Nickerson, 1999](#)) and confirmation bias ([Edwards and Smith, 1996](#)) - which was a central design goal of BCause.

We employed various methods in our investigation: quantitative user survey, social network analysis, topic modelling, which involves a methodological triangulation to study the same phenomenon. This gave us a more comprehensive and well-rounded understanding of the topic in question and contributes to make the findings above more rigorous.

## 7.4 Summary

In this chapter, we ran an extrinsic evaluation of the two artefacts described in Chapters [5](#) and [6](#) when present in a real live deliberation platform. We undertook a large-scale evaluation study, carefully examining an array of variables regarding the individual Sensemaking, perceived quality of deliberation and social dynamics. Through systematic analysis and interpretation of the data, we examined the effect of automatic synopsis and SciArgRecSys in Sensemaking, Engagement, Mutual understanding, Aesthetics, Social Dynamics in online discussion and quality of discussion. After having examined the effect of the two artefacts both solely deployed

---

and in combination, we can conclude that utilising computational argumentation artefacts shows improvements in both participants' perceptions of the deliberation quality, the social dynamics facilitated by the deliberation process, and topics diversity and coherence in the resulting deliberation content. These findings suggest that argument computation approaches and technologies show potential to improve online deliberation.

# Chapter 8

## Conclusions

“The more chords you know, the more courage it takes to not play them.”

---

Jacob Collier

### 8.1 Discussion

The main objective of this doctoral work was to assess if incorporating computational argumentation tools has the potential to improve the quality of online deliberation. This is attained by supporting the individual Sensemaking of participants in this highly complex cognitive process and advancing the quality of debate by promoting the use of scientific arguments in the discussion.

My inquiry began with exploring the ill phenomena observed in deliberation platforms today and the concerns and aspirations of users regarding the use of technologies to improve public deliberation and decision making. At the same time, a focus on argument computation was explored, by looking to what concerns and aspirations could be best addressed by argument computation technologies. This lead to the design and development of two proposed artefacts to improve the process of sensemaking and the quality of deliberation.

The two artefacts proposed were: (i) a synoptical summariser of the long discussions happening online and (ii) a recommender of scientific arguments to promote evidence-based reasoning. The development and evaluation of those two technological enhancements was carried out with a user-centric perspective, as those were ultimately evaluated by real users as part of a real deliberation platform. In the investigation, we focused on a series of variables to assess our two proposed artefacts: such as: human sensemaking, engagement with the platform, mutual understanding and social dynamics of the debate. In the following, we discuss our findings by organising them around the research questions driving the overall investigation (Section 1.3.1).

**RQ1: What are the higher-level guidelines for the design of deliberative platforms for organisational decision-making?** We followed an inductive approach to explore the issues surrounding the use of social media platforms for online deliberation and, in parallel, uncover the long-term aspirations of users and experts. For that, we interviewed experts from various domains (n=14) and through thematic analysis, we uncovered major aspirations and concerns regarding the use of social media in online deliberation setups. Based on the set of aspirations and fears, we proposed design guidelines (*DG*) centred around the themes of *accountability*, *argument-structure*, *collective decision-making* and *privacy*, presented in Section 2.2.6. These guidelines, though induced from the set of interviews conducted with domain experts, highly correlate to previous design guidelines found in the literature. For example, they match the “Design for Quality Content” principle of (Towne and Herbsleb, 2012) and adapt it to the modern era for the application of novel technological artefacts. They are also broad enough to cover a wide range of online deliberation as well, in contrast to most guidelines, e.g. Esau et al. (2017) that specialise only on news platforms or Aragón et al. (2017b) that focus on the single use case of citizen participation.

Further to the proposed guidelines, we explored possible computational elements grounded on users’ aspirations that can contribute to the improvement of individual

Sensemaking and perceived quality of deliberation. From those, we propose that a synoptical summariser and a recommender of scientific arguments have the highest potential toward this aspiration.

**RQ2: What automated reporting approaches are more appropriate for online deliberation?** Through *Study II*, we have shown that summarisation approaches have an effect on Sensemaking and the perceived quality of debate. We conclude that a blend of templated and abstractive summaries embodied with argumentative elements has the highest potential to computationally aid online deliberation. Similarly, argumentative highlight has more interesting attributes, assisting the participant to *Reflect*, to gain *Insights*, to *Focus* and find structure and organise information (*Argumentation*) in a better way. This approach is in alignment with other approaches that promote healthy engagement with online debates such as [Rieger et al. \(2022\)](#) where main incentive is the prevention of cognitive biases by facilitating understanding (a lighter flavour of Sensemaking) and bringing in persuasive arguments.

We then propose a set of 6 design principles for the design of automated reports (see Section 4.4), that we use for the development of the Synoptical summariser artefact. This set of guidelines offers a novel framework for designing a summariser of online discussions purely through a human Sensemaking lens. This is in continuation of previous attempts to bring a Sensemaking focus on summary evaluation such as [Mani et al. \(2002\)](#) (news articles) or other attempts to understand human preferences for summary design ([Santhan et al., 2016](#)).

**RQ3: To what extent can an AI-generated abstractive synopsis of an online discussion provide a quality summary of the discussion and significantly improve Sensemaking?** To answer this question, we cross-compared various state-of-the-art automatic summarising models when applied to long discussions, and performed the evaluation on human metrics of quality of the output summary (*Study III*). We also assessed the effect of the presence of each summary output on

human Sensemaking. We deduced that prompting Large Language models (LLM) is the best method according to intrinsic human evaluation metrics. In particular, it far exceeds other methods in terms of Fluency, Coherence, Relevance, Concision and Readability; however, LLMs perform slightly lower, but still competitive, in terms of Comprehensiveness, Accuracy, Adequacy and Factuality. We speculate that in the former metrics, LLMs outperform other methods because of their ability to generate human believable text, while in the later metrics, LLMs suffer from the well-known phenomenon of “hallucinations” (Bender et al., 2021) or their output does not produce such accurate text as other summarisation models exclusively designed and trained for this task (due to the stochastic process of generation). However, the drop in accuracy and factuality seems to be so small that is tolerable or even not attained by users. This discovery aligns with recent findings of other research within the realm of news article summarisation (Tam et al., 2022); yet, it consists of a novel result in the context of applying LLMs to the synthesis of online discussions.

Overall, the task of summarising long discussions is quite distinct from the widely explored problem of summarising news articles that has been the main focus of the summarisation research community so far. This is because a discussion is substantially semantically richer than a news article. In our approach, we propose techniques of how to represent the discussion as a long text document that is then used as an input prompt. Furthermore, we explored prompt templates that specifically target Sensemaking features. This offers simple and actionable ways of addressing the challenge of utilising LLMs for summarising online discussions.

Regarding the effect in Sensemaking, prompting LLMs has -again- the highest effect in almost all Sensemaking factors with the exception of the Focus factor. Interestingly, not only the factors about the early stages of sensemaking (SM1-SM4) were improved, as in the case of the initial study shown in *Study II*, but also the later stages of sensemaking, e.g. Distinguish, Assess assumptions, Change assumptions, were positively affected with a great effect. This reaffirms the somewhat persuasive power of LLMs to affect users’ views - as shown for example in a recent study that opinionated LLMs can propagate their bias to participants (Jakesch et al., 2023).



This signals the need for the research community and industry to engage in the oversight of unintended bias emerging from the application of LLMs, to defer from recklessly calibrating models and rather targeting beneficial social interventions.

**RQ4: To what extent does the provision of quality external scientific arguments improve sensemaking and the overall quality of the online deliberation process?** We commenced the investigation of developing a scientific argument recommender (*SciArgRecSys*) by examining the feasibility of scientific argument extraction in large scale (*Study IV*). For that, we chose the simple -yet effective for our case - argumentation scheme of ArguminSci (Lauscher et al., 2018a) with only three argumentative classes. We then compared the original model of ArguminSci to a finetuned discriminatory LLM (BERT) and generative LLM (GPT-3) in various configurations. We deduced that discriminatory models are more appropriate for argument component identification compared to generative models - as expected for classification tasks, e.g. shown for sentiment analysis (Mullick et al., 2022) or for hate speech detection where LLMs are no better than random guessing (Liang et al., 2022). This implies that generative models regardless of their impressive performance in synthesizing, composing, and in general creative tasks (among those summarisation as shown in the previous study), still underperform when compared with discriminatory models.

As an alternative unit of recommendation, we proposed the use of LLMs to produce paper’s main argument summary. For that, we show methods to prompt generative models (GPT-3) to produce a condensed yet largely accurate representation of a research paper abstract that conveys its main argument (Section 6.2.3). Overall, this novel approach comprises a noteworthy contribution, even to the domain of research paper recommendation systems (Beel et al., 2016), which is distinct to the field of this work.

We used this argument extractor component to build test corpora - along with other two corpora of paper abstracts and paper’s main argument transformed (Section 6.3). We used these corpora to test different methods of querying based on a user’s position

on a topic. We compared the different units of recommendation, using TF-IDF and kNN query methods using three different embeddings (sBERT, SPECTER, GPT-3). We conclude that recommending argument summaries is the preferred recommendation unit in various dimensions of the intrinsic recommendation quality. Specifically, it shows the highest performance in Accuracy, Adequacy, Perceived Usefulness and Satisfaction factors. In terms of embeddings, GPT-3 based embeddings perform better in Diversity and Satisfaction, however, SPECTER embeddings (that have been trained exclusively on scientific text) perform better in Accuracy regardless the unit of recommendation.

**RQ5: To what extent does the automated reporting and provision of scientific arguments in combination improve Sensemaking and the quality of the online deliberation process?** To answer this question we have undertaken an exhaustive study described in Chapter 7. The conclusion is that both automated reporting and provision of scientific arguments have a largely positive effect in Sensemaking and the quality of online deliberation. Specifically, both synoptical summarisation as a method of automated reporting and provision of scientific argument through a recommender had a positive effect on the Engagement variable (Perceived Usability, Endurability, and Novelty factors) when deployed on their own, however, when both simultaneously were present, there was a small decrease in improvement. This could be explained as an outcome of *limited attention* and *information overload* cognitive limitation (Eppler and Mengis, 2008), i.e. providing more features and information can overwhelm our cognitive capacity to process and evaluate it all (Schick et al., 1990). Regarding Sensemaking, there was a statistically significant improvement in Focus, Argumentation, Distinguishing, Assess assumptions in the case of the *SciArgRecSys* artefact and in Reflection and Insights in the case of the *Synoptical summariser* artefact. Moreover, a positive effect when both artefacts are deployed is observed in Mutual Understanding.

This is a significant improvement vis-à-vis other argumentation tools for public deliberation proposed over the years (e.g. Benn and Macintosh (2011) or Klein

(2012) that do not adequately support these aspects of Sensemaking and signals for a rebirth of argumentation-centric solutions.

Via social network analysis, we also derived important conclusions on the impact of our technological enhancements on *social dynamics*. Specifically, we can deduce that including computational argumentation artefacts exerts a positive effect on both the participants' perception of online discussion and the social interaction enabled by the discussion platform. Furthermore, including these aids does not undermine *reciprocal participation*; on the contrary, it enhances aesthetic perception and engagement in online debate. Most notably, the reduction of the number of components and increased occurrences of fully connected interaction graphs indicate substantial *reduction of polarisation*, isolated ideas and non-constructive conflict. Along with other research (e.g. interface design Bossens et al. (2021)), this contribution can potentially have a great impact in our modern society where incivility prevails online and ways to tackle polarization are needed.

Expanding our analysis, we employed - a third method of - topic modelling to examine the quality of the discussion generated during this study. We conclude that the discussion corpus is more diverse and coherent if any of the two artefacts are present, a positive effect that is amplified when both are deployed. This discussion enrichment effect would be expected with the use of a component that facilitates or moderates discussion (e.g. in Sasaki et al. (2021)), i.e. it has been designed exclusively for this cause. Given recent research findings indicating the positive impact of diversity on a range of collective intelligence processes (Ito et al., 2022), the ability of our argument computation aids to improve diversity could be both cause and effect of further beneficial improvements in the quality of deliberation (Woolley et al., 2010).

## 8.2 Limitations

We recognise that the studies and the work overall presented in this dissertation have their limitations. Firstly, most of the studies (with the exception of *study I*), were carried out in controlled environments as part of a between-subject factorial

design experiment. This is the most rigorous way to isolate the target variables you are interested in, however, is not a naturalistic setup, and therefore it inevitably reduce the external validity of our study i.e. you would expect if the experiment is carried out in-the-wild (Rogers and Marshall, 2017) different results due to unforeseen real-world events. However, our methodological choice has strengthened the internal validity of our results, and together with the implementation of measures to introduce relevant elements of realism, we believe we have provided our studies with a strong degree of ecological validity.

In studies around the development of *Summariser* and *SciArgRecSys* artefacts (*Study III* and *Study IV*) the focus was not to compare our proposed method performance against existing state-of-the-art techniques, but instead, our investigation was dedicated to optimise for performance as assessed by human evaluators, as this was the target variable of interest. This was also underscored by the shortage of datasets that perfectly suited our task. For example, while there is a plethora of datasets for ‘argument retrieval for comparative questions’, e.g. the Touché corpus (Bondarenko et al., 2022) or scientific argument corpus e.g. SciArg corpus (Binder et al., 2022), there is no suitable dataset specifically designed for “scientific argument retrieval for controversial questions” as the *SciArgRecSys* artefact development required.

Finally, regarding the quality of deliberation, as this was used in *Study II* and *Study V*, this was in reality the perceived quality of debate as expressed by the subjective views of participants in the discussion. We could have elaborated a different method to assess the quality of deliberation in more objective way, such as by recurring to expert annotation or using quantitative computational metrics. However, automated metrics are not reliable as they often struggle to capture the nuances of argumentative discussion we were after (Klein, 2010). Moreover they would have introduced the bias of the model used to detect those. In addition, annotation by experts is very costly. Our intention was to capture the internal perception of participants about the quality - if they feel their opinions are heard, if they find the messages clear if they interpret opinions shared well reasoned, etc - and therefore such metrics were

deemed the most cost-effective for our purposes.

### 8.3 Future work

Overall, the work presented in this dissertation stands in the middle ground of four different research areas: (i) Human-Computer Interaction (HCI), (ii) Natural Language Processing (NLP), (iii) Online Deliberation and (iv) Argumentation.

We propose the creation of specialised resources that can accommodate the continuation of this interdisciplinary research. For example, a dedicated dataset comprising debate data, scientific evidence and user interaction analytics, can accommodate the further refinement of scientific argument recommender components. It can take into account other features not considered in the creation of *SciArgRecSys*, such as previous user interactions (e.g. likes on other positions), the inferred stance of the user on the debated topic as inputs, differentiation to the purpose of recommending (e.g. to challenge user’s position instead of only supporting it).

Regarding the two components developed, there is a range of ways that can be extended or further evaluated. For instance, the summariser artefact could benefit from the utilisation of more capable LLMs - particularly those with significantly larger context windows. Additionally, a comprehensive examination of the prompt design can be employed. Furthermore, the summaries originally given by users in *study V* as an attention check, can be used for the creation of a rich dataset of human summaries-long discussion pairs. This dataset may be further used for creating a fine-tuned LLM specialising in this task. Likewise, the survey platform implemented for *Study V* can also be reused to get human judgments on different summary variations that can be harnessed later in training an LLM within an RLHF framework. Additional techniques of improving argumentative models warrant investigation. For instance, finetuning with embeddings - in alignment with the paradigm reported in Section 6.2.2) - as it is poised to yield improved performance due to the complex semantics of the underlying data. Overall, the *SciArgRecSys* artefact serves as a proof of concept, demonstrating the potential for scaling out

to bigger corpora. The prospect of querying massive scholarly literature databases (such as CORE (Knoth et al., 2023) or OpenAlex<sup>1</sup>, which encompass the majority of the whole of scientific literature) offers promising avenues for future research. Moving our approach to web-scale may encounter unforeseen phenomena but also accommodate serendipitous discoveries, expanding the knowledge in the domain.

An alternative pathway for the recommender component would be to explore variants of argumentation schemes to identify the most appropriate for the task of recommending scientific evidence to participants of online discussion. In the current work, for the development of SciArgRecSys we used the -rather straightforward- ArguminSci scheme, however, employing a richer scheme may yield a different impact on the downstream task of improving sensemaking or quality of deliberation.

As improving Sensemaking or quality of deliberation is the main objective of our research, we may harness the power of LLMs in different ways than just summariser and recommender to promote this. In Chapter 2 we identified a selection of computational artefacts that have a potential for this purpose but other interventions are made possible with the advent of powerful LLMs such as *automatic argument generation*. In Section 7.1.2, we showed the conceptualisation of such intervention, nevertheless due to the limitations of models available at the time, we withheld further exploration of this concept.

Moreover, an active way of utilising LLMs -compared to the ways described in this thesis that require explicit user action or just passively appear on the side- to engage citizens in online deliberation can be explored by the use of intelligent agents. We can envision a system where discussion agents, like e.g. in Ito et al. (2022), monitors and facilitates discussion, intervening where appropriate to promote the discussion. Naturally, though, any system of this nature requires an elevated level of security and safety measures, due to the potential risks; nevertheless is an extraordinarily compelling research trajectory with the potential to yield transformative effects in the field of online deliberation.

---

<sup>1</sup><https://openalex.org/>

## 8.4 Final Remarks

This thesis shed light on the complex interplays that the inclusion of computational artefacts has in the perceptions, experiences and behaviors of participants of an online deliberation process. The main endeavour of this thesis was to elucidate exactly these complex processes of incorporating computational artefacts in online discussion platforms. I thereby emphasised the importance of careful and well-informed execution of such integrations.

Throughout this work (*Study II*)(*Study III*) and (*Study IV*)(*Study V*), we followed a Human-Computer Interaction (HCI) perspective providing insights on the effect of argument computations enhancements on human understanding (*Sensemaking*) and perception of *quality of deliberation*. This offers a more human-behaviour oriented perspective than the conventional practice of NLP technology assessment, which primarily concentrates on the improvement of specific internal quality metrics, without foreseeing the overall context that computational elements live in. This kind of thinking is in line with earlier work on “The Pragmatic Web” which argued that Web technologies only become useful when applied in well-understood communities of use (Singh, 2002). We aspire for these tools to significantly reduce the cognitive load associated with participating in online discussion and overall enhance the efficiency and effectiveness of deliberation platforms, as long as AI-safety mechanisms are in place. The sky being the limit in the promises being made by the developers of novel language technologies, we hope we have made the case that the human factor is equally important as advanced technological features. Along the lines of Douglas Engelbart (Engelbart, 1962), we advocate that AI-powered systems should be developed to *augment* human intelligence, rather than to replace it.

In the course of this research, we made extensive use of generative Large Language Models in large-scale scenarios. I hope that the approach presented in this work will allow for a more accurate representation of the practical implications and applicability of LLMs, therefore contributing to the public discourse on the use of generative AI. I argue that the uptake of Generative Large Language Models creates a paradigm shift,

signalling a transition from relying on tons of labelled data and heavily supervised methods to models capable of convincingly creating human-like results with little effort. Regardless of their -countless- limitations, and the need for explainable AI, or methods to hold accountable such capable systems before engaging in large-scale deployments, our findings suggest the emergence of a more fruitful research path where AI's focus is not dry (e.g. concerning often marginal computational improvement in prediction), but rather applied to enhance human sensemaking and collaboration.



# Bibliography

- Aakhus, M. and Lewiński, M. (2017). Advancing polylogical analysis of large-scale argumentation: Disagreement management in the fracking controversy. *Argumentation*, 31:179–207. [80](#)
- Abdi, A., Idris, N., Alguliev, R. M., and Aliguliyev, R. M. (2015). Automatic summarization assessment through a combination of semantic and syntactic information for intelligent educational systems. *Information Processing & Management*, 51(4):340–358. [120](#)
- Adams, G., Alsentzer, E., Ketenci, M., Zucker, J., and Elhadad, N. (2021). What’s in a summary? laying the groundwork for advances in hospital-course summarization. In *Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting*, volume 2021, page 4794. NIH Public Access. [119](#)
- Aladalah, M., Cheung, Y., and Lee, V. C. (2016). Delivering public value: Synergistic integration via gov 2.0. In *2016 49th Hawaii International Conference on System Sciences (HICSS)*, pages 3000–3009. IEEE. [52](#)
- Albrecht, S. (2006). Whose voice is heard in online deliberation?: A study of participation and representation in political debates on the internet. *Information, Community and Society*, 9(1):62–82. [2](#)
- Alemaný, L. A. and Fuentes, M. (2003). Cohesion and coherence for automatic summarization. In *Student Research Workshop*. [120](#)

- Allen, J. (1995). *Natural language understanding*. Benjamin-Cummings Publishing Co., Inc. [92](#)
- Almahy, I. and Salim, N. (2014). Web discussion summarization: study review. In *Proceedings of the First International Conference on Advanced Data and Information Engineering (DaEng-2013)*, pages 649–656. Springer. [119](#)
- Alsufiani, K., Attfield, S., and Zhang, L. (2017). Towards an instrument for measuring sensemaking and an assessment of its theoretical features. In *Proceedings of the 31st British Computer Society Human Computer Interaction Conference*, page 86. BCS Learning and Development Ltd. [102](#), [103](#)
- Amini Parsa, V., Salehi, E., Yavari, A. R., and van Bodegom, P. M. (2019). Evaluating the potential contribution of urban ecosystem service to climate change mitigation. *Urban Ecosystems*, 22(5):989–1006. [147](#)
- Ammar, W., Groeneveld, D., Bhagavatula, C., Beltagy, I., Crawford, M., Downey, D., Dunkelberger, J., Elgohary, A., Feldman, S., Ha, V., et al. (2018). Construction of the literature graph in semantic scholar. *arXiv preprint arXiv:1805.02262*. [151](#)
- Aragón, P., Gómez, V., and Kaltenbrunner, A. (2017a). Detecting platform effects in online discussions. *Policy and Internet*, 9(4):420–443. [62](#)
- Aragón, P., Kaltenbrunner, A., Calleja-López, A., Pereira, A., Monterde, A., Barandiaran, X. E., and Gómez, V. (2017b). Deliberative platform design: The case study of the online discussions in decidim barcelona. In *Social Informatics*. [223](#)
- Ashby, F. G., Isen, A. M., et al. (1999). A neuropsychological theory of positive affect and its influence on cognition. *Psychological review*, 106(3):529. [201](#)
- Association, A. P. et al. (1990). Critical thinking: A statement of expert consensus for purposes of educational assessment and instruction. *ERIC document ED*, 315:423.

- Azzollini, A. and Pomponio, A. (2019). Positive energy static solutions for the chern-simons-schrödinger system under a large-distance fall-off requirement on the gauge potentials. [86](#)
- Bächtiger, A. and Parkinson, J. (2019). *Mapping and measuring deliberation: towards a new deliberative quality*. Oxford University Press. [62](#)
- Baeza-Yates, R., Ciaramita, M., Mika, P., and Zaragoza, H. (2008). Towards semantic search. In *Natural Language and Information Systems: 13th International Conference on Applications of Natural Language to Information Systems, NLDB 2008 London, UK, June 24-27, 2008 Proceedings 13*, pages 4–11. Springer. [150](#)
- Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*. [85](#)
- Bahdanau, D., Cho, K., and Bengio, Y. (2016). Neural machine translation by jointly learning to align and translate. [87](#)
- Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., Joseph, N., Kadavath, S., Kernion, J., Conerly, T., El-Showk, S., Elhage, N., Hatfield-Dodds, Z., Hernandez, D., Hume, T., Johnston, S., Kravec, S., Lovitt, L., Nanda, N., Olsson, C., Amodei, D., Brown, T., Clark, J., McCandlish, S., Olah, C., Mann, B., and Kaplan, J. (2022). Training a helpful and harmless assistant with reinforcement learning from human feedback. [90](#)
- Bakshy, E., Messing, S., and Adamic, L. A. (2015). Exposure to ideologically diverse news and opinion on facebook. *Science*, 348(6239):1130–1132. [2](#), [19](#)
- Baldwin, P. and Price, D. (2008). Debategraph. [4](#)
- Banerjee, S. and Lavie, A. (2005). Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72. [94](#)

- Banfield, R., Lombardo, C. T., and Wax, T. (2015). *Design sprint: A practical guidebook for building great digital products.* " O'Reilly Media, Inc.". [179](#)
- Barberá, P., Jost, J. T., Nagler, J., Tucker, J. A., and Bonneau, R. (2015). Tweeting from left to right: Is online political communication more than an echo chamber? *Psychological science*, 26(10):1531–1542. [19](#)
- Barbour, R. S. (2018). Doing focus groups. *Doing focus groups*, pages 1–224. [21](#)
- Batista, J., Ferreira, R., Tomaz, H., Ferreira, R., Dueire Lins, R., Simske, S., Silva, G., and Riss, M. (2015). A quantitative and qualitative assessment of automatic text summarization systems. In *Proceedings of the 2015 ACM Symposium on Document Engineering*, pages 65–68. [91](#)
- Beck, J., Neupane, B., and Carroll, J. M. (2018). Managing conflict in online debate communities: Foregrounding moderators' beliefs and values on kialo. [69](#)
- Beel, J., Gipp, B., Langer, S., and Breitingner, C. (2016). Paper recommender systems: a literature survey. *International Journal on Digital Libraries*, 17:305–338. [226](#)
- Behrendt, M. and Harmeling, S. (2021). Arguebert: How to improve bert embeddings for measuring the similarity of arguments. In *Proceedings of the 17th Conference on Natural Language Processing (KONVENS 2021)*, pages 28–36. [91](#)
- Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big?. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623. [137](#), [225](#)
- Benn, N. and Macintosh, A. (2011). Argument visualization for eparticipation: towards a research agenda and prototype tool. In *Electronic Participation: Third IFIP WG 8.5 International Conference, ePart 2011, Delft, The Netherlands, August 29–September 1, 2011. Proceedings 3*, pages 60–73. Springer. [227](#)

- Bentahar, J., Moulin, B., and Bélanger, M. (2010). A taxonomy of argumentation models used for knowledge representation. *Artificial Intelligence Review*, 33(3):211–259. [77](#)
- Bhatia, S., Biyani, P., and Mitra, P. (2014). Summarizing online forum discussions—can dialog acts of individual messages help? In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 2127–2131. [119](#)
- Binder, A., Verma, B., and Hennig, L. (2022). Full-text argumentation mining on scientific publications. *arXiv preprint arXiv:2210.13084*. [229](#)
- Bishop, C. M. and Nasrabadi, N. M. (2006). *Pattern recognition and machine learning*, volume 4. Springer. [89](#)
- Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., Donahue, C., Doumbouya, M., Durmus, E., Ermon, S., Etchemendy, J., Ethayarajh, K., Fei-Fei, L., Finn, C., Gale, T., Gillespie, L., Goel, K., Goodman, N., Grossman, S., Guha, N., Hashimoto, T., Henderson, P., Hewitt, J., Ho, D. E., Hong, J., Hsu, K., Huang, J., Icard, T., Jain, S., Jurafsky, D., Kalluri, P., Karamcheti, S., Keeling, G., Khani, F., Khattab, O., Koh, P. W., Krass, M., Krishna, R., Kuditipudi, R., Kumar, A., Ladhak, F., Lee, M., Lee, T., Leskovec, J., Levent, I., Li, X. L., Li, X., Ma, T., Malik, A., Manning, C. D., Mirchandani, S., Mitchell, E., Munyikwa, Z., Nair, S., Narayan, A., Narayanan, D., Newman, B., Nie, A., Niebles, J. C., Nilforoshan, H., Nyarko, J., Ogut, G., Orr, L., Papadimitriou, I., Park, J. S., Piech, C., Portelance, E., Potts, C., Raghunathan, A., Reich, R., Ren, H., Rong, F., Roohani, Y., Ruiz, C., Ryan, J., Ré, C., Sadigh, D., Sagawa, S., Santhanam, K., Shih, A., Srinivasan, K., Tamkin, A., Taori, R., Thomas, A. W., Tramèr, F., Wang, R. E., Wang, W., Wu, B., Wu, J., Wu, Y., Xie, S. M., Yasunaga, M., You, J., Zaharia, M., Zhang, M., Zhang, T., Zhang, X., Zhang, Y., Zheng,

- L., Zhou, K., and Liang, P. (2022). On the opportunities and risks of foundation models. [86](#)
- Bonabeau, E. (2009). Decisions 2.0: The power of collective intelligence. *MIT Sloan management review*, 50(2):45. [63](#)
- Bondarenko, A., Fröbe, M., Kiesel, J., Syed, S., Gurcke, T., Beloucif, M., Panchenko, A., Biemann, C., Stein, B., Wachsmuth, H., et al. (2022). Overview of touché 2022: argument retrieval. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 13th International Conference of the CLEF Association, CLEF 2022, Bologna, Italy, September 5–8, 2022, Proceedings*, pages 311–336. Springer. [148](#), [229](#)
- Bondarenko, A., Gienapp, L., Fröbe, M., Beloucif, M., Ajjour, Y., Panchenko, A., Biemann, C., Stein, B., Wachsmuth, H., Potthast, M., et al. (2021). Overview of touché 2021: argument retrieval. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 12th International Conference of the CLEF Association, CLEF 2021, Virtual Event, September 21–24, 2021, Proceedings 12*, pages 450–467. Springer. [148](#)
- Bossens, E., Storms, E., and Geerts, D. (2021). Improving the debate: Interface elements that enhance civility and relevance in online news comments. In *Human-Computer Interaction–INTERACT 2021: 18th IFIP TC 13 International Conference, Bari, Italy, August 30–September 3, 2021, Proceedings, Part IV 18*, pages 433–450. Springer. [228](#)
- Boxell, L., Gentzkow, M., and Shapiro, J. M. (2017). Greater internet use is not associated with faster growth in political polarization among us demographic groups. *Proceedings of the National Academy of Sciences*, 114(40):10612–10617. [19](#)
- Boyd, D. M. and Ellison, N. B. (2007). Social network sites: Definition, history, and scholarship. *Journal of computer-mediated Communication*, 13(1):210–230. [18](#)

- Bozdag, C. (2020). Managing diverse online networks in the context of polarization: Understanding how we grow apart on and through social media. *Social Media+ Society*, 6(4):2056305120975713. [2](#)
- Brandes, U. and Wagner, D. (2004). Analysis and visualization of social networks. *Graph drawing software*, pages 321–340. [209](#)
- Braun, V. and Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative research in psychology*, 3(2):77–101. [25](#)
- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Nee-lakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901. [86](#), [87](#), [143](#)
- Cabrio, E. and Villata, S. (2013). A natural language bipolar argumentation approach to support users in online debate interactions. *Argument & Computation*, 4(3):209–230. [83](#)
- Cabrio, E. and Villata, S. (2018). Five years of argument mining: a data-driven analysis. In *IJCAI*, volume 18, pages 5427–5433. [83](#), [102](#)
- Cañas, A. J., Hill, G., Carff, R., Suri, N., Lott, J., Gómez, G., Eskridge, T. C., Arroyo, M., and Carvajal, R. (2004). Cmaptools: A knowledge modeling and sharing environment. [4](#)
- Carr, C. S. (2000). *The effect of computer-supported collaborative argumentation (CSCA) on argumentation skills in second-year law students*. The Pennsylvania State University. [78](#)
- Carroll, J., Minnen, G., Canning, Y., Devlin, S., and Tait, J. (1998). Practical simplification of english newspaper text to assist aphasic readers. In *Proceedings of the AAAI-98 Workshop on Integrating Artificial Intelligence and Assistive Technology*, pages 7–10. [84](#)

- Caswell, D. and Dörr, K. (2018). Automated journalism 2.0: Event-driven narratives: From simple descriptions to real stories. *Journalism practice*, 12(4):477–496. [85](#)
- Chakrabarty, T., Hidey, C., Muresan, S., McKeown, K., and Hwang, A. (2020). Ampersand: Argument mining for persuasive online discussions. *arXiv preprint arXiv:2004.14677*. [83](#)
- Chandrasekar, R., Doran, C., and Srinivas, B. (1996). Motivations and methods for text simplification. In *Proceedings of the 16th conference on Computational linguistics-Volume 2*, pages 1041–1044. Association for Computational Linguistics. [84](#)
- Chen, H.-t. (1997). Applying mixed methods under the framework of theory-driven evaluations. *New directions for evaluation*, 1997(74):61–72. [25](#)
- Chen, H. T. (2012). *Theory-driven evaluation: Conceptual framework, application and advancement*. Springer. [25](#)
- Chernodub, A., Oliynyk, O., Heidenreich, P., Bondarenko, A., Hagen, M., Biemann, C., and Panchenko, A. (2019). Targer: Neural argument mining at your fingertips. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 195–200. [83](#)
- Cho, K.-L. and Jonassen, D. H. (2002). The effects of argumentation scaffolds on argumentation and problem solving. *Educational Technology Research and Development*, 50(3):5. [75](#)
- Chopra, S., Auli, M., and Rush, A. M. (2016). Abstractive sentence summarization with attentive recurrent neural networks. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 93–98. [85](#)
- Clark, H. H. and Brennan, S. E. (1991). Grounding in communication. [200](#)



- Cohan, A., Feldman, S., Beltagy, I., Downey, D., and Weld, D. S. (2020). Specter: Document-level representation learning using citation-informed transformers. *arXiv preprint arXiv:2004.07180*. [151](#)
- Cohen, L., Manion, L., and Morrison, K. (2017). *Research methods in education*. routledge. [21](#)
- Conklin, J. and Begeman, M. L. (1988). gibis: A hypertext tool for exploratory policy discussion. *ACM Transactions on Information Systems (TOIS)*, 6(4):303–331. [63](#), [79](#)
- Conover, M., Ratkiewicz, J., Francisco, M., Gonçalves, B., Menczer, F., and Flammini, A. (2011). Political polarization on twitter. In *Proceedings of the international aaai conference on web and social media*, volume 5, pages 89–96. [19](#)
- Convertino, G., Hong, L., Nelson, L., Pirolli, P., and Chi, E. H. (2009). Activity awareness and social sensemaking 2.0: Design of a task force workspace. In *Foundations of Augmented Cognition. Neuroergonomics and Operational Neuroscience: 5th International Conference, FAC 2009 Held as Part of HCI International 2009 San Diego, CA, USA, July 19-24, 2009 Proceedings 5*, pages 128–137. Springer. [201](#)
- Convertino, G., Mentis, H. M., Ting, A. Y., Rosson, M. B., and Carroll, J. M. (2007). How does common ground increase? In *Proceedings of the 2007 international ACM conference on Supporting group work*, pages 225–228. [201](#)
- Creswell, J. W. and Creswell, J. D. (2017). *Research design: Qualitative, quantitative, and mixed methods approaches*. Sage publications. [21](#), [99](#)
- Crotty, M. J. (1998). The foundations of social research: Meaning and perspective in the research process. *The foundations of social research*, pages 1–256. [13](#)
- Damay, J., Lojico, G., Lu, K., Tarantan, D., and Ong, E. (2006). Simtext: Text simplification of medical literature. In *Proceedings of the 3rd National Natural Lan-*

- guage Processing Symposium-Building Language Tools and Ressources*, volume 48. 84
- De Liddo, A. and Buckingham Shum, S. (2010). Cohere: A prototype for contested collective intelligence. 68
- De Liddo, A., Sándor, Á., and Shum, S. B. (2012). Contested collective intelligence: Rationale, technologies, and a human-machine annotation study. *Computer Supported Cooperative Work (CSCW)*, 21(4-5):417–448. 63, 186
- De Liddo, A., Souto, N. P., and Plüss, B. (2020). Let’s replay the political debate: Hypervideo technology for visual sensemaking of televised election debates. *International Journal of Human-Computer Studies*, 145:102537. 102
- De Lin, O., Gottipati, S., Ling, L. S., and Shankararaman, V. (2021). Mining informal & short student self-reflections for detecting challenging topics—a learning outcomes insight dashboard. In *2021 IEEE Frontiers in Education Conference (FIE)*, pages 1–9. IEEE. 151
- Deibert, R. J., Abdulrazzak, B., Marczak, B., Scott-Railton, J., and Mckune, S. (2019). The citizen lab. 70
- Dencheva, S., Prause, C. R., and Prinz, W. (2011). Dynamic self-moderation in a corporate wiki to improve participation and contribution quality. In *ECSCW 2011: Proceedings of the 12th European Conference on Computer Supported Cooperative Work, 24-28 September 2011, Aarhus Denmark*, pages 1–20. Springer. 66
- Dervin, B. (1999). Chaos, order and sense-making: A proposed theory for information design. *Information design*, pages 35–57. 54
- Dervin, B., Foreman-Wernet, L., and Lauterbach, E. (2003). *Sense-making methodology reader: Selected writings of Brenda Dervin*. Hampton Pr. 54
- Dervin, B. and Frenette, M. (2000). Sense-making methodology: Communicating communicatively. *Public communication campaigns*, page 69. 55

- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*. [86](#), [87](#), [95](#), [142](#)
- DiCicco-Bloom, B. and Crabtree, B. F. (2006). The qualitative research interview. *Medical education*, 40(4):314–321. [21](#)
- Dinno, A. (2015). Nonparametric pairwise multiple comparisons in independent groups using dunn’s test. *The Stata Journal*, 15(1):292–300. [106](#)
- Donaldson, L. (2005). Organization theory as a positive science. [14](#)
- Dou, Z.-Y., Liu, P., Hayashi, H., Jiang, Z., and Neubig, G. (2020). Gsum: A general framework for guided neural abstractive summarization. *arXiv preprint arXiv:2010.08014*. [101](#)
- Edwards, K. and Smith, E. E. (1996). A disconfirmation bias in the evaluation of arguments. *Journal of personality and social psychology*, 71(1):5. [220](#)
- Ehrgott, M., Klamroth, K., and Schwehm, C. (2004). An mcdm approach to portfolio optimization. *European Journal of Operational Research*, 155(3):752–770. [52](#)
- Ellison, N. B., Steinfield, C., and Lampe, C. (2007). The benefits of facebook “friends:” social capital and college students’ use of online social network sites. *Journal of computer-mediated communication*, 12(4):1143–1168. [19](#)
- Engelbart, D. C. (1962). Augmenting human intellect: A conceptual framework. *Menlo Park, CA*, 21. [232](#)
- Eppler, M. J. and Mengis, J. (2008). The concept of information overload-a review of literature from organization science, accounting, marketing, mis, and related disciplines (2004) the information society: An international journal, 20 (5), 2004, pp. 1–20. *Kommunikationsmanagement im Wandel: Beiträge aus 10 Jahren= mcm institute*, pages 271–305. [227](#)

- Esau, K., Friess, D., and Eilders, C. (2017). Design matters! an empirical analysis of online deliberation on different news platforms. *Policy & Internet*, 9:321–342. [223](#)
- Evans, J. S. B., Over, D. E., and Manktelow, K. I. (1993). Reasoning, decision making and rationality. *Cognition*, 49(1-2):165–187. [63](#)
- Faddoul, M., Chaslot, G., and Farid, H. (2020). A longitudinal analysis of youtube’s promotion of conspiracy videos. *arXiv preprint arXiv:2003.03318*. [2](#)
- Fan, A., Lewis, M., and Dauphin, Y. (2018). Hierarchical neural story generation. *arXiv preprint arXiv:1805.04833*. [89](#)
- Fishkin, J. (2009). *When the people speak: Deliberative democracy and public consultation*. Oup Oxford. [51](#)
- Fowler Jr, F. J. (2013). *Survey research methods*. Sage publications. [21](#)
- Friedman, B., Kahn, P., and Borning, A. (2002). Value sensitive design: Theory and methods. *University of Washington technical report*, 2:12. [179](#)
- Fuchs, C. (2008). Wikinomics: How mass collaboration changes everything. *International Journal of Communication*, 2:1–11. [51](#)
- Galletta, A. (2013). *Mastering the semi-structured interview and beyond: From research design to analysis and publication*, volume 18. NYU press. [21](#)
- Gambhir, M. and Gupta, V. (2017). Recent automatic text summarization techniques: a survey. *Artificial Intelligence Review*, 47:1–66. [92](#)
- Ganai, A. F. and Khursheed, F. (2019). Predicting next word using rnn and lstm cells: Stastical language modeling. In *2019 Fifth International Conference on Image Information Processing (ICIIP)*, pages 469–474. IEEE. [88](#)
- Gatt, A. and Krahmer, E. (2018). Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, 61:65–170. [92](#)

- Gaved, M., Calderón Lüning, E., Unteidig, A., Davies, G., and Stevens, J. (2019). Power, roles and adding value: reflecting on the challenges of bridging across research and action on an international community networking project. [34](#)
- Gillespie, T. (2020). Content moderation, ai, and the question of scale. *Big Data & Society*, 7(2):2053951720943234. [6](#)
- Glaser, B. G. (1978). *Theoretical sensitivity*. University of California,. [26](#)
- Glaser, B. G. and Strauss, A. L. (2017). *Discovery of grounded theory: Strategies for qualitative research*. Routledge. [25](#)
- Gliwa, B., Mochol, I., Biesek, M., and Wawer, A. (2019). Samsum corpus: A human-annotated dialogue dataset for abstractive summarization. *arXiv preprint arXiv:1911.12237*. [122](#)
- Goggins, S. and Xing, W. (2016). Building models explaining student participation behavior in asynchronous online discussion. *Computers and Education*, 94:241–251. [59](#)
- Golbeck, J., Ashktorab, Z., Banjo, R. O., Berlinger, A., Bhagwan, S., Buntain, C., Cheakalos, P., Geller, A. A., Gergory, Q., Gnanasekaran, R. K., et al. (2017). A large labeled corpus for online harassment research. In *Proceedings of the 2017 ACM on web science conference*, pages 229–233. [2](#)
- Graham, T. and Witschge, T. (2003). In search of online deliberation: Towards a new method for examining the quality of online discussions. *COMMUNICATIONS-SANKT AUGUSTIN THEN BERLIN-*, 28(2):173–204. [58](#), [61](#), [71](#), [102](#)
- Green, N. (2014). Towards creation of a corpus for argumentation mining the biomedical genetics research literature. In *Proceedings of the first workshop on argumentation mining*, pages 11–18. [141](#)
- Green, N. (2018). Proposed method for annotation of scientific arguments in terms

- of semantic relations and argument schemes. In *Proceedings of the 5th Workshop on Argument Mining*, pages 105–110. [142](#)
- Green, N., Ashley, K. D., Litman, D., Reed, C., and Walker, V. (2014). Proceedings of the first workshop on argumentation mining. In *Proceedings of the First Workshop on Argumentation Mining*. [80](#)
- Gregar, J. (1994). Research design (qualitative, quantitative and mixed methods approaches). *Book published by SAGE Publications*, 228. [21](#)
- Gretz, S., Friedman, R., Cohen-Karlik, E., Toledo, A., Lahav, D., Aharonov, R., and Slonim, N. (2020). A large-scale dataset for argument quality ranking: Construction and analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7805–7813. [148](#)
- Griffith, T. L. (1999). Technology features as triggers for sensemaking. *Academy of Management review*, 24(3):472–488. [3](#)
- Grootendorst, M. (2022). Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*. [216](#)
- Guntuku, S. C., Yaden, D. B., Kern, M. L., Ungar, L. H., and Eichstaedt, J. C. (2017). Detecting depression and mental illness on social media: an integrative review. *Current Opinion in Behavioral Sciences*, 18:43–49. [3](#)
- Gutmann, A. and Thompson, D. F. (2004). *Why deliberative democracy?* Princeton University Press. [51](#)
- Habermas, J. (1985). *The theory of communicative action: Volume 1: Reason and the rationalization of society*, volume 1. Beacon press. [6](#)
- Habermas, J. (2015). *The theory of communicative action: Lifeworld and systems, a critique of functionalist reason*, volume 2. John Wiley and Sons. [58](#)
- Habernal, I. and Gurevych, I. (2017). Argumentation mining in user-generated web discourse. *Computational Linguistics*, 43(1):125–179. [80](#), [83](#)

- Hahn, U. and Mani, I. (2000). The challenges of automatic summarization. *Computer*, 33(11):29–36. [91](#), [119](#)
- Halavais, A. (2016). Computer-supported collaborative learning. *The International Encyclopedia of Communication Theory and Philosophy*, pages 1–5. [68](#)
- Haque, M., Pervin, S., Begum, Z., et al. (2013). Literature review of automatic multiple documents text summarization. *International Journal of Innovation and Applied Studies*, 3(1):121–129. [119](#)
- Hardaker, C. (2010). Trolling in asynchronous computer-mediated communication: From user discussions to academic definitions. [66](#)
- Hassan, H., Aue, A., Chen, C., Chowdhary, V., Clark, J., Federmann, C., Huang, X., Junczys-Dowmunt, M., Lewis, W., Li, M., Liu, S., Liu, T.-Y., Luo, R., Menezes, A., Qin, T., Seide, F., Tan, X., Tian, F., Wu, L., Wu, S., Xia, Y., Zhang, D., Zhang, Z., and Zhou, M. (2018). Achieving human parity on automatic chinese to english news translation. [86](#)
- Heitz, L., Lischka, J. A., Birrer, A., Paudel, B., Tolmeijer, S., Laugwitz, L., and Bernstein, A. (2022). Benefits of diverse news recommendations for democracy: A user study. *Digital Journalism*, 10(10):1710–1730. [2](#)
- Helm, K. C. (2017). Investigation of the impacts of cognitive style and teamwork on ideation effectiveness. [66](#)
- Hermann, K. M., Kočiský, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., and Blunsom, P. (2015). Teaching machines to read and comprehend. [92](#)
- Hokamp, C. and Liu, Q. (2017). Lexically constrained decoding for sequence generation using grid beam search. *arXiv preprint arXiv:1704.07138*. [89](#)
- Holtzman, A., Buys, J., Du, L., Forbes, M., and Choi, Y. (2019). The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*. [89](#)

- Iandoli, L., Salado, A., and Zollo, G. (2020). The role of aesthetic reasoning in knowledge management: the case of elegant systems architecture design. *Knowledge Management Research & Practice*, 18(1):93–109. [201](#)
- Ito, T., Hadfi, R., and Suzuki, S. (2022). An agent that facilitates crowd discussion: A crowd discussion support system based on an automated facilitation agent. *Group Decision and Negotiation*, pages 1–27. [228](#), [231](#)
- Jackson, S. K. and Kuehn, K. M. (2016). Open source, social activism and” necessary trade-offs” in the digital enclosure: A case study of platform co-operative, loomio.org. *tripleC: Communication, Capitalism and Critique. Open Access Journal for a Global Sustainable Information Society*, 14(2):413–427. [70](#)
- Jakesch, M., Bhat, A., Buschek, D., Zalmanson, L., and Naaman, M. (2023). Co-writing with opinionated language models affects users’ views. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–15. [225](#)
- Jennings, F. J., Suzuki, V. P., and Hubbard, A. (2021). Social media and democracy: Fostering political deliberation and participation. *Western Journal of Communication*, 85(2):147–167. [18](#)
- Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., and Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38. [123](#)
- Jiménez-Aleixandre, M. P. and Puig, B. (2012). Argumentation, evidence evaluation and critical thinking. *Second international handbook of science education*, pages 1001–1015. [180](#)
- Jonsson, M. E. and Åström, J. (2014). The challenges for online deliberation research: A literature review. *International Journal of E-Politics (IJEP)*, 5(1):1–15. [3](#)
- Kahneman, D. (2011). *Thinking, fast and slow*. macmillan. [6](#)



- Kay, R. H. (2006). Developing a comprehensive metric for assessing discussion board effectiveness. *British Journal of Educational Technology*, 37(5):761–783. [59](#), [61](#), [102](#), [103](#)
- Kirschner, P. A., Buckingham-Shum, S. J., and Carr, C. S. (2012). *Visualizing argumentation: Software tools for collaborative and educational sense-making*. Springer Science Business Media. [63](#)
- Klein, G., Kim, Y., Deng, Y., Senellart, J., and Rush, A. M. (2017). Opennmt: Open-source toolkit for neural machine translation. [89](#)
- Klein, M. (2010). Using metrics to enable large-scale deliberation. In *Collective intelligence in organizations: A workshop of the ACM Group 2010 Conference*, pages 103–233. [229](#)
- Klein, M. (2011). How to harvest collective wisdom on complex problems: An introduction to the mit deliberatorium. *Center for Collective Intelligence working paper*. [79](#)
- Klein, M. (2012). Enabling large-scale deliberation using attention-mediation metrics. *Computer Supported Cooperative Work (CSCW)*, 21:449–473. [1](#), [2](#), [227](#)
- Klein, M. (2015). A critical review of crowd-scale online deliberation technologies. *Available at SSRN 2652888*. [5](#), [63](#)
- Knoth, P., Herrmannova, D., Cancellieri, M., Anastasiou, L., Pontika, N., Pearce, S., Gyawali, B., and Pride, D. (2023). Core: a global aggregation service for open access papers. *Scientific Data*, 10(1):366. [231](#)
- Koehn, P. (2009). *Statistical machine translation*. Cambridge University Press. [85](#)
- Krueger, R. A. (2014). *Focus groups: A practical guide for applied research*. Sage publications. [20](#)
- Kunz, W. and Rittel, H. (1970a). *Issues as Elements of Information Systems*. Number no. 131 in California. University. Center for Planning and Development Research.

- Working paper, no. 131. Institute of Urban and Regional Development, University of California. [79](#)
- Kunz, W. and Rittel, H. W. (1970b). *Issues as elements of information systems*, volume 131. Citeseer. [121](#)
- Kwak, N., Lee, C. W., and Kim, J. H. (2005). An mcdm model for media selection in the dual consumer/industrial market. *European Journal of Operational Research*, 166(1):255–265. [52](#)
- Lashkari, F., Bagheri, E., and Ghorbani, A. A. (2019). Neural embedding-based indices for semantic search. *Information Processing & Management*, 56(3):733–755. [150](#)
- Lauscher, A., Glavaš, G., and Eckert, K. (2018a). Arguminsci: A tool for analyzing argumentation and rhetorical aspects in scientific writing. In *Proceedings of the 5th Workshop on Argument Mining*, pages 22–28. [141](#), [226](#)
- Lauscher, A., Glavaš, G., and Ponzetto, S. P. (2018b). An argument-annotated corpus of scientific publications. Association for Computational Linguistics. [141](#)
- Lawrence, J. and Reed, C. (2020). Argument mining: A survey. *Computational Linguistics*, 45(4):765–818. [83](#)
- Lazer, D. M., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., et al. (2018). The science of fake news. *Science*, 359(6380):1094–1096. [3](#)
- Lear, J. (1986). *Aristotle and logical theory*. CUP Archive. [76](#)
- Lee, L. W., Dabirian, A., McCarthy, I. P., and Kietzmann, J. (2020). Making sense of text: artificial intelligence-enabled content analysis. *European Journal of Marketing*. [6](#)
- Lemaire, B., Mandin, S., Dessus, P., and Denhière, G. (2005). Computational cognitive models of summarization assessment skills. In *Proceedings of the 27th*

- Annual Meeting of the Cognitive Science Society (CogSci'2005)*, pages 1266–1271. 119
- Lewiński, M. and Aakhus, M. (2014). Argumentative polylogues in a dialectical framework: A methodological inquiry. *Argumentation*, 28:161–185. 79
- Lewiński, M. et al. (2010). *Internet political discussion forums as an argumentative activity type: A pragma-dialectical analysis of online forms of strategic manoeuvring in reacting critically*. SicSat Amsterdam. 180
- Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*. 92, 122
- Liang, G.-S. and Wang, M.-J. J. (1994). Personnel selection using fuzzy mcdm algorithm. *European journal of operational research*, 78(1):22–33. 52
- Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., Zhang, Y., Narayanan, D., Wu, Y., Kumar, A., Newman, B., Yuan, B., Yan, B., Zhang, C., Cosgrove, C., Manning, C. D., R'e, C., Acosta-Navas, D., Hudson, D. A., Zelikman, E., Durmus, E., Ladhak, F., Rong, F., Ren, H., Yao, H., Wang, J., Santhanam, K., Orr, L. J., Zheng, L., Yuksekgonul, M., Suzgun, M., Kim, N. S., Guha, N., Chatterji, N. S., Khattab, O., Henderson, P., Huang, Q., Chi, R., Xie, S. M., Santurkar, S., Ganguli, S., Hashimoto, T., Icard, T. F., Zhang, T., Chaudhary, V., Wang, W., Li, X., Mai, Y., Zhang, Y., and Koreeda, Y. (2022). Holistic evaluation of language models. *Annals of the New York Academy of Sciences*. 226
- Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81. 94
- Lippi, M. and Torroni, P. (2016a). Argumentation mining: State of the art and emerging trends. *ACM Transactions on Internet Technology (TOIT)*, 16(2):1–25. 6, 83

- Lippi, M. and Torroni, P. (2016b). Margot: A web server for argumentation mining. *Expert Systems with Applications*, 65:292–303. [83](#), [102](#), [104](#)
- Liu, F., Flanigan, J., Thomson, S., Sadeh, N., and Smith, N. A. (2018). Toward abstractive summarization using semantic representations. *arXiv preprint arXiv:1805.10399*. [85](#)
- Liu, Y. and Liu, P. (2021). Simcls: A simple framework for contrastive learning of abstractive summarization. *arXiv preprint arXiv:2106.01890*. [92](#)
- Llinas, J. (2014). A survey of automated methods for sensemaking support. *Next-Generation Analyst II*, 9122:47–59. [5](#)
- Lugini, L. and Litman, D. (2021). Contextual argument component classification for class discussions. *arXiv preprint arXiv:2102.10290*. [83](#)
- Ma, S., Sun, X., Li, W., Li, S., Li, W., and Ren, X. (2018). Query and output: Generating words by querying distributed word representations for paraphrase generation. *arXiv preprint arXiv:1803.01465*. [84](#)
- Mahyar, N., Liu, W., Xiao, S., Browne, J., Yang, M., and Dow, S. P. (2017). Consensus: Visualizing points of disagreement for multi-criteria collaborative decision making. In *Companion of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, pages 17–20. [52](#)
- Malone, T. W., Laubacher, R., and Dellarocas, C. (2009). Harnessing crowds: Mapping the genome of collective intelligence. [52](#)
- Mani, I., Klein, G., House, D., Hirschman, L., Firmin, T., and Sundheim, B. (2002). Summac: a text summarization evaluation. *Natural Language Engineering*, 8(1):43–68. [224](#)
- Mao, M., Heinzl, T., Klemba, K., and Li, Q. (2009). A sensemaking-based information foraging and summarization system in business environments. In *CSREA EE*, pages 384–390. [119](#)

- Matias, J. N., Johnson, A., Boesel, W. E., Keegan, B., Friedman, J., and DeTar, C. (2015). Reporting, reviewing, and responding to harassment on twitter. *arXiv preprint arXiv:1505.03359*. [2](#)
- Mehdad, Y., Carenini, G., and Ng, R. (2014). Abstractive summarization of spoken and written conversations based on phrasal queries. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1220–1230. [119](#)
- Mercer, N. (2000). *Words and minds: How we use language to think together*. Psychology Press. [6](#)
- Mercier, H. and Sperber, D. (2011). Why do humans reason? arguments for an argumentative theory. *Behavioral and brain sciences*, 34(2):57–74. [75](#)
- Mihalcea, R. (2005). Language independent extractive summarization. In *ACL*, volume 5, pages 49–52. [85](#)
- Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., and Miller, K. J. (1990). Introduction to wordnet: An on-line lexical database. *International journal of lexicography*, 3(4):235–244. [84](#)
- Mishra, A., Laha, A., Sankaranarayanan, K., Jain, P., and Krishnan, S. (2019). Storytelling from structured data and knowledge graphs: An nlg perspective. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 43–48. [85](#)
- Mitchell, M. and Krakauer, D. C. (2023). The debate over understanding in ai’s large language models. *Proceedings of the National Academy of Sciences*, 120(13):e2215907120. [126](#)
- Mochales, R. and Moens, M.-F. (2011). Argumentation mining. *Artificial Intelligence and Law*, 19(1):1–22. [82](#), [83](#)

- Moens, M.-F., Boiy, E., Palau, R. M., and Reed, C. (2007). Automatic detection of arguments in legal texts. In *Proceedings of the 11th international conference on Artificial intelligence and law*, pages 225–230. [83](#)
- Mohammed, D. (2016). Goals in argumentation: A proposal for the analysis and evaluation of public political arguments. *Argumentation*, 30(3):221–245. [76](#)
- Monk, A. and Watts, L. (2000). Peripheral participation in video-mediated communication. *International Journal of Human-Computer Studies*, 52(5):933–958. [201](#)
- Muhlberger, P. and Stromer-Galley, J. (2009). Automated and hand-coded measurement of deliberative quality in online policy discussions. In *Proceedings of the 10th Annual International Conference on Digital Government Research: Social Networks: Making Connections between Citizens, Data and Government*, pages 35–41. [59](#)
- Mullick, D., Fyshe, A., and Ghanem, B. (2022). Discriminative models can still outperform generative models in aspect based sentiment analysis. *arXiv preprint arXiv:2206.02892*. [226](#)
- Murphy, E. (2004). Recognising and promoting collaboration in an online asynchronous discussion. *British Journal of Educational Technology*, 35(4):421–431. [63](#)
- Musi, E. and Aakhus, M. (2018). Discovering argumentative patterns in energy polylogues: A macroscope for argument mining. *Argumentation*, 32:397–430. [80](#)
- Naess, A. and Hannay, A. (1968). Communication and argument. elements of applied semantics. [59](#)
- Nazari, N. and Mahdavi, M. (2019). A survey on automatic text summarization. *Journal of AI and Data Mining*, 7(1):121–135. [91](#)

- Nguyen, T. T., Hui, P.-M., Harper, F. M., Terveen, L., and Konstan, J. A. (2014). Exploring the filter bubble: the effect of using recommender systems on content diversity. In *Proceedings of the 23rd international conference on World wide web*, pages 677–686. [6](#)
- Nickerson, R. S. (1999). How we know—and sometimes misjudge—what others know: Imputing one’s own knowledge to others. *Psychological bulletin*, 125(6):737. [220](#)
- Nisioi, S., Štajner, S., Ponzetto, S. P., and Dinu, L. P. (2017). Exploring neural text simplification models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 85–91. [85](#)
- Novikova, J., Dušek, O., Curry, A. C., and Rieser, V. (2017). Why we need new evaluation metrics for nlg. *arXiv preprint arXiv:1707.06875*. [95](#)
- O’Brien, H. L. and Toms, E. G. (2010). The development and evaluation of a survey to measure user engagement. *Journal of the American Society for Information Science and Technology*, 61(1):50–69. [201](#)
- O’Leary, K., Wobbrock, J. O., and Riskin, E. A. (2013). Q-methodology as a research and design tool for hci. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1941–1950. [179](#)
- OpenAI (2023). Gpt-4 technical report. [86](#)
- Palau, R. M. and Moens, M.-F. (2009). Argumentation mining: the detection, classification and structure of arguments in text. In *Proceedings of the 12th international conference on artificial intelligence and law*, pages 98–107. [82](#)
- Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318. [94](#)
- Papushina, I., Maksimenkova, O., and Kolomiets, A. (2016). Digital educational mind maps: A computer supported collaborative learning practice on marketing

- master program. In *International Conference on Interactive Collaborative Learning*, pages 17–30. Springer. [68](#)
- Pariser, E. (2011). *The filter bubble: What the Internet is hiding from you*. penguin UK. [2](#)
- Parveen, D., Mesgar, M., and Strube, M. (2016). Generating coherent summaries of scientific articles using coherence patterns. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 772–783. [119](#)
- Pasquetto, I. V., Swire-Thompson, B., Amazeen, M. A., Benevenuto, F., Brashier, N. M., Bond, R. M., Bozarth, L. C., Budak, C., Ecker, U. K., Fazio, L. K., et al. (2020). Tackling misinformation: What researchers could do with social media data. *The Harvard Kennedy School Misinformation Review*. [20](#)
- Paul, R. and Elder, L. (2006). Critical thinking: The nature of critical and creative thought. *Journal of Developmental Education*, 30(2):34. [6](#)
- Pearce, L. (2005). The value of public participation during a hazard, impact, risk and vulnerability (hirv) analysis. *Mitigation and Adaptation Strategies for Global Change*, 10(3):411–441. [52](#)
- Peppers, K., Tuunanen, T., Rothenberger, M. A., and Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of management information systems*, 24(3):45–77. [12](#)
- Pelteret, M. and Ophoff, J. (2016). A review of information privacy and its importance to consumers and organizations. *Informing Science*, 19:277–301. [38](#)
- Perez, O. (2008). Complexity, information overload, and online deliberation. *ISJLP*, 5:43. [3](#)
- Perret-Clermont, A.-N., Schär, R., Greco, S., Convertini, J., Iannaccone, A., and Rocci, A. (2019). Shifting from a monological to a dialogical perspective on



- children's argumentation. *Argumentation in Actual Practice: Topical studies about argumentative discourse in context*, 17. [80](#)
- Pinto, R. C. (2010). The uses of argument in communicative contexts. *Argumentation*, 24(2):227–252. [75](#)
- Pirolli, P. and Card, S. (1999). Information foraging. *Psychological review*, 106(4):643. [5](#), [120](#)
- Pirolli, P. and Card, S. (2005). The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. In *Proceedings of international conference on intelligence analysis*, volume 5, pages 2–4. McLean, VA, USA. [3](#), [55](#), [56](#), [98](#), [120](#)
- Pontis, S. and Blandford, A. (2015). Understanding “influence:” an exploratory study of academics' processes of knowledge construction through iterative and interactive information seeking. *Journal of the Association for Information Science and Technology*, 66(8):1576–1593. [137](#)
- Ponto, J. (2015). Understanding and evaluating survey research. *Journal of the advanced practitioner in oncology*, 6(2):168. [21](#)
- Potter, S. (2006). *Doing postgraduate research*, volume 13. Sage. [76](#)
- Pu, P., Chen, L., and Hu, R. (2011). A user-centric evaluation framework for recommender systems. In *Proceedings of the fifth ACM conference on Recommender systems*, pages 157–164. [156](#)
- Quinto, I., Iandoli, L., De Liddo, A., et al. (2021). Designing online collaboration for the individual and social good: A collective argumentation approach. In *AMCIS*. [4](#)
- Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al. (2018). Improving language understanding by generative pre-training. [87](#)

- Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. [86](#), [87](#)
- Reed, C. and Walton, D. (2003). Argumentation schemes in argument-as-process and argument-as-product. [76](#)
- Reimers, N., Schiller, B., Beck, T., Daxenberger, J., Stab, C., and Gurevych, I. (2019). Classification and clustering of arguments with contextualized word embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 567–578, Florence, Italy. Association for Computational Linguistics. [83](#)
- Reiter, E. and Dale, R. (1997). Building applied natural language generation systems. *Natural Language Engineering*, 3(1):57–87. [85](#)
- Ribeiro, B., Gonçalves, C., Pereira, F., Pereira, G., Santos, J., Gonçalves, R., and Au-Yong-Oliveira, M. (2019). Digital bubbles: living in accordance with personalized seclusions and their effect on critical thinking. In *New Knowledge in Information Systems and Technologies: Volume 3*, pages 463–471. Springer. [3](#)
- Ricci, F., Rokach, L., and Shapira, B. (2010). Introduction to recommender systems handbook. In *Recommender systems handbook*, pages 1–35. Springer. [156](#)
- Rieger, A., Shaheen, Q.-U.-A., Sierra, C., Theune, M., and Tintarev, N. (2022). Towards healthy engagement with online debates: An investigation of debate summaries and personalized persuasive suggestions. In *Adjunct Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization*, pages 192–199. [224](#)
- Rittel, H. (1972). On the planning crisis: Systems analysis of the ‘first and second generations’. *Bedriftskonomen*, 8:390–396. [75](#)
- Rittel, H. W. and Webber, M. M. (1973). Dilemmas in a general theory of planning. *Policy sciences*, 4(2):155–169. [52](#)

- Roberts, N. (1997). Public deliberation: An alternative approach to crafting policy and setting direction. *Public Administration Review*, pages 124–132. [62](#)
- Rogers, Y. and Marshall, P. (2017). Research in the wild. *Synthesis Lectures on Human-Centered Informatics*, 10:i–97. [229](#)
- Rose, J. and Øystein Sæbø (2010). Designing deliberation systems. *The Information Society*, 26(3):228–240. [19](#)
- Rosenberg, A. and Hirschberg, J. (2007). V-measure: A conditional entropy-based external cluster evaluation measure. In *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*, pages 410–420. [154](#)
- Rush, A. M., Chopra, S., and Weston, J. (2015). A neural attention model for abstractive sentence summarization. *arXiv preprint arXiv:1509.00685*. [85](#)
- Russell, D. M., Stefik, M. J., Pirolli, P., and Card, S. K. (1993). The cost structure of sensemaking. In *Proceedings of the INTERACT’93 and CHI’93 conference on Human factors in computing systems*, pages 269–276. ACM. [54](#), [55](#)
- Sahlgren, M. (2006). *The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces*. PhD thesis. [84](#)
- Sanchan, N., Aker, A., and Bontcheva, K. (2017). Automatic summarization of online debates. *arXiv preprint arXiv:1708.04587*. [117](#)
- Sanchan, N., Bontcheva, K., and Aker, A. (2016). Understanding human preferences for summary designs in online debates domain. *Polibits*, (54):79–85. [224](#)
- Sasaki, C., Oyama, T., Oshima, C., Kajihara, S., and Nakayama, K. (2021). Online discussion support system with facilitation function. *International Journal of Advanced Computer Science and Applications*, 12(8). [228](#)

- Schaeffer, M. and Yilmaz, S. (2008). *Strengthening local government budgeting and accountability*. The World Bank. [52](#)
- Schick, A. G., Gordon, L. A., and Haka, S. (1990). Information overload: A temporal approach. *Accounting, organizations and society*, 15(3):199–220. [227](#)
- Schneider, D. C., Voigt, C., and Betz, G. (2007). Argunet: A software tool for collaborative argumentation analysis and research. In *7th Workshop on Computational Models of Natural Argument (CMNA VII)*. [4](#)
- Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. (2017). Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*. [91](#)
- See, A., Liu, P. J., and Manning, C. D. (2017). Get to the point: Summarization with pointer-generator networks. *arXiv preprint arXiv:1704.04368*. [85](#), [91](#), [101](#), [104](#)
- Shane, P. M. (2004). *Democracy Online: The Prospects for Political Renewal Through the Internet*. Psychology Press. Google-Books-ID: vtFu8KgEB68C. [51](#)
- Shani, G. and Gunawardana, A. (2011). Evaluating recommendation systems. *Recommender systems handbook*, pages 257–297. [156](#)
- Shu, K., Sliva, A., Wang, S., Tang, J., and Liu, H. (2017). Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter*, 19(1):22–36. [6](#)
- Siddharthan, A. (2006). Syntactic simplification and text cohesion. *Research on Language and Computation*, 4(1):77–109. [84](#)
- Singh, M. P. (2002). The pragmatic web. *IEEE Internet Computing*, 6(03):4–5. [232](#)
- Smiraglia, R. P. (2014). *Cultural synergy in information institutions*. Springer. [52](#)
- Stab, C., Miller, T., Schiller, B., Rai, P., and Gurevych, I. (2018). Cross-topic argument mining from heterogeneous sources. In *Proceedings of the 2018 Conference*

- on Empirical Methods in Natural Language Processing*, pages 3664–3674, Brussels, Belgium. Association for Computational Linguistics. [148](#)
- Stromer-Galley, J. (2007). Measuring deliberation’s content: A coding scheme. *Journal of public deliberation*, 3(1):12. [3](#), [58](#), [61](#), [71](#), [103](#)
- Stunkel, L., Benson, M., McLellan, L., Sinaii, N., Bedarida, G., Emanuel, E., and Grady, C. (2010). Comprehension and informed consent: assessing the effect of a short consent form. *Irb*, 32(4):1. [120](#)
- Sun, C., Qiu, X., Xu, Y., and Huang, X. (2019). How to fine-tune bert for text classification? In *Chinese Computational Linguistics: 18th China National Conference, CCL 2019, Kunming, China, October 18–20, 2019, Proceedings 18*, pages 194–206. Springer. [143](#)
- Sunstein, C. R. (2018). *# Republic: Divided democracy in the age of social media*. Princeton University Press. [2](#)
- Suran, S., Pattanaik, V., and Draheim, D. (2020). Frameworks for collective intelligence: A systematic literature review. *ACM Comput. Surv.*, 53(1). [19](#)
- Suran, S., Pattanaik, V., Kurvers, R., Hallin, C. A., De Liddo, A., Krimmer, R., and Draheim, D. (2022). Building global societies on collective intelligence: Challenges and opportunities. *Digit. Gov.: Res. Pract.*, 3(4). [18](#), [20](#)
- Surowiecki, J. (2005). *The wisdom of crowds*. Anchor. [2](#), [52](#)
- Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27. [86](#), [92](#)
- Taber, C. S. and Lodge, M. (2006). Motivated skepticism in the evaluation of political beliefs. *American journal of political science*, 50(3):755–769. [59](#)
- Talboom, S. and Pierson, J. (2013). Understanding trust within online discussion boards: trust formation in the absence of reputation systems. In *Trust Management*

- VII: 7th IFIP WG 11.11 International Conference, IFIPTM 2013, Malaga, Spain, June 3-7, 2013. *Proceedings 7*, pages 83–99. Springer. [181](#)
- Tam, D., Mascarenhas, A., Zhang, S., Kwan, S., Bansal, M., and Raffel, C. (2022). Evaluating the factual consistency of large language models through summarization. *arXiv preprint arXiv:2211.08412*. [225](#)
- Tampe, I., Mendoza, M., and Milios, E. (2022). Neural abstractive unsupervised summarization of online news discussions. In *Proceedings of SAI Intelligent Systems Conference*, pages 822–841. Springer. [119](#)
- Tarrant, M., Branscombe, N. R., Warner, R. H., and Weston, D. (2012). Social identity and perceptions of torture: It’s moral when we do it. *Journal of Experimental Social Psychology*, 48(2):513–518. [220](#)
- Tay, Y., Dehghani, M., Bahri, D., and Metzler, D. (2022). Efficient transformers: A survey. *ACM Computing Surveys*, 55(6):1–28. [92](#)
- Telford, J. K. (2007). A brief introduction to design of experiments. *Johns Hopkins apl technical digest*, 27(3):224–232. [194](#)
- Teufel, S., Siddharthan, A., and Batchelor, C. (2009). Towards discipline-independent argumentative zoning: evidence from chemistry and computational linguistics. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3-Volume 3*, pages 1493–1502. Association for Computational Linguistics. [82](#)
- Thomas, D. R. (2006). A general inductive approach for analyzing qualitative evaluation data. *American journal of evaluation*, 27(2):237–246. [25](#)
- Thornton, L., Knowles, B., and Blair, G. (2022). The alchemy of trust: The creative act of designing trustworthy socio-technical systems. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1387–1398. [181](#)

- Tigelaar, A. S., Den Akker, R. O., and Hiemstra, D. (2010). Automatic summarisation of discussion fora. *Natural Language Engineering*, 16(2):161–192. [117](#), [119](#)
- Torrey, L. and Shavlik, J. (2010). Transfer learning. In *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*, pages 242–264. IGI global. [143](#)
- Toulmin, S. E. (2003). *The uses of argument*. Cambridge university press. [xix](#), [77](#), [78](#), [142](#)
- Towne, W. B. and Herbsleb, J. D. (2012). Design considerations for online deliberation systems. *Journal of Information Technology & Politics*, 9(1):97–115. [223](#)
- Treem, J. W. and Leonardi, P. M. (2013). Social media use in organizations: Exploring the affordances of visibility, editability, persistence, and association. *Annals of the International Communication Association*, 36(1):143–189. [19](#)
- van der Lee, C., Gatt, A., van Miltenburg, E., and Krahmer, E. (2021). Human evaluation of automatically generated text: Current trends and best practice guidelines. *Computer Speech & Language*, 67:101151. [93](#), [95](#)
- Van Eemeren, F. H. and Grootendorst, R. (2016). *Argumentation, communication, and fallacies: A pragma-dialectical perspective*. Routledge. [59](#)
- Van Eemeren, F. H., Grootendorst, R., and Kruiger, T. (2019). *Handbook of argumentation theory: A critical survey of classical backgrounds and modern studies*, volume 7. Walter de Gruyter GmbH and Co KG. [75](#)
- van Hillegersberg, J. and Koenen, S. (2016). Adoption of web-based group decision support systems: experiences from the field and future developments. *International journal of information systems and project management*, 4(1):49–64. [4](#)
- Vargha, A. and Delaney, H. D. (1998). The kruskal-wallis test and stochastic homogeneity. *Journal of Educational and behavioral Statistics*, 23(2):170–192. [106](#)

- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30. [86](#), [87](#), [92](#)
- Velikanov, C. and Prosser, A. (2017). Mass online deliberation in participatory policy-making—part i. In *Beyond Bureaucracy*, pages 209–234. Springer. [62](#)
- Vosoughi, S., Roy, D., and Aral, S. (2018). The spread of true and false news online. *science*, 359(6380):1146–1151. [19](#)
- Wachsmuth, H., Potthast, M., Al Khatib, K., Ajjour, Y., Puschmann, J., Qu, J., Dorsch, J., Morari, V., Bevendorff, J., and Stein, B. (2017). Building an argument search engine for the web. In *Proceedings of the 4th Workshop on Argument Mining*, pages 49–59. [148](#)
- Walton, D. (2009). Argumentation theory: A very short introduction. In *Argumentation in artificial intelligence*, pages 1–22. Springer. [76](#)
- Walton, D., Reed, C., and Macagno, F. (2008). *Argumentation schemes*. Cambridge University Press. [76](#), [142](#)
- Wang, G., Gill, K., Mohanlal, M., Zheng, H., and Zhao, B. Y. (2013). Wisdom in the social crowd: an analysis of quora. In *Proceedings of the 22nd international conference on World Wide Web*, pages 1341–1352. [70](#)
- Wang, X. and Wan Wart, M. (2007). When public participation in administration leads to trust: An empirical assessment of managers’ perceptions. *Public administration review*, 67(2):265–278. [52](#)
- Watanabe, W. M., Junior, A. C., Uzêda, V. R., Fortes, R. P. d. M., Pardo, T. A. S., and Aluísio, S. M. (2009). Facilita: reading assistance for low-literacy readers. In *Proceedings of the 27th ACM international conference on Design of communication*, pages 29–36. [84](#)



- Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., et al. (2022). Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*. [90](#)
- Weick, K. E. (1995). *Sensemaking in organizations*, volume 3. Sage. [53](#), [73](#), [103](#)
- Weick, K. E., Sutcliffe, K. M., and Obstfeld, D. (2005). Organizing and the process of sensemaking. *Organization science*, 16(4):409–421. [54](#)
- Welleck, S., Kulikov, I., Roller, S., Dinan, E., Cho, K., and Weston, J. (2019). Neural text generation with unlikelihood training. *arXiv preprint arXiv:1908.04319*. [88](#)
- Whiting, L. S. (2008). Semi-structured interviews: guidance for novice researchers. *Nursing Standard (through 2013)*, 22(23):35. [21](#)
- Whittaker, S., Terveen, L., Hill, W., and Cherny, L. (1998). The dynamics of mass interaction. In *Proceedings of the 1998 ACM conference on Computer supported cooperative work*, pages 257–264. [201](#)
- Widyassari, A. P., Rustad, S., Shidik, G. F., Noersasongko, E., Syukur, A., Affandy, A., et al. (2022). Review of automatic text summarization techniques & methods. *Journal of King Saud University-Computer and Information Sciences*, 34(4):1029–1046. [92](#), [118](#)
- Willig, C. and Rogers, W. S. (2017). *The SAGE handbook of qualitative research in psychology*. Sage. [25](#)
- Wilson, V. (2012). Research methods: interviews. *Evidence Based Library and Information Practice*, 7(2):96–98. [22](#)
- Woolley, A. W., Chabris, C. F., Pentland, A., Hashmi, N., and Malone, T. W. (2010). Evidence for a collective intelligence factor in the performance of human groups. *science*, 330(6004):686–688. [228](#)
- Wright, S. and Street, J. (2007). Democracy, deliberation and design: the case of online discussion forums. *New media & society*, 9(5):849–869. [34](#)

- Wu, J., Ouyang, L., Ziegler, D. M., Stiennon, N., Lowe, R., Leike, J., and Christiano, P. (2021). Recursively summarizing books with human feedback. *arXiv preprint arXiv:2109.10862*. [119](#)
- Xu, J., Gan, Z., Cheng, Y., and Liu, J. (2019). Discourse-aware neural extractive model for text summarization. *arXiv preprint arXiv:1910.14142*. [91](#)
- Yaffe, P. (2020). The secrets of writing a truly useful executive summary. *Ubiquity*, 2020(November):1–4. [119](#)
- Yan, Y., Qi, W., Gong, Y., Liu, D., Duan, N., Chen, J., Zhang, R., and Zhou, M. (2020). Prophetnet: Predicting future n-gram for sequence-to-sequence pre-training. *arXiv preprint arXiv:2001.04063*. [101](#)
- Yang, A., Liu, K., Liu, J., Lyu, Y., and Li, S. (2018). Adaptations of rouge and bleu to better evaluate machine reading comprehension task. *arXiv preprint arXiv:1806.03578*. [95](#)
- Yang, J., Jin, H., Tang, R., Han, X., Feng, Q., Jiang, H., Yin, B., and Hu, X. (2023). Harnessing the power of llms in practice: A survey on chatgpt and beyond. [87](#)
- Zhang, F., Yuan, N. J., Lian, D., Xie, X., and Ma, W.-Y. (2016a). Collaborative knowledge base embedding for recommender systems. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 353–362. [149](#)
- Zhang, J., Zhao, Y., Saleh, M., and Liu, P. (2020). Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR. [92](#), [101](#)
- Zhang, R., Li, W., Liu, N., and Gao, D. (2016b). Coherent narrative summarization with a cognitive model. *Computer Speech & Language*, 35:134–160. [120](#)
- Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. (2019). Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*. [95](#)

- Zhao, Y. and Karypis, G. (2001). Criterion functions for document clustering: Experiments and analysis. [154](#)
- Zhao, Y., Khalman, M., Joshi, R., Narayan, S., Saleh, M., and Liu, P. J. (2022). Calibrating sequence likelihood improves conditional language generation. *arXiv preprint arXiv:2210.00045*. [92](#)
- Zhao, Z., Cohen, S. B., and Webber, B. (2020). Reducing quantity hallucinations in abstractive summarization. *arXiv preprint arXiv:2009.13312*. [137](#)
- Zhong, M., Liu, P., Wang, D., Qiu, X., and Huang, X. (2019). Searching for effective neural extractive summarization: What works and what’s next. *arXiv preprint arXiv:1907.03491*. [91](#)
- Zhou, L. and Hovy, E. H. (2006). On the summarization of dynamically introduced information: Online discussions and blogs. In *AAAI Spring symposium: Computational approaches to analyzing weblogs*, page 237. [119](#)
- Zhu, M., Bergner, Y., Zhang, Y., Baker, R., Wang, Y., and Paquette, L. (2016). Longitudinal engagement, performance, and social connectivity: a mooc case study using exponential random graph models. In *Proceedings of the Sixth International Conference on Learning Analytics and Knowledge*, pages 223–230. [59](#)
- Ziegler, D. M., Stiennon, N., Wu, J., Brown, T. B., Radford, A., Amodei, D., Christiano, P., and Irving, G. (2020). Fine-tuning language models from human preferences. [90](#)
- Zimmerman, J. and Forlizzi, J. (2014). Research through design in hci. *Ways of Knowing in HCI*, pages 167–189. [12](#), [14](#)
- Zimmerman, J., Forlizzi, J., and Evenson, S. (2007). Research through design as a method for interaction design research in hci. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 493–502. [12](#)

# Appendices

## A Study I - Interview Sketch

**Participant's background / views** 1. What do you think online discussion tools are for? What is your experience with online discussion technologies? What tools/platforms have you used for such purposes in the past, as an individual or in your organisation?

2. How important do you think they are to make good decisions? What do you know about current practice for making “big” group decisions? (in your organisation or in your general knowledge)

3. Current decision-making processes Do you have any need to realise group decision in your organisational role, or in your current research or everyday life? (or is just decisions behind closed doors or in small groups usually enough?)

Do you identify any problems in the current processes to carry on public consultation and deliberation in your organisation/research or general experience? Do you have any established protocol or process for collective deliberation?

Can you give us some practical examples of some cases in which you personally needed to consult or involve a larger group of people in a decision? How did you carry on the decision? Did you use any technology to support it?

**Issues Exploration - Technological aspirations** Explore strengths and weaknesses, expectations and fears on: Privacy and accountability online polarisation, division and conflict vs consensus building quality of discussion and collective deci-

sions

What sort of tools, if realised, would you bring into the process to help you?

## B Study I - Codebook

abuse of democracy	DM protocols	online tools not suitable for DM
abusing the system	DM unsuitability	only results not reasoning explained
accountability	dynamic consent	organisational structure
accountability = responsibility	e-democracy	participation in DM
accountability concern	each media has it problems	participatory
accountability definition	ease of engagement	participatory budgeting
accountability dynamic consent	ease of use	personal experience
accountability example	email repurposing	personal understanding
accountability importance	email ubiquity	personla use
accountability in traditional media	emotional discussion	pessimistic
accountability is a multi party concept	emotional motivation	philosophy and technology disconnect
accountability party	engagement	plethora of tools
accountability party\Idea	engagement feature	polarization
accountability ties with privacy	equal participation	policing by the community
added value to beat inertia	equal participation dilemma	preparation use
adoption	established huge competition	presenting facts problems
adoption difficulty	established market	privacy concern
AI practicality	expectations from discussion	privacy importance
anonymity trust reputation	experience	professional use
area of interest	expertise	professional use concern
argument by authority	facilitated decision making	prominent figures attract audience
argument mapping dislike	facilitated discussion	protocol for meetings

argumentation motivation	facts and evidence	public vs restricted participation
argumentative discussion	facts by authority	purpose over functionality
argumentative discussion use case	fake news	reason for non engagement
aversion of deep discussion	familiarity	reflection
avoiding discussion	flexibility	reputatio manipulation mechanisms
big audience requirement	forced choice of tool	reputation
blind followers	forums use	reputation system
censorship	free speech or regularization dilemma	resources sharing technologies
challenge	generic use	responsibility
chat analysis	good deliberation impact	responsibility of the platform
cognition overload	good discussion can happen in not good platforms	sandbox exercise
collective decision making importance	good DM comes with sacrifice	self bias
collective decision making use case	good in one aspect only	self promotion
commercial solutions	good justification	self reflection
communication	guidance	seniority matters
communication\Idea	hard facts still intertreable	sensemaking factors
communication use	hate speech	sensitive of content shared
complexity should be an obstacle	highly qualified audience	sequenced decision making
computer supported DM enactment	human barriers not technology	shallow explanation
concern for repurposing my information	impact of discussion	shift of paradigm
confirmation bias awareness	impersonal	simplicity
confirmation bias discussion	importance in DM	SM ambiguity

consultation in DM	impossibility of democracy	SM do not really do DM
context is important	improvement	SM ease of use
context is resources - time	improvement via technology	SM engagement
contrast to face-to-face deliberation	inclusiveness	SM for DM critique
critical thinking	indirect	SM is media
daily workflow use	indirect democracy	SM purpose
data analytics	infancy stage	SM uses
data use intention not clear	information cascading	small group discussion is feasible
decision making use	information overflow	social media fear
deliberation context	informed consent	solution tried but failed
deliberation systems	innovative platforms	source of evidence is important
democracy workflow protocol	interaction complexity	source reliability
design choice	interaction problems	specific solutions
design guideline	interactive conversation	spontaneous use
desirable feature	interface design difficulty	strong opinions
desirable function	interface difficulty	structured discussion
different functions	interface importance	suitability - appropriateness
different level knowledge people	involvement	switching context is difficult
different levels of entry in discussion	issues with sequential systems	tech platform implementing physical procedures
different paradigms of discussion	knowing how your data is used	technology appropriateness
different terminology	lack of confidence	technology complementary
different tools different functions	lack of regulation	technology improvement
different tools for different groups	lack of time	technology is not the solution



different tools for different use cases	learning curve	technology per se is not the solution
different use	legal comprehension	technology scepticism
difficult to implement	levels of discussion	technology shift difficulty
difficulty	liability	technology speed up
difficulty of DM protocols	lotteries mechanisms	technology support
difficulty to adopt	making sense by explaining	technology use
difficulty to follow long discussions	maturity of face-to-face deliberation	tool ambiguity
difficulty to use	media literacy	tools limitations
discussion connection to problem solving	meticulous	tools overlapping
discussion needs to be positioned at first	motivation for choosing a tool	tools overload
discussion problem	multiple channels of communication	TOS obsufication
discussion quality elements	narrative connotation	traceability
discussion tool example	necessity	traceability example
discussion tool feature	need of summarisation	tracking use
discussion tool function	need of training	traditional tech can be very effective
discussion tools ambiguity	nice!	trolling
discussion tools function	no moderation	trolling battling
discussion tools importance	no willingness to deliberate	trust
discussion tools purpose	non argumentative discussion	trust of not misuse
discussion tools to support DM	not fancy but useful	trustworthiness
discussion tools type	not multi-faceted	unbiased discussion
discussion tools use	not responding conversation	understanding process and content
discussionm problem	online discussion difference to natural dialog	use case example

---

dissemination	online discussion in not like physical discussion	virtual proximity
distinction of DM DT Delib- eration	online discussion reflection of physical	virtual replicates real world
DM is part of Organisation culture	online tool problem	voting mechanisms
DM open feasibility	online tools complementary use	what is discussion tools
DM protocol complexity	online tools is a reflection of the physical world	who is responsible

## C Study II - Questionnaire

**Understanding online discussion task** This survey examines online discussion systems and how people understand and comprehend opinions and arguments about a debated issue.

**Task** You are asked to read a debate about the topic: “Higher education should be publicly funded.”

Head out to the external page: <https://mturk-survey-01.web.app/b/debate/overall/-LwIc1ttIsZE5DwgPHy/st/1/user/61961>

Click join debate and you will be presented with a simple user interface where an issue is debated with pro and con arguments.

Devote at least 5 minutes to read and comprehend the debate

### Questions

Could you please summarise, in 5 to 10 lines max, your position regarding this debate, and while explaining your position please link/cite as much as you can the discussion you just read (by citing/reusing other people positions and evidence expressed in the debate) A good summary has as many links to existing evidence and ideas in the debate as possible. (required)

Copy-Paste the statement (argument) that is overall the strongest in your opinion (regardless if you agree or not). (required)

Provide a brief justification why you think this is the strongest (required)

**The next set of questions is about the usefulness of the summary box on**

**top of the discussion**

After your exploration of the debate, how much do you agree to each of the following statements:

The summary box presented on the top of the page helped me to accomplish the task more quickly (required)

1. Strongly disagree 2. Disagree 3. Neutral 4. Agree 5. Strongly agree

The summary box presented on the top of the page did not helped me to finish the task of locating strong arguments faster (required)

1. Strongly disagree 2. Disagree 3. Neutral 4. Agree 5. Strongly agree

The summary box presented on the top of the page was overall useful (required)

1. Strongly disagree 2. Disagree 3. Neutral 4. Agree 5. Strongly agree

It was easy to use the summary box presented on top of the page (required)

1. Strongly disagree 2. Disagree 3. Neutral 4. Agree 5. Strongly agree

I was not able to flexibly interact with it (required)

1. Strongly disagree 2. Disagree 3. Neutral 4. Agree 5. Strongly agree

I found the summary box on top of the page overall easy to use (required)

1. Strongly disagree 2. Disagree 3. Neutral 4. Agree 5. Strongly agree

**The next set of questions is about your understanding of the debate**

I was able to reflect on the debated question(required)

1. Strongly disagree 2. Disagree 3. Neutral 4. Agree 5. Strongly agree

I was provided with unexpected insights on what is the question and what are the

main arguments for and against.(required)

1. Strongly disagree 2. Disagree 3. Neutral 4. Agree 5. Strongly agree

I was not able to focus on different aspects of the debate(required)

1. Strongly disagree 2. Disagree 3. Neutral 4. Agree 5. Strongly agree

I was able to find structure in the information provided in this debate and find a way to organise it (required)

1. Strongly disagree 2. Disagree 3. Neutral 4. Agree 5. Strongly agree

I was not able to identify the main points raised in this debate (required)

1. Strongly disagree 2. Disagree 3. Neutral 4. Agree 5. Strongly agree

I was able to assess facts and evidence provided in this debate (required)

1. Strongly disagree 2. Disagree 3. Neutral 4. Agree 5. Strongly agree

I was able to distinguish between different people's claims(required)

1. Strongly disagree 2. Disagree 3. Neutral 4. Agree 5. Strongly agree

I was not able to assess my initial assumptions about this debate (required)

1. Strongly disagree 2. Disagree 3. Neutral 4. Agree 5. Strongly agree

Some initial assumptions I had about this question changed (required)

1. Strongly disagree 2. Disagree 3. Neutral 4. Agree 5. Strongly agree

**The next set of questions is about your opinion on the quality of the debate**

The messages were presented with clarity, i.e. they were clear (required)

1. Strongly disagree 2. Disagree 3. Neutral 4. Agree 5. Strongly agree

The messages quality was poor (required)

1. Strongly disagree 2. Disagree 3. Neutral 4. Agree 5. Strongly agree

I was presented with new knowledge (required)

1. Strongly disagree 2. Disagree 3. Neutral 4. Agree 5. Strongly agree

External resources were used (required)

1. Strongly disagree 2. Disagree 3. Neutral 4. Agree 5. Strongly agree

With the arguments presented, I believe there is not going to be a resolution in the debate (required)

1. Strongly disagree 2. Disagree 3. Neutral 4. Agree 5. Strongly agree

Opinions presented were reasoned, expressed as claims with evidence that can be confirmed or denied (required)

1. Strongly disagree 2. Disagree 3. Neutral 4. Agree 5. Strongly agree

The evidences or references provided were trustworthy (required)

1. Strongly disagree 2. Disagree 3. Neutral 4. Agree 5. Strongly agree

Opinions and views present were overlapping and indistinct (required)

1. Strongly disagree 2. Disagree 3. Neutral 4. Agree 5. Strongly agree

There was no opinion/argument dominating the overall discussion (required)

1. Strongly disagree 2. Disagree 3. Neutral 4. Agree 5. Strongly agree

Arguments stayed on topic and did not diverged from the defined issue (required)

1. Strongly disagree 2. Disagree 3. Neutral 4. Agree 5. Strongly agree

There was engagement among participants of the debate, i.e. arguments were responding to other arguments and not just isolated opinions. (required)

1. Strongly disagree 2. Disagree 3. Neutral 4. Agree 5. Strongly agree

## D Study II - Focus-Group Interview Sketch

**Purpose** This focus group aims at exploring the value and applicability of computationally powered summarised reports when deployed in online deliberation systems and their effect on participants' sensemaking (their capability to make sense of the debate) and their perception of the overall quality of the debate.

**Engagement questions (probe)** [ expected duration: 15 min ]

How *familiar* you are with online discussion/deliberation/debate platforms?

Have you used any in the past for *making a decision*? In what *scenario* would you resort to such platforms?

What would be your *platform of preference*? E.g. Quora for generic question? Stackoverflow for programming? Trello (!) to just sort pro and con and make a decision?

Are you aware of more *deliberation-specific* platforms? In particular have you ever come across an argumentation tool? If yes, what did you use it for? (consider.it, pol.is, coggle, kialo, loomio, debatehub) ?

Short break up here. Tell them what they are expected to do in this focus group and then introduce the focus of discussion.

**Exploration questions (follow-up)** [ expected duration : 45 min ]

Are you *familiar* with the topic of whether “Higher education should be publicly funded” Can you suspect why is a highly *controversial* topic? Do you understand what *sides* exist in this topic and what are the *arguments* used by each side?

10 minute TASK - 3min each interface

Explore the debate in the following links (same content differently presented):

Try to approach each like you were *naturally* jumping into a discussion.

Also, try to read not only the summary but the *whole* discussion

Regarding the summary box present on top of the discussion:

Have you read the summary at the top of the debate? Did you read it only once?

Ask *theme exploration* questions like:



- when did you read it?
- At the beginning?
- Did you go back to the summary while you were half way or while reading the debate?
- How often?

Was it useful? What was useful for? (Then while they answer you can make more specific questions on understanding, simplicity etc: Did it help to understand the points mentioned above faster? With more ease?)

Do you think the summaries were useful to make sense of the debate? (What exactly did you find out? What was the most interesting insight that it gave you? Did you have any “Aha!” moments?)

Do you think the summaries helped you to assess the overall quality of the debate? Do you think they affected your perception of the quality of the debate? In what way? Did you like this or not and why?

---

**One by one examination** Let's look at them one by one.

What are the *pros* of the first summary? What did you like of it?

What are the *disadvantages* of the first? What you did not like?

Where would you like to “see” this summary on the screen? Would you put it anywhere else? (How would you position/*design* such an “assistance” summary?)

Are there any missing features? Things you would like to add?

What do you think would help you to read the summary during the navigation of the debate?

Would you even want/like that? (What elements do you expect to be present to assist you while navigating a debate?) ask simple, specific UI/UX questions

Repeat these for the other 2 summaries

## Comparative examination

Now let's have a discussion about the 3 summaries in comparison.

Comparing the 3 different types of summary presented, which one did you like the most and **why**? (then you can ask exploratory questions on how do you evaluate each? What are the pros and cons of each solution?)

Which summary was *easier* to explore? (Did you manage to *interact* with it? Did you have to spend time to *familiarise* with it (esp. In case 3)

Which one do you think helped you more to *understand* the debate?

Which one do you think helped you more to *assess* the overall *quality* of the debate?

**Exit questions** [ expected duration : 5 min ]

Is there anything I haven't asked that I should have?