

University of Central Florida

**STARS**

---

Graduate Thesis and Dissertation 2023-2024

---

2024

## **Advancing Policy Insights: Opinion Data Analysis and Discourse Structuring Using LLMs**

Aaditya Bhatia

*University of Central Florida*

Find similar works at: <https://stars.library.ucf.edu/etd2023>

University of Central Florida Libraries <http://library.ucf.edu>

This Masters Thesis (Open Access) is brought to you for free and open access by STARS. It has been accepted for inclusion in Graduate Thesis and Dissertation 2023-2024 by an authorized administrator of STARS. For more information, please contact [STARS@ucf.edu](mailto:STARS@ucf.edu).

---

### **STARS Citation**

Bhatia, Aaditya, "Advancing Policy Insights: Opinion Data Analysis and Discourse Structuring Using LLMs" (2024). *Graduate Thesis and Dissertation 2023-2024*. 204.

<https://stars.library.ucf.edu/etd2023/204>

ADVANCING POLICY INSIGHTS: OPINION DATA ANALYSIS AND DISCOURSE  
STRUCTURING USING LLMS

by

AADITYA “SONNY” BHATIA  
B.Sc. San Jose State University, 2009

A thesis submitted in partial fulfilment of the requirements  
for the degree of Master of Science  
in the Department of Computer Science  
in the College of Engineering and Computer Science  
at the University of Central Florida  
Orlando, Florida

Spring Term  
2024

Major Professor: Gita Sukthankar

© 2024 Aaditya Bhatia

## **ABSTRACT**

The growing volume of opinion data presents a significant challenge for policymakers striving to distill public sentiment into actionable decisions. This study aims to explore the capability of large language models (LLMs) to synthesize public opinion data into coherent policy recommendations. We specifically leverage Mistral 7B and Mixtral 8x7B models for text generation and have developed an architecture to process vast amounts of unstructured information, integrate diverse viewpoints, and extract actionable insights aligned with public opinion. Using a retrospective data analysis of the Polis platform debates published by the Computational Democracy Project, this study examines multiple datasets that span local and national issues with 1600 statements posted and voted upon by over 3400 participants. Through content moderation, topic modeling, semantic structure extraction, insight generation, and argument mapping, we dissect and interpret the comments, leveraging voting data and LLMs for both quantitative and qualitative insights. A key contribution of this thesis is demonstrating how LLM reasoning techniques can enhance content moderation. Our content moderation approach shows performance improvements using comment deconstruction in multi-class classification, underscoring the trade-offs between moderation strategies and emphasizing a balance between precision and cautious moderation. Using comment clustering, we establish a hierarchy of semantically linked topics, facilitating an understanding of thematic structures and the generation of actionable insights. The generated argument maps visually represent the relationships between topics and insights, and highlight popular opinions. Future work will leverage advanced semantic extraction and reasoning techniques to enhance insight generation further. We also plan to generalize our techniques to other major discussion platforms, including Kialo. Our work contributes to the understanding of using LLMs for policymaking and offers a novel approach to structuring complex debates and translating public opinion into actionable policy insights.

This thesis is dedicated to my wife, Navdeep, whose patience, unwavering support, and brilliant ideas during our extensive brainstorming were instrumental in making this research possible.

## **ACKNOWLEDGMENTS**

I am immensely grateful to my advisor, Dr. Gita Sukthankar, for her guidance and support throughout this research journey. Her expertise and critical insights have been invaluable in shaping both the direction and the success of this work. I am also very thankful to my committee members, Dr. Mary Jean Amon and Dr. Ozlem Garibay, for their thoughtful and highly constructive feedback that enriched my research efforts. Special gratitude is owed to Jack Park, a long-time friend whose mentorship has profoundly influenced my thinking and approach for over a decade and a half.

## TABLE OF CONTENTS

LIST OF FIGURES . . . . .	xi
LIST OF TABLES . . . . .	xv
CHAPTER 1: INTRODUCTION . . . . .	1
Online Debate Platforms and Public Opinion . . . . .	2
Context . . . . .	4
Problem Statement . . . . .	5
Research Objective . . . . .	5
Scope and Limitations . . . . .	7
CHAPTER 2: LITERATURE REVIEW . . . . .	9
Evolution of Collective Intelligence Technologies . . . . .	9
Theoretical Foundation and Technological Enablers . . . . .	11
Knowledge Graphs and Semantic Structure . . . . .	12
Role of Argument Mapping in Deliberative Democracy . . . . .	14
Semantic Analysis and Topic Modeling . . . . .	15

Software Agents and Large Language Models . . . . .	17
Online Deliberation Platforms . . . . .	18
Analysis of current approaches . . . . .	20
Identification of gaps in the literature . . . . .	21
Challenges and Ethical Considerations . . . . .	22
<b>CHAPTER 3: METHODOLOGY . . . . .</b>	<b>24</b>
Datasets . . . . .	25
Data Extraction . . . . .	27
Embeddings . . . . .	27
Comment Moderation . . . . .	28
Topic Modeling . . . . .	33
Statements and Embeddings . . . . .	34
Dimensionality Reduction . . . . .	35
Clustering . . . . .	36
Vectorizer and c-TF-IDF . . . . .	36
Topic Representation . . . . .	37
Topic Labeling . . . . .	37



Argument Generation . . . . .	38
Argument-Statement Association . . . . .	39
Argument Mapping . . . . .	41
<b>CHAPTER 4: RESULTS . . . . .</b>	<b>44</b>
Comment Moderation as a Multi-class Classification Task . . . . .	44
Evaluation Metrics . . . . .	45
Accuracy . . . . .	45
F1 Score . . . . .	46
Unsure Rate . . . . .	46
False Positive Rate . . . . .	47
Overall Results . . . . .	47
Baseline Three-class Classification . . . . .	49
Three-class Classification with Examples . . . . .	50
Three-class Classification with Second-Thought Reasoning . . . . .	51
Three-class Classification with Second-Thought Reasoning and Examples . . . . .	51
Seven-class Classification . . . . .	51
Seven-class Classification with Thought Statement . . . . .	51

Seven-class Classification with Comment Deconstruction and Chain-of-Thought Reasoning . . . . .	52
Seven-class Classification with Comment Deconstruction . . . . .	52
Three-class Classification with Comment Deconstruction and Chain-of-Thought Reasoning . . . . .	52
Comparative Analysis of Results . . . . .	52
Challenges in Comment Moderation . . . . .	54
Topic Modeling . . . . .	55
Embedding and Preprocessing . . . . .	56
Dimensionality Reduction . . . . .	56
Clustering . . . . .	62
Argument Generation . . . . .	65
Argument-Statement Association . . . . .	67
Argument Mapping . . . . .	71
CHAPTER 5: CONCLUSION . . . . .	73
Summary of Findings . . . . .	73
Theoretical Implications . . . . .	74
Practical Implications . . . . .	75

Limitations and Challenges . . . . . 76

Future Research Directions . . . . . 78

    Semantic Extraction and Reasoning in Earlier Stages . . . . . 78

    Connections Across Topics . . . . . 79

    More Platforms . . . . . 79

Final Contemplation . . . . . 80

LIST OF REFERENCES . . . . . 81

## LIST OF FIGURES

Figure 3.1: Data processing pipeline illustrating the flow of information. Upon importing user-generated comments, we calculate embeddings and experiment with various moderation techniques. Topic modeling identifies areas of interest, which provide structure for argument generation and mapping. Scoring those arguments yields actionable insights. . . . .	24
Figure 3.2: Topic modeling involves several algorithms chained together that identify clusters of user-generated statements and help describe those hot topics. . . .	34
Figure 3.3: Argument generation pipeline that demonstrates how user-generated comments are used to synthesize actionable insights, which in turn are scored and sorted to generate argument maps. . . . .	38
Figure 4.1: Performance of comment moderation techniques across different datasets. The left graph shows the results for the <i>american-assembly.bowling-green</i> dataset, while the right graph shows the results for the <i>scoop-hivemind.biodiversity</i> dataset. The x-axis represents the various experimental configurations, while the y-axis shows the accuracy, f1 score, unsure rate, and false positive rate. . . .	49
Figure 4.2: UMAP <i>n_neighbors</i> parameter. Smaller neighborhoods capture finer details while sacrificing global structure. Large values overly emphasize global structure but lose the finer details. Optimum values of 4, 8, and 16 show a reasonably balanced spread of data points. . . . .	58

Figure 4.3: UMAP *min\_dist* parameter. These figures show the impact of varying minimum distance between each projected data point. Higher values result in visually appealing charts but destroy the structural information that enables clustering. Lower values are better in the context of clustering data points based on similarity with 0.0 yielding the best results. . . . . 59

Figure 4.4: UMAP *n\_components* parameter. These graphs show a 2D projection of embedding vectors that had already been reduced to *n\_components* dimensions. While we run these vectors through UMAP twice, we expect that the changes made to the global structure of these embeddings from the first run will be appropriately captured in the final 2D projection. The colors of individual data points are still computed using PCA on the original embeddings, maintaining relevance to the global structure of the data. . . . . 60

Figure 4.5: UMAP *metrics* parameter. All metrics appear to yield similar results. We select *cosine* to ensure that vectors are compared based on their semantic similarity but not magnitude. . . . . 61

Figure 4.6: Final UMAP Projections of Statement Embeddings for the *american-assembly* dataset. The first one shows the data points in groups somewhat consistent with their colors, which shows a balance between global structure and local relationships. The second image is a network connectivity graph that shows the distance between various clusters. A high-density core with several offshoots is consistent with a scale-free network and is considered a strong candidate for clustering. A hairball-type projection would have implied a poorly projected set of vectors. . . . . 62

Figure 4.7: Distribution of statements colored by topic projected in a 2D plane using PCA. The upper image shows the statement distribution before the outlier assignment while the lower one is updated after the assignment process. While the gray data points disappear in the process, each topic is inherently changed and represented with a different set of labels that include the newly assigned statements. . . . . 64

Figure 4.8: The top graph shows the distribution of statements across various topics immediately upon clustering. The first bar represents a large proportion of statements that do not belong to a topic and are considered outliers. The bottom graph shows a relatively even topic distribution after assigning each outlier to a topic based on its embedding proximity to the topic centroids. . . . . 65

Figure 4.9: Topic Hierarchy dendrogram depicting inter-topic distance among centroids of statement clusters. Topics in proximity have common themes and can be safely merged into a larger topic. This hierarchical map can be used to organize the argument map in more nuanced layers and aid in further analysis. 66

Figure 4.10 The set of instructions and prompts used to generate arguments from comments using the LLM. Text highlighted in green is produced by the language model, while everything else is inserted into the LLM context by us programmatically. . . . . 67

Figure 4.11 The set of instructions and prompts used to characterize the relationship between arguments and comments using the LLM. Text highlighted in green is produced by the language model, while everything else is inserted into the LLM context by us programmatically. . . . . 69

Figure 4.12 Distribution of argument-comment mappings with a SUPPORT relationship.

The x-axis shows individual arguments clustered by topic while the y-axis indicates the number of SUPPORT links between that argument and corresponding statements. The tallest bars represent arguments with the most supporting statements but not necessarily the most agreed-upon positions. The initial topics have more statements and thus a larger number of arguments. This distribution skews towards the initial and less controversial topics that are more heavily represented through statements and votes. . . . . 70

Figure 4.13 Argument maps generated using *american-assembly.bowling-green* dataset,

covering two different topics. The upper map addresses healthcare and the opioid epidemic, while the lower focuses on community enrichment programs. The middle row in each map contains LLM-generated insights, while the lowest row has human-authored statements that support one or more of the arguments. We use Polis voting data to estimate the support for each generated argument. . . . . 71

## LIST OF TABLES

Table 3.1: Summary of Experimental Configurations for Comment Moderation. The first four and the last one are framed as three-class classification tasks, while the remainder allow the classification of each statement into one of the seven categories. We use different techniques, including chain-of-thought reasoning and comment deconstruction to comprehensively analyze the model performance in identifying spam or irrelevant statements. Due to the combinatorial space presented by these parameters, we limit our selection to these nine configurations. . . . . 31

Table 4.1: Comment moderation results for the *american-assembly.bowling-green* dataset. 48

Table 4.2: Comment moderation results for the *scoop-hivemind.biodiversity* dataset. . . 48

Table 4.3: Comment moderation results for the *scoop-hivemind.freshwater* dataset. . . . 48

Table 4.4: Comment moderation results for the *scoop-hivemind.taxes* dataset. . . . . 50

Table 4.5: Comment moderation results for the *scoop-hivemind.affordable-housing* dataset. 50

Table 4.6: UMAP parameters tested along with their considered and optimal values. . . 57

Table 4.7: The number of argument-statement pairs listed by each dataset, and the number of pairs that are labeled as SUPPORT or REFUTE. The model tends to classify very few links as REFUTE. . . . . 69



## CHAPTER 1: INTRODUCTION

While democratic processes have strived to interpret public opinion for millennia, the advent of information technology fundamentally changed the nature of real-time communication. Empowering citizens to voice their support or opposition and reach consensus brings the promise of strengthening democratic institutions through fair and equitable participation. Leaders and policymakers have an opportunity to leverage consensus in a massive conglomeration of human intelligence toward solving complex problems. However, human interaction on social platforms is producing data at an ever-growing rate, and it takes extensive effort to analyze that data to produce useful results. The integration of artificial intelligence tools into public discourse and policy decision-making carries the potential to revolutionize the way we comprehend and engage with societal issues. By harnessing the power of AI, we can augment human intellect, supercharge collaboration, provide transparency in processes, enable real-time synthesis of diverse opinions from multiple information streams and provide timely insights that are crucial for informed policymaking.

Collective intelligence is the enhanced capacity and intelligence that emerges when a diverse group of individuals works together to brainstorm, collaborate, and innovate to solve wicked problems and make decisions [11]. Collective efforts enable faster ideation by pooling and coordinating the creative and analytical skills of several brains, ideally allowing each individual to bring their best abilities into the mix, leading to a more efficient and deliberate understanding of the situation and decision-making. In policymaking, the concept of collective intelligence enables tapping into the “wisdom of the masses” and allows the public to participate in the decision-making process [12]. A diverse set of perspectives allows for identifying and analyzing problems with precision. A population that includes expert citizens aids in a deeper understanding of issues, along with the discovery of innovative and creative solutions. This meaningfully includes the citizens in the democratic processes and have their opinions heard, as policymakers are able to make more

informed and democratic decisions based on public opinion. Additionally, the ability to quantify and visualize the data helps explain the sentiment of the population, which is a powerful tool for any public leader. As our world evolves, public opinion takes a new shape and collective intelligence tools present the opportunity to rapidly adapt the new policies to those changes and thoroughly understand the risks involved. The resulting consensus from engaging with the public at a personal level and allowing voices to be heard grants legitimacy to the involved organizations and strengthens democracy. This is particularly important in democratic systems where legitimacy and public acceptance of decisions are crucial.

However, all the public deliberation tools require working with a massive amount of public opinion data, which tends to be chaotically unstructured, opinionated, and often not backed by evidence. This type of deliberation is also very vulnerable to misinformation campaigns by threat actors. Due to this, many existing technological solutions require significant effort for manual moderation and analysis of individual comments. In this thesis, we discuss the technological evolution that enables next-generation tools capable of collectively solving complex problems. Specifically, we explore the use of large language models (LLM) and propose a software pipeline for real-time facilitation and retrospective data analysis of historical deliberation events. We demonstrate content moderation techniques and produce argument maps that provide a visual summary to inform policy decisions.

## Online Debate Platforms and Public Opinion

Many citizens desire to actively participate in their community, local governance, and national democratic discussions and choose to express their opinions in various ways. The traditional methods such as surveys and polls, while often accurate, could not qualitatively and quantitatively capture the dynamic aspects of public viewpoints. For example, thinking about a question

from a certain perspective begins to impact an individual's opinion on the subject. Discussing the subject leads to the formation of new ideas and raises more interesting questions. This evolution of thought process is not easy to capture without real-time communication and an ability to freely express thoughts and ideas.

Over the past few decades, there has been a massive emergence of online debate platforms that revolutionized the way public opinion is captured, analyzed, and reported. Modern web-based platforms, such as Polis and Kialo [24, 27], make these tools accessible to the broad public, enabling gathering diverse perspectives and insights through real-time analysis and deliberation of complex social, economical, and societal issues. This significant technological advancement represents an opportunity to revamp democratic engagement and public policy processes. Beyond supporting public interaction, these platforms structure and analyze the massive amount of opinion data generated from those deliberations by identifying certain patterns, trends, and biases in the opinion landscape. This capacity for analysis is highly beneficial to anyone seeking to understand the state of public opinion on any current issue.

The integration of AI in processing public opinions introduces unprecedented challenges, including the potential for misuse and the propagation of biases inherent in training datasets. These biases are not merely statistical anomalies but deeply ingrained prejudices that can skew insights and reinforce existing societal inequalities, a concern that has been extensively discussed in the literature [5, 28]. A massive challenge faced by such a system is turning a massive volume of unstructured and chaotic data into meaningful insights by collecting, structuring, analyzing, and interpreting the data in a way that enables unbiased and fair analysis. Before making this system available, it is crucial to address key issues such as data privacy, content moderation, equitable representation, and the integrity of analytical methods. The establishment of transparent, fair governance and careful consideration throughout the implementation is imperative. These systems represent the latest technology that is being made available to the public and carry the potential of

either strengthening or destroying the public trust in current and future AI-based systems [31].

Therefore, it is crucial to consider the potential, impact, and limitations of these systems. This thesis research is focused on not only the technical implementation of such a system but also an exploration of engaging and informing the public, which is essential for democracy.

## Context

The global challenges and existential crises that we presently face are increasingly characterized as “wicked problems”. Conklin defines them as complex, multifaceted, and resistant to traditional problem-solving methods due to their interconnectedness and ever-changing contexts [3]. Attempting to solve a wicked problem changes the environment to the degree that it is not possible to reverse the solution and try another one. The traditional linear problem-solving methods are not capable of addressing the complexity. These unique challenges require novel approaches often discovered through the wisdom of the masses.

Public policy and executive decision-making are complex issues that meet most of Conklin’s six criteria [3], which qualifies them as wicked problems. Policymakers typically get one attempt at implementing the best solution or making the correct decision, and their decisions have long-lasting impacts, so it is in their best interest to gather and analyze as many actionable insights and diverse perspectives as possible. Optimizing such a solution would require a massive amount of human effort and computational resources.

Successful leaders make decisions by leveraging their prior knowledge and experience along with their current understanding of the information environment. Traditional top-down leadership approaches struggle as the cost of considering additional information grows exponentially. These approaches rely on the leadership team to solicit feedback, consider a diverse set of perspectives,

and balance the best interests of all involved parties. Often, these approaches often fail to capture the multidimensional complexity of wicked problems.

Engaging a broader audience, including decision-makers, stakeholders, experts, citizens etc., in dialogue results in a more diverse and inclusive set of perspectives. However, it results in a massive amount of information that requires a significant amount of manual effort to analyze and interpret. The time cost of human involvement grows exponentially as the amount of information grows, which makes it inherently difficult to hold lengthy and detailed dialogues about decision-making in most contexts. A collective intelligence system is expected to coordinate the efforts of the participants, enable fair moderation, and provide tools to sift through and synthesize vast amounts of data. Do today's technological advances present an opportunity to increase the effectiveness and reduce operating costs of such a system?

### Problem Statement

How can large language models enable us to ingest massive streams of unstructured information, incorporate diverse perspectives and distill them into actionable insights that align with public opinion?

### Research Objective

The primary research objective of this study is to explore the use of Natural Language Processing (NLP) and Large Language Models (LLMs) and related technologies in providing actionable insights into large troves of public opinion data, specifically from publicly hosted online debate forums. This study is crucial in understanding the potential of LLMs in facilitating informed policy decisions for complex problems on a global scale.

*How effectively can LLMs structure and enable access to large amounts of opinion data?*

Firstly, we aim to explore the efficacy of NLP and LLMs in structuring and providing access to massive volumes of opinion data. This involves investigating how these technologies can be leveraged for topic modeling, measuring consensus or discord among opinions, and generating executive summaries and visualizations of the opinion landscape. A critical aspect of this objective is to compare the efficiency of automated processes enabled by LLMs against traditional manual methods.

*What metrics and insights can we generate from embeddings?*

Secondly, we focus on the metrics and insights that can be derived from the embeddings created by LLMs. This involves analyzing how LLMs cluster comments and opinions, facilitate topic modeling, and identify outliers. Clustering can help gauge the strengths and weaknesses of prevailing viewpoints, and assist in moderation of comments that are irrelevant or malicious. Vector search can speed up search and help create a structure around the otherwise unstructured set of comments. We also examine the LLMs' ability to identify diverse opinions and attempt to quantify the diversity within public opinions, which is essential for understanding public sentiment.

*What are the inherent risks associated with the deployment of LLMs?*

Finally, we delve into the challenges posed by the application of LLMs in a domain that significantly impacts societal norms and democratic processes. This research aims to address concerns related to data bias, ethical implications, the potential for misinformation, and the overall integrity of AI-driven systems in shaping public policy. The question is rooted in the understanding that while LLMs offer transformative capabilities in analyzing vast, unstructured datasets, they also

carry the potential to skew public discourse or influence policy decisions in unintended ways. Addressing these risks is crucial for the responsible and equitable use of AI in public policy formulation and democratic deliberation.

### Scope and Limitations

This thesis is primarily centered on the application of NLP and LLMs in analyzing public opinion data derived from online debate platforms. This encompasses an in-depth examination of the capabilities of LLMs in topic modeling, sentiment analysis, consensus measurement, and opinion summarization. The study will focus on data from established platforms like Polis, which have been instrumental in aggregating public opinions on a range of issues.

This research is subject to certain limitations. The data from these online debate platforms can be skewed by several factors, such as internet access and technology literacy, and is unlikely to fully represent the broader public opinion. Additionally, while LLMs appear to be performing advanced cognitive tasks, their performance heavily relies on the quality of the training data, and any biases or inconsistencies can appear in the output. Besides that, LLMs are known to hallucinate when producing text in an unfamiliar context, which makes it difficult to guarantee the accuracy of the output content.

Despite these limitations, the study aims to provide valuable insights into the potential of NLP and LLMs in shaping public policy decisions and contributing to the growing body of research in AI Argumentation.

The document is organized into the following key sections. Chapter 2 presents a critical analysis of the relevant literature on crowd-scale platforms and the underlying technologies. Following the literature review, Chapter 3 describes the technical implementation of our system and the ex-

periments designed to evaluate it. Subsequently, Chapter 4 discusses and analyzes the results of our experiments based on the methodologies described earlier. Finally, Chapter 5 summarizes the significant results, discusses their broader implications and outlines potential directions for future research.



## CHAPTER 2: LITERATURE REVIEW

This literature review examines the intersection of existing programs designed for collective intelligence, online platforms for public deliberation, the use of symbolic and neural AI techniques, and their application to policymaking and complex problem-solving. The review derives insights from a range of studies, focusing on the potential and challenges of leveraging AI in public discourse. The goal is to explore how LLMs can structure, interpret, and facilitate access to vast amounts of public opinion data, ultimately contributing to more informed and democratic policy decisions.

### Evolution of Collective Intelligence Technologies

Initially, the concept of collective intelligence involved humans collaborating to solve problems and make decisions. In the realm of public policy, this was often facilitated by town hall meetings or brainstorming sessions. Over time, technological advances have increased our capacity for harnessing the creative and analytical skills of the masses. Early technological approaches were mostly focused on structured argumentation processes within small groups. Tools like gIBIS [4], developed in 1988, provided a structured approach to capturing the “issues, ideas, and arguments” in a debate, creating a more organized method for groups to discuss and solve complex problems. These technologies required certain expertise in the way of processing information and were typically limited to specific research contexts.

As the internet became more accessible in the early 2000s, the potential for human collaboration expanded drastically. Social media provided online discussion capabilities to a larger and more diverse group of participants in a largely unstructured manner. The MIT Collaboratorium project [14] and tools like Cohere [25] aimed to support collaborative deliberation at a larger scale, allow-

ing users to introduce structure and make connections between various ideas and arguments across discussions and documents. These tools began integrating visual representations of argumentation and more sophisticated user interfaces. Each of these tools was created and distributed as a standalone desktop application that required operating system support, binary distribution, and manual installation, which created a barrier for the public to explore and use them.

Around the 2010s, Web 2.0 technologies came to prominence and emphasized user-generated content, usability, interoperability, and real-time updates to web applications. Static web pages were replaced with dynamic web applications that could query data and update parts of the document without ever reloading the webpage. Tools for argumentation and collective decision-making became more user-friendly and integrated with other web services. Platforms like the MIT Deliberatorium [12] leveraged these technologies to harness the collective intelligence of large groups for complex problem-solving. Additionally, the focus shifted towards scalability and real-time collaboration.

The introduction of deep learning and neural network architectures in the 2010s revolutionized natural language processing, leading to models that could learn representations of language directly from large data corpora without extensive feature engineering. This period saw the emergence of transformer models like Generative Pre-trained Transformer (GPT), Bidirectional Encoder Representations from Transformers (BERT), and their variants, which significantly enhanced the ability of machines to understand and generate human language with a high degree of proficiency [20].

By around 2015, platforms like Polis [27] started to combine machine learning with real-time voting and commenting to identify consensus among large groups, providing a scalable way to gauge public opinion on various issues. The dramatic rise of social media also provided unprecedented data and opportunities for understanding collective behavior and opinions, along with new challenges related to misinformation and polarization.

More recently, collective intelligence technologies have started to see the integration of advanced artificial intelligence tools. In particular, LLMs can enable analysis of large volumes of text, identify trends and patterns, summarize ideas, semantically extract information, and even participate in discussions, potentially leading to more informed and effective collective decision-making. For instance, systems like “Coherent Mode” have been designed to synthesize public discourse into coherent narratives, helping to identify common ground among diverse viewpoints [19].

Some of the more recent works have specifically explored the opportunities and risks associated with LLMs in this context, emphasizing the potential of these technologies while also raising awareness of potential misuse or unintended issues [28]. One notable risk is the amplification of misinformation where LLMs, without rigorous checks, might inadvertently prioritize sensational or biased information and fuel false narratives. Additionally, there is a risk of creating echo chambers, where the technology reinforces existing biases in the data, further polarizing public opinion instead of fostering a diverse and inclusive discourse. These instances underscore the critical need for implementing mechanisms that ensure data integrity, mitigate biases and promote the responsible use of AI to safeguard against such unintended consequences. As we continue to innovate responsibly, the public trust in technology increases and the potential of collective intelligence grows dramatically, paving the way for a more informed and cohesive society.

### Theoretical Foundation and Technological Enablers

The concept of deliberative democracy, introduced by Hadfi et al. [8], offers a solid framework for collective intelligence in policymaking and problem-solving by empowering citizens through technological means. Building upon the potential role of AI, structured frameworks, and autonomous conversational agents in enriching and scaling deliberative processes, there is considerable potential to realize more informed, equitable, and participatory democratic processes. As the con-

vergence of collective intelligence and deliberative democracy evolves, the exploration of how technological tools and AI can enhance these processes becomes increasingly essential. In this section, we will explore the theoretical foundations of deliberative democracy, its challenges, and the role of technology in actualizing its potential.

Symbolic AI relies on symbols and rules for logical reasoning and is ideal for applications requiring clear explanations, such as expert systems. Neural networks mimic the human brain's structure, learning complex patterns from large datasets, which makes them effective for tasks like image recognition and natural language processing. Probabilistic models, such as Bayesian networks and Markov models, handle uncertainty by predicting future events based on statistical assumptions, which is useful in decision-making under uncertainty. Each AI approach has distinct strengths and is often integrated with others in complex systems to enhance performance.

### *Knowledge Graphs and Semantic Structure*

The concept of an Issue-Based Information System (IBIS) was developed by Horst Rittel in the 1970s as a method to capture the complexity of planning and policymaking processes [15]. The IBIS framework is designed to handle “wicked problems” characterized by high levels of complexity and uncertainty. To provide a clear visualization of ongoing discourse, Conklin and Begeman developed a graphical application, gIBIS [4], in 1988 that was capable of framing real-time discussions in IBIS terms. Conklin's core idea is to provide a structured way to document issues, ideas, and arguments related to a decision-making process. It proposed a structured approach to capture the complexity and nuances of deliberations and debates. IBIS provided a clear and navigable structure to trace, strengthen and refute arguments while avoiding repetition. Conklin further discusses how gIBIS was leveraged by several corporations and government organizations to improve group understanding and boost productivity [3].

The integration of IBIS with emerging technologies like XML (Extensible Markup Language) topic maps was a further step in creating a distributed system of knowledge. Park, in his works [22, 23], delved into the potential of XML topic maps in creating independent systems or “knowledge federations” that collected structured data. Topic maps in XML format offered a way to digitally represent the relationships and connections within complex sets of data and discussions. This approach improved the accessibility and manageability of information. But more importantly, it facilitated the federation of independent systems that shared their knowledge by linking and sharing distributed knowledge, enabling a more collaborative and comprehensive understanding of topics [23]. Symbolic AI has historically been a cornerstone in this domain by offering precision and interpretability, crucial for applications where understanding the “why” behind a decision or recommendation is as important as the outcome itself. In the case of structured conversations, expert systems have been used to discover relationships between entities and normalize data for interoperability [21]. The real potential of a knowledge federation lies within agent-based expert systems that, while conducting focused research, stumble upon previously unnoticed connections across these otherwise disjoint collections.

Knowledge graphs are graph-based data structures that link interconnected entities. Due to their tremendous scalability, knowledge graphs are increasingly being used to organize and navigate vast amounts of information. Most of the initial attempts at mapping discourse were backed by knowledge graphs due to their ability to represent complex relationships and dependencies, making them particularly valuable for tasks like semantic search, data integration, and knowledge discovery, all of which are essential for dealing with large amounts of unstructured data. These properties make knowledge graphs particularly suitable for use with symbolic AI techniques that enable the traversal of these vast trees to surface latent connections across the corpus. The work of Pan et al. [20] delves into the integration of LLMs with knowledge graphs, highlighting the potential of this combination in enhancing the understanding and interaction with complex data. This inte-

gration enables the use of LLMs to map relationships between entities and store them for future retrieval, while also providing an information lookup capability, often called retrieval augmented generation (RAG), to augment the training data and provide more relevant responses. These technologies provide a robust foundation for handling complex information structures, which opens up several new opportunities to map arguments and claims, examine individual beliefs, enable discourse, improve societal cohesion, and support coherent analysis of public opinion.

### *Role of Argument Mapping in Deliberative Democracy*

Deliberative democracy constitutes discourse that focuses on evidence and reasoning, which encourages participants to reflect on various perspectives and form an informed opinion. It assumes that through rational discourse, participants can arrive at decisions that are more legitimate and informed [31]. Such discourse can be effectively organized into an argument map, which visually represents the flow of information [11], as further explored by Klein [13], emphasizing the significance of crowd-scale deliberation in tackling complex problems. Here, we explore the interplay between technological advancements and conceptual frameworks that collectively empower and augment the deliberative democratic processes.

According to Fishkin [6], the practical realization of deliberative democracy faces several challenges. While effective deliberation can be facilitated in small groups, scaling this to a larger population becomes difficult. Maintaining inclusivity and equality in discussions is especially critical to ensure that all voices are heard and considered. Most importantly, the moderation efforts needed to structure and manage complex discussions are essential to keep them focused and productive, and this time commitment grows significantly with the number of participants.

Klein and Iandoli [14] elaborate on supporting collaborative deliberation through technological platforms that facilitate the aggregation, organization, and analysis of collective inputs, ensuring

that the deliberative process is efficient, scalable, and inclusive. These platforms also provide mechanisms for visualizing arguments and tracing the evolution of ideas, which is crucial in managing the complexity of a large-scale discussion. The development of structured and systematic frameworks for debate, as further discussed by Klein [12, 13], marks a significant stride towards organizing and harvesting collective wisdom. The MIT Deliberatorium, with its argumentation-based approach, provides a structured platform for systematic analysis and selection of solutions, ensuring that collective deliberation is both comprehensive and focused. Such structured approaches minimize redundancy and encourage clarity, essential for productive and meaningful deliberation.

### *Semantic Analysis and Topic Modeling*

Online discussion platforms generate a massive amount of data that is not readily structured for large-scale analysis. NLP gives computers the ability to parse and extract information from text through semantic parsing, role labeling and relationship extraction. This enables computers to identify themes and patterns in the given text, providing a foundation for higher-level tasks such as discourse analysis, topic modeling, and argument mining. Latent Dirichlet Allocation (LDA) and Non-negative Matrix Factorization (NMF) are distinct approaches for topic modeling to discover latent topics within text [1]. LDA, a probabilistic model, assumes documents are mixtures of topics and iteratively updates distributions, leading to varying results with multiple runs. In contrast, NMF is a matrix factorization technique that also identifies topics in text data by decomposing the term-document matrix into lower-dimensional matrices, representing potential topics and their significance within each document. While both of these methods enable thematic analysis of unstructured text, they rely on a bag-of-words technique. BERTopic is a newer algorithm that relies on transformer embeddings that take the context around the words into account [7].

In the context of this thesis, we focus on leveraging topic modeling to extract keywords from individual comments on the Polis platform. This method is instrumental in identifying and visualizing prevalent themes and opinions, crucial for understanding collective intelligence. The application of topic modeling in such online platforms, where vast numbers of participants express their views, enables the identification of core issues and perspectives, thus providing a clearer picture of collective viewpoints. Beyond revealing majority opinions, topic modeling facilitates AI argumentation by pinpointing the main points of contention or agreement among various parties, including highlighting divergent minority opinions that might otherwise remain unnoticed. This is particularly valuable for policymakers as it allows them to map out the public sentiment while remaining inclusive and aware of the diversity of opinions, thereby aiding informed decision-making that reflects the opinion of the populace. Moreover, effective topic modeling contributes to maintaining the quality of discourse by identifying and filtering out the noise in the form of irrelevant content that is distracting and does not contribute meaningfully to the discussions. This ensures that the focus remains on relevant and substantive contributions, enhancing the overall utility and integrity of the debate.

Text embeddings are numerical representations that map words or phrases into high-dimensional vectors of real numbers that capture the semantic relationships within a body of text. These embeddings allow for the quantification of textual similarities and differences by placing words or sentences with similar meanings close to each other in the vector space. This makes it possible to apply mathematical operations to our documents and derive insights from natural language. BERTopic [7] is our choice of topic modeling technique and software framework that uses text embeddings to cluster comments by semantic similarity by employing a statistical procedure, Term Frequency-Inverse Document Frequency (TF-IDF), to extract relevant keywords from large text data. This method creates embeddings for each comment using pre-trained language models and clusters them using a statistical density-based algorithm, where each cluster is considered a dis-



tinct topic. BERTopic then applies a statistical procedure, Term Frequency-Inverse Document Frequency (TF-IDF), to evaluate the importance of each word within a given set of comments. The TF part of the algorithm calculates the frequency of a word in a specific document. The IDF part measures the rarity of the word across the entire corpus, reducing the weight of common words that appear in many documents. By multiplying these two values, TF-IDF assigns a score to each word, enabling identification of those that are characteristic of a particular document while considering their commonality in the entire corpus. This method is widely used in tasks like search engine optimization, information retrieval, and text mining. BERTopic further enhances the TF-IDF method by treating all comments in a cluster as a single class, thus making the algorithm focus on clusters rather than individual comments. This class-based TF-IDF helps identify the unique topic words and extracts representative keywords for each cluster. By doing so, BERTopic streamlines the identification and analysis of individual topics, generates visual representations of these topic clusters and outliers across all comments, and organizes those topics in the hierarchy, facilitating further exploration of inter-topic relationships.

### *Software Agents and Large Language Models*

Software agents carry the potential to drastically enrich the deliberative process through argument mapping [2]. These agents not only facilitate the capture and organization of complex discussions but also ensure that related discourse across topics is interlinked and easily navigable, thereby enhancing the depth and breadth of analysis. Moreover, the role of automated facilitation agents in enhancing the deliberative process is notably illustrated in several works [8, 10]. These agents assist in structuring online discussions, extracting discussion themes, and guiding the deliberation toward coherence and productivity. The potential of conversational agents to boost deliberation in social media contexts through consensus, as envisioned by Hadfi and Ito [8], demonstrates a transformative approach towards integrating AI in public discourse, promising a more dynamic

and inclusive participation.

LLMs are very capable tools that can generate large amounts of coherent-appearing text with rudimentary reasoning capabilities. Webb et al. [30] evaluated an older model, Generative Pre-trained Transformer 3 (GPT-3), and concluded that LLMs are able to solve several analogical reasoning problems at a human-like level of performance. The advent of LLMs and their application in platforms like Polis has opened new avenues for scaling up deliberative processes. Small et al. [28] emphasized that these LLMs hold the potential to significantly enhance the efficiency of moderation and summarization in public engagements. This advancement not only extends the reach of deliberation but also enriches the quality of discourse by providing diverse perspectives and insights.

While the augmentation of democratic deliberation through conversational agents and LLMs offers a pathway to more scalable, inclusive, and informed democratic processes, the integration of such technology requires careful consideration of potential risks and limitations. Any biases or misinformation in the training data has a likelihood of being amplified in the generated text, though often subtle and hard to detect. A strong understanding and consideration of any potential pitfalls is a must, including the contextual understanding and biases of LLMs, which could influence the depth and integrity of discussions and erode public trust.

### *Online Deliberation Platforms*

In the context of public deliberation, various online platforms have emerged to facilitate discourse and argument mapping, each with distinct features and approaches. gIBIS [4] and Compendium [26] are the early forerunners in this domain. gIBIS, developed in the late 1980s, was a pioneering tool in structuring debates around complex issues. It laid the groundwork for argumentation-based rationale, providing a graphical interface to depict the interconnectedness of issues, positions, and

arguments. Compendium evolved from the gIBIS framework and extended these capabilities into a more versatile tool that integrated hypermedia and collaborative functionalities. Besides mapping arguments from multiple concurrent users, it provided a powerful visualization framework to display the relationships and dependencies, helping the group to effectively converge on the overarching idea.

More recent platforms like Polis [27] and Kialo [24] represent the evolution of online deliberation tools, incorporating advanced technologies and broader accessibility. Polis stands out for its statistical algorithms used to analyze voting patterns and visualize the opinion landscape in real time. This feature allows it to handle large-scale deliberations, making it an effective tool for gauging public opinion on diverse issues [28]. Kialo, on the other hand, is an educational tool that emphasizes structured, clear, and rational debates. It allows for the organization of arguments in a pro/con structure, fostering critical thinking and enabling users to explore the depth of each argument and its underlying premises.

The contrast between these platforms lies primarily in their approach and technological sophistication. gIBIS and Compendium, as precursors in the field, focus on issue mapping to develop a representation of discourse structure with a limited number of users. They are more oriented towards providing a visual framework for understanding complex discussions for relatively smaller groups. Polis and Kialo, however, leverage contemporary web technologies to enhance user interaction and engagement. Polis uses advanced data analytics to capture and reflect the dynamic nature of public opinion, while Kialo's structured debating format offers a more organized and digestible way of carrying complex argumentation across the Internet. This evolution to highly dynamic, interactive platforms reflects the growing complexity and requirements of online public deliberation in the digital age.

## Analysis of current approaches

Modern approaches to effective and inclusive policymaking entail gathering substantial public opinion data from diverse sources, particularly online debate platforms such as Polis and Kialo. The Computational Democracy Project developed Polis [27] as an open-source, real-time system enabled by advanced statistics to collect, analyze, and comprehend the perspectives of large groups as expressed in their own words. Polis has been employed in various contexts, demonstrating its versatility and impact. For instance, in Taiwan, Polis played a pivotal role in a national-scale deployment in 2014, contributing to a public movement that successfully influenced legislative changes. In Austria, it was used in 2022 to engage thousands in a national climate assembly, and between 2016 and 2019, a national newspaper in New Zealand leveraged what they called “Hive-Mind,” based on Polis, to solicit public feedback on several topics and assist in policy development. Also, the city of Bowling Green, Kentucky, employed Polis in 2018 to find common ground on local issues amidst political discord. This platform offers valuable datasets of public opinions and discourse, crucial for understanding collective viewpoints on various issues. Its analysis aids in deciphering complex and often polarized public stances on policy issues [28]. Polis promotes a more engaging and responsive form of policy deliberation by the published reports of these events that demonstrate scientific rigor and fairness. This approach encourages participatory engagement, where stakeholders can immediately visualize and comprehend public opinion in a meaningful way.

There is a growing emphasis on addressing algorithmic bias and ethical considerations in these methodologies. As LLMs are trained on extensive datasets, there is a risk of inheriting biases present within the data. Contemporary methods are frequently focused on mitigating these biases to ensure that the insights generated for policymaking are equitable and representative of diverse viewpoints [8]. Another critical aspect is the scalability and computational efficiency of these ap-

proaches, particularly vital when addressing global-scale issues with large number of participants. The application of LLMs and AI-driven tools needs to be weighed against computational costs and the practicality of implementing these solutions on a large scale [5].

### Identification of gaps in the literature

The integration of LLMs in the realm of policy decision-making, particularly in conjunction with the principles of collective intelligence carries significant potential yet remains underexplored. While there is a robust body of literature on the use of LLMs for a variety of applications [20] and on the fundamentals of collective intelligence [17], the effective amalgamation of these two domains in policymaking contexts is not extensively studied. The work of [28] touches upon the capabilities of LLMs to process, structure, and interpret large-scale public opinion data for informing policy decisions. However, this area is still nascent and offers ample scope for further exploration.

A notable gap exists in the comprehensive evaluation of various LLMs prompting techniques in the context of extracting and interpreting public opinion for policymaking. Although there are established benchmarks and evaluation methods for assessing LLMs in cognitive tasks, the specific performance of these models in understanding complex societal issues and their influence on policy decisions is largely uncharted territory [28]. The field of AI argumentation would greatly benefit from in-depth research into the development of specific metrics and methodologies to evaluate the effectiveness of LLM outputs for policymaking [5]. Further research should focus on measuring the reliability and validity of insights derived from public opinion data.

Additionally, there is a critical research gap in examining the computational costs and scalability associated with using LLMs for policy deliberations. This is particularly relevant when handling

massive datasets from platforms like Polis or Kialo [10]. Despite declining costs of LLMs owing to improvements in algorithms and cheaper hardware, understanding these factors is vital for the scalability and practical application of such technologies. Furthermore, research on the performance of LLMs across different languages and cultural contexts in policy deliberation is limited. The challenge of integrating and understanding diverse global perspectives in policymaking through AI-driven platforms has not been sufficiently addressed [31].

Finally, there is a scarcity of comprehensive literature that addresses the ethical implications and the necessity for transparency in the use of AI and collective intelligence in policy formulation [8]. This includes critical concerns about data privacy, representativeness, and accountability in AI-mediated decision-making processes. Addressing these issues is crucial for the responsible and equitable implementation of AI in policymaking.

### Challenges and Ethical Considerations

The integration of LLMs in policymaking presents multifaceted challenges and ethical considerations that primarily revolve around the potential for data bias in LLMs. Trained on extensive datasets, these models may inherit and inadvertently amplify inherent biases in their outputs, raising questions about the representativeness and fairness of the insights they provide, particularly in cases where decisions have widespread impacts [8]. This concern underscores the importance of ensuring equity and inclusiveness in AI-driven policymaking, emphasizing the need to make the benefits of AI accessible to all societal segments and to prevent the exacerbation of existing inequalities.

Another significant challenge lies in the opacity of many neural AI systems, especially LLMs, which complicates transparency and accountability. Tracing how these models infer and conclude

is often challenging, impeding efforts to hold them accountable for errors or biased outputs [8]. The ethical and responsible use of public data in training and deploying LLMs for policy purposes also demands careful consideration. Issues related to consent, privacy, and the proper use of data from discussion platforms are critical. A lack of transparency can rapidly erode public trust in AI-assisted decision-making processes [10].

The most significant ethical concern is the degree to which policy decisions should rely on AI-based systems since over-dependence can undermine human autonomy in decision-making processes. Maintaining human decision-makers in the loop and command is crucial to ensure that AI acts as a tool to augment, not replace, human judgment.

Deliberative democracy emerges as a promising framework to harness collective intelligence, with technology playing a pivotal role in facilitating and enhancing these processes. However, as we stand at the cusp of significant technological integration in policymaking, it is imperative to navigate these advancements with a thorough awareness of their ethical implications and practical challenges. Future research should aim to bridge these gaps by developing robust, transparent, and accountable AI systems that can effectively support and enrich the democratic process. In doing so, we can harness the full potential of AI and collective intelligence in crafting policies that are not only data-driven and efficient but also equitable, inclusive, and representative of diverse public sentiments. The path forward is one of cautious optimism, where technology is leveraged as a tool for empowerment and progress within the carefully considered bounds of ethical and democratic principles.

## CHAPTER 3: METHODOLOGY

Our experimental approach, as illustrated in Figure 3.1, leverages natural language processing and language models to systematically analyze unstructured opinion data. The process starts with the importing and pre-processing of raw statements from the downloaded dataset, which we store in a Python DataFrame. For each statement, we generate embeddings — numerical vectors that represent semantic similarities — using transformer models. In parallel, we run a comment moderation routine to experiment with our ability to identify and filter low-quality statements. The embeddings facilitate the clustering of similar data points and the identification of latent topics within the dataset through topic modeling. For each topic, we identify problems and proposed solutions, and use that to craft actionable insights, referred to as *arguments*, aimed at driving policy changes. Then we visualize these insights using argument maps, a set of discussion trees that organize these statements hierarchically where each AI-generated argument is supported by a collection of human-generated statements, enabling a deep understanding of the discourse flow and the meaningful categorization of statements. Finally, we run a scoring process to surface the insights. Our methodology is designed to uncover and identify insights with the highest levels of public acceptance by exploring the structure, topics, and dynamics of the discussion tree, thereby facilitating a

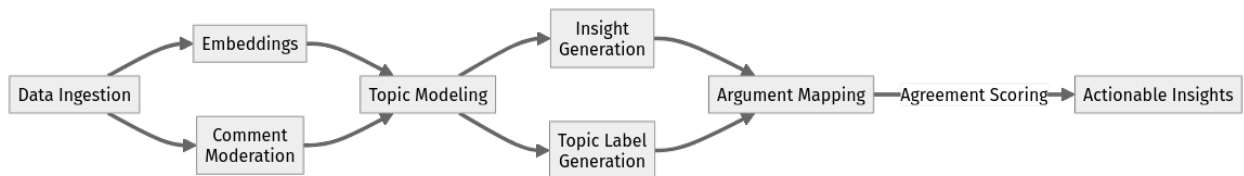


Figure 3.1: Data processing pipeline illustrating the flow of information. Upon importing user-generated comments, we calculate embeddings and experiment with various moderation techniques. Topic modeling identifies areas of interest, which provide structure for argument generation and mapping. Scoring those arguments yields actionable insights.



comprehensive understanding of the public opinion data and aiding in well-informed conclusions.

## Datasets

In our study on the impact of natural language processing on policymaking, we analyze publicly available data from Polis, a major online deliberation platform that has been used by various organizations worldwide. Polis is a highly innovative tool designed for facilitating public deliberation and collective intelligence. It stands out in democratic engagement, policy formulation, and social research due to a uniquely inventive of capturing and analyzing public opinion data. By blending social media features, survey technology, and data analytics, Polis creates a dynamically interactive space where diverse groups can express and share their views on different subjects. Its ability to cluster participants by their responses and visualize opinion groups is essential for discerning public opinion trends and pinpointing areas of agreement or division. This capability is crucial in today's digital age, fostering a more informed and participatory public dialogue by addressing the challenge of finding common ground in online discussions. Polis enriches our analysis by providing detailed comments, voting data, and participant metrics, enabling a thorough exploration of public opinion and discourse dynamics, as highlighted by Small et al. [28]. The following tables are included with each dataset:

- **Summary Statistics** includes basic descriptive information about the conversations, such as the number of voters, commenters, and the topic of conversation. This provides an overview and helps in understanding the scale and scope of the discussion.
- **Stats History** presents a timeline of various metrics like the number of votes, comments, visitors, voters, and commenters. Such temporal data can be crucial for understanding how the conversation evolved over time and identifying patterns of engagement.

- **Participants and Votes Data** is structured as a table of participants, detailing each participant's contributions in terms of comments and votes. It also includes clustering assignments (group-id) from K-means clustering, providing a basis for segmenting and analyzing different participant groups based on their interaction patterns.
- **Comments Data** contains the complete set of comments, including each comment's unique ID, author ID, body of the comment, and the number of agrees and disagrees it received. Analyzing this data can reveal which topics or opinions garnered the most consensus or dissent.
- **Votes Data** records the complete voting history, including voter ID, comment ID, timestamp, and vote (agree/disagree). This data is essential for analyzing voting patterns and understanding which comments resonated most with the participants.

We find two specific datasets from the Polis repository particularly interesting and examine them more closely: *american-assembly.bowling-green* and *scoop-hivemind.biodiversity*.

The *american-assembly.bowling-green* dataset originates from a conversation facilitated by the American Assembly in Bowling Green, Kentucky. This dataset explores local and regional perspectives on what is important to residents and probes the narrative of a divided America. It comprises 896 comments, of which 289 were moderated, leaving a substantial 607 comments for analysis. This dataset is particularly valuable for understanding community priorities and perceptions at a local level.

The *scoop-hivemind.biodiversity* dataset, on the other hand, stems from a conversation conducted by New Zealand's Public Engagement Projects (PEP) in partnership with the news outlet Scoop regarding biodiversity. It consists of 316 comments submitted by 96 people, with 162 comments moderated out, resulting in 154 remaining comments. This dataset provides insights into public

opinions and knowledge about protecting and restoring New Zealand’s biodiversity, an area of global environmental concern that drives significant policy decisions. Besides, we use three additional datasets, *scoop-hivemind.freshwater*, *scoop-hivemind.taxes*, and *scoop-hivemind.affordable-housing*, that were all collected by the same organization in different contexts.

## Data Extraction

Our thesis uses a streamlined pipeline as shown in Figure 3.1 to analyze public opinion data, combining language models and argument mapping. Starting with statement extraction and storage in Polars DataFrames, we employ transformer models for embeddings, then use dimensionality reduction and clustering to identify statement clusters based on semantic similarities. The BERTopic framework facilitates topic identification and hierarchical organization, leading to a discussion tree that maps out issues, ideas, and arguments. This structured data ultimately facilitates visual analysis through a deliberation map in Argdown syntax.

The first step of our experiment is to gather comments and their metadata from the publicly accessible Polis datasets. All of our work is conducted in Python programming language. We download and process these datasets into a versatile storage format using the Polars DataFrame library, selected for its robust performance and scalability in managing extensive datasets. Polars library uses the Apache Arrow format, which ensures smooth integration across different languages and platforms, facilitating future work without requiring specific technologies.

## Embeddings

After importing the data into our environment, we prepare it for analysis. Using the Sentence Transformers library, we calculate embeddings for each statement within the dataset. These state-

ments can range from a few words to 3-4 sentences, but the participants are required to limit each statement to one problem or idea. Considering this, we calculate an embedding for each statement based on the mean pooling method, where each dimension of the embeddings is averaged across the sentences to capture the essential semantic meaning in a unified representation. This approach is particularly effective for concise statements focused on a single idea, as it allows the pooled representation to reflect the core semantic content effectively. Each embedding is then coupled with the original statement and consolidated into the same DataFrame for later retrieval. Our approach involves employing multiple language models to generate these embeddings. Notably, we use *infloate5-mistral-7b-instruct*, a Mistral-based model, alongside *WhereIsAI/UAE-Large-V1*[16], a model known for its efficiency and small memory footprint. Both models have been consistently near the top of HuggingFace Massive Text Embedding Benchmark leaderboard specifically for clustering tasks and are openly available via the HuggingFace Hub. While selecting these models, we considered several other models including OpenAI’s *text-embedding-ada-002* and *text-embedding-3-large*, but ultimately decided to work with open models. These text embeddings play a significant role in semantically interpreting the comments, thus laying the groundwork for topic modeling and subsequent analysis.

### Comment Moderation

For our set of experiments on comment moderation, we use the Polis moderation guidelines to design prompts for various language models. This method aims to identify and label irrelevant and overly complex statements. Specifically, each statement undergoes individual analysis by the language model for classification purposes. The efficacy of this strategy is evaluated by comparing the outcomes of the language model’s spam detection against a gold standard—moderation data previously labeled by the moderators of each Polis study dataset. This comparison seeks to

ascertain the spam detection accuracy across various language models.

To manage text generation from our language models, we use the *guidance* framework originally developed by Microsoft. This library represents a unique programming paradigm that enhances control and efficiency for a language model by constraining generation through regular expressions and context-free grammar. Developers can freely add text to the context window at any point between text generations, effectively interleaving control and generation seamlessly using traditional programming paradigms such as conditionals and loops.

The Polis project has proposed the following moderation guidelines. Each organization conducting a user study ultimately decides its policy for moderation but generally follows these guidelines.

- Spam: Comments devoid of relevance to the discussion.
- Duplicative: Comments restating a previously made point.
- Complex: Comments articulating multiple ideas or problems.

Our experiment considers the effect of several variables on moderation outcomes, including:

1. Class labels used for classification by the language model, varying across experiments between a simple set (ACCEPT, UNSURE, REJECT) and a more detailed set (ACCEPT, UNSURE, SPAM, IRRELEVANT, UNPROFESSIONAL, SCOPE, COMPLEX).
2. A set of examples provided to the language model to help it better understand the task.
3. Chain-of-thought reasoning techniques that ask the model to state a pre-decision “thought” statement and potentially enhance the model’s reasoning abilities, as demonstrated by the ReAct strategy.[32]

4. Application of additional reasoning techniques, like statement deconstruction, to dissect and understand statements in depth.

We took steps to ensure that the model is not too strict in its moderation, since there is more harm done from falsely rejecting a statement than from falsely accepting it. A statement that is allowed into discussion will still be voted upon and moderated by the community. For example, a statement that adds no value to the discussion is likely to be treated as noise and see a strong disagreement. However, a statement that is incorrectly rejected is less likely to be seen again. Moderating legitimate statements discourages community participation and damages the trust people place in the system. This is also a place where inherent biases in language models can surface. Therefore, we prefer false positives to false negatives.

Our techniques do not detect duplicate statements, which is one of the Polis moderation categories. This is an inherently difficult task for a language model and better suited for other techniques such as clustering statements using their text embeddings and semantic analysis. We consider this out of scope for the experiment and defer it as future work.

We structured our methodological approach to leverage the capabilities of transformer-based language models, specifically focusing on their application in comment moderation tasks. Using the *transformers* library and specifically chat-tuned versions of Mixtral model, we defined the task using instructional prompts and user statements for the model to respond to in a chat paradigm. We carefully crafted our inputs to align with our experimental objectives by providing context, task and expected output format. The use of the *guidance* framework allowed us to retain control over most aspects of the model's output, ensuring adherence to our structured generation format.

Our experimental design consisted of nine distinct experimental configurations to evaluate the efficacy of language models in moderating user-generated content. These experiments varied

Configuration	Target Classes	Examples	Deconstruction	CoT Technique
1	3	No	No	N/A
2	3	Yes	No	N/A
3	3	No	No	Thought after rejection
4	3	Yes	No	Thought after rejection
5	7	No	No	N/A
6	7	No	No	Thought before decision
7	7	No	Yes	Thought before decision
8	7	No	Yes	N/A
9	3	No	Yes	Thought before decision

Table 3.1: Summary of Experimental Configurations for Comment Moderation. The first four and the last one are framed as three-class classification tasks, while the remainder allow the classification of each statement into one of the seven categories. We use different techniques, including chain-of-thought reasoning and comment deconstruction to comprehensively analyze the model performance in identifying spam or irrelevant statements. Due to the combinatorial space presented by these parameters, we limit our selection to these nine configurations.

from simple three-class classifications to more intricate tasks requiring reasoning and semantic extraction. The overarching goal was to understand how different moderation strategies impact the model’s moderation performance, balancing between accuracy and reasoning. A summary of these experimental configurations is provided in Table 3.1 and described below.

**Baseline Three-class Classification:** As a foundational experiment, we applied a basic instructional prompt to categorize comments into three labels: ACCEPT, UNSURE, or REJECT. When the model rejected a statement, we asked it to further select a reason that was either SPAM or COMPLEX. This experiment did not incorporate examples or reasoning techniques. For explainability, we requested the model to output an explanation for each rejection after it had made the classification.

**Three-class Classification with Examples:** Enhancing our baseline approach, this experiment introduced a set of example statements and classifications in user and assistant messages after the instructional prompt. This technique typically improves the language model’s interpretation of the

task and the expected output format through practical illustrations.

**Three-class Classification with Second-Thought Reasoning:** Implementing a “second-thought” reasoning mechanism, this variant encouraged the model to reconsider its initial REJECT decision. Since one of our goals is to minimize false positives, we allowed the model to indicate when it was unsure about rejecting a statement. In this case, we asked the model for a rationale for its initial choice and allowed for reclassification to UNSURE if the model showed uncertainty, thereby hypothetically improving results.

**Three-class Classification with Second-Thought Reasoning and Examples:** This combined experiment integrated the second-thought reasoning process with examples.

**Seven-class Classification:** We introduced a more complex seven-class classification task, enabling the model to categorize comments into a wider array of labels. Detailed instructions were provided for each label to guide the model’s decision-making, aiming for more precise and explainable moderation decisions.

**Seven-class Classification with Thought Statement:** Adding chain-of-thought reasoning, this setup required the model to clearly articulate the thought before each decision, aiming for a higher moderation accuracy and transparency despite lower efficiency.

**Seven-class Classification with Comment Deconstruction and Thought Statement:** In this approach, we introduced the concept of comment deconstruction. Focusing on the semantic content of each comment, we tasked the model with deconstructing statements to identify their requisite components and the number of ideas introduced. We require each comment to mention at least one problem or solution and no more than one unique idea. Combining it with chain-of-thought reasoning, this approach deviated from standard moderation guidelines and instead judged comments based on their relevance to the ongoing conversation.



**Seven-class Classification with Comment Deconstruction:** In this iteration, we focused on comment deconstruction without the explicit thought statement for comparative analysis, assessing the impact of this technique on moderation accuracy.

**Three-class Classification with Comment Deconstruction and Thought Statement:** In the final approach, we aimed to evaluate the model’s proficiency in semantic extraction tasks in a simpler context.

Throughout these experiments, we carefully considered the implications of false positives on the moderation outcome, with a particular emphasis on minimizing false positives to foster inclusive community discussions. These experiments contribute valuable insights into the potential of language models to streamline and enhance the comment moderation process, contributing to the ongoing development of more effective and nuanced AI moderation tools.

## Topic Modeling

To uncover the latent structures within the debate statements, we employed the BERTopic library [7], leveraging an advanced pipeline based on transformer embeddings and class-based Term Frequency-Inverse Document Frequency (TF-IDF) statistics. This process, depicted in Figure 3.2, involves several stages, starting with generating vector representations of the statements using transformer models, which allows us semantically distinguish user-generated comments. It is followed by dimensionality reduction using Uniform Manifold Approximation and Projection (UMAP) [18], which translates the high-dimensional embeddings into low-dimensional vectors that allow the identification of clusters in the dataset using Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN). The resulting clusters were then analyzed for their most representative keywords using a count vectorizer and a class-based TF-IDF tech-

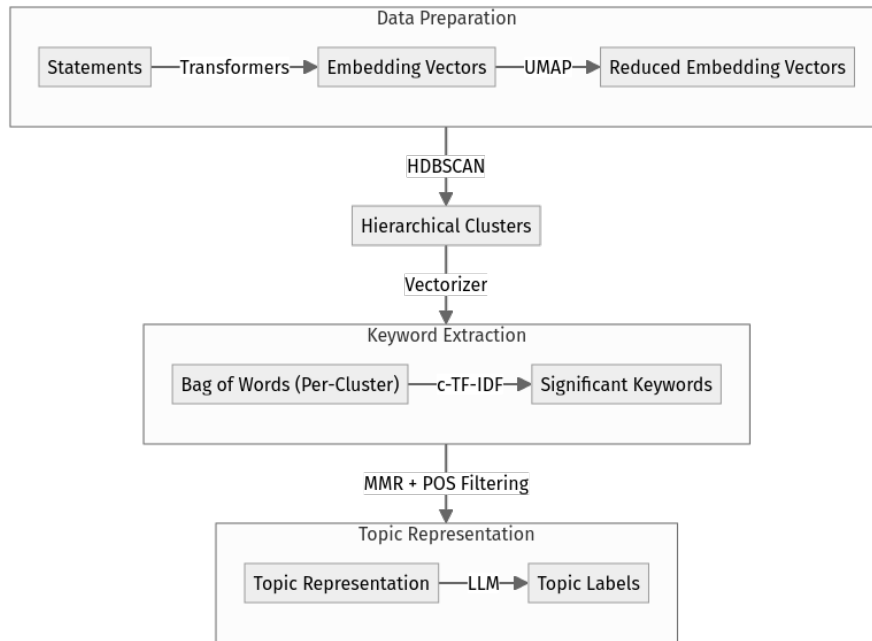


Figure 3.2: Topic modeling involves several algorithms chained together that identify clusters of user-generated statements and help describe those hot topics.

nique, further refined through part-of-speech filtering and Maximal Marginal Relevance (MMR) techniques to ensure accurate topic representation. Finally, topics were assigned descriptive labels using a language model for easy identification and interpretation.

### *Statements and Embeddings*

We use our previously stored DataFrame that includes embeddings computed during data ingestion. For this experiment, we continued to use the *WhereIsAI/UAE-Large-V1* [16] embedding model.

## *Dimensionality Reduction*

We used the Uniform Manifold Approximation and Projection (UMAP) algorithm to reduce the text embedding vector dimensions. The embeddings from *intfloat/e5-mistral-7b-instruct* model have 4096 dimensions while *WhereIsAI/UAE-Large-V1* model outputs 1024 dimensions. We employed a density-based clustering algorithm, which performs poorly in high-dimensional spaces and fails to identify any high-density regions. We translate these embeddings to a vector space lower than 100 dimensions to allow better identification of high-density clusters in the data.

UMAP uses several hyperparameters that we optimized through experimentation, including *n\_neighbors*, *min\_dist*, *n\_components*, and *metric*, with the goal of balancing local detail against global data structure, preserving local structure, and determining the optimal number of dimensions for the reduced space.

**n\_neighbors** constrains the size of local neighborhoods, balancing local details against the global structure of data. Higher values favor global structure. Clustering algorithms typically depend on the local relationships between points, so a lower value is preferred. We experimented with values [2, 4, 8, 16, 32, 64].

**min\_dist** specifies how closely the points can be located in a low-dimensional space. Higher values typically result in more evenly distributed points at the expense of local structure, which look great in visualizations. Lower values allow clumping of data points, which is not ideal for visualization but is more highly desirable for clustering. We experimented with values [0.0, 0.1, 0.25, 0.5, 0.8, 0.99] to demonstrate the effect of this parameter.

**n\_components** specifies the number of dimensions in the reduced space. Since the 2D graphs are always limited to two dimensions, we first reduce the data to our selected value of *n\_components* and then further reduce it to 2D for visualization. We experimented with values [8, 16, 32, 64, 128, 256]

to visualize the effect of this parameter. UMAP documentation recommends setting this anywhere from 10 to 100. An excessively high value might defeat the purpose of dimensionality reduction.

**metric** specifies the distance metric used to measure the distance between points in the input space. We experimented with Euclidean, Minkowski, and cosine metrics. All three metrics are valid for our data, however, the cosine metric does not take the magnitude of the vectors into account, which risks losing potentially useful information.

### *Clustering*

The reduced dimensionality text embedding vectors were clustered using HDBSCAN, a density-based clustering algorithm well-suited for data with noise and varying densities. We tuned the parameters for HDBSCAN using trial and error to optimize the number of clusters, cluster distribution, and outlier count. While we calculated certain metrics including silhouette score and relative validity, none of these metrics appeared to correlate with the quality of the clusters when generalized to other datasets. We ultimately settled on a set of parameters that produced the most coherent clusters for our dataset without further experimentation.

### *Vectorizer and c-TF-IDF*

We used a count-based vectorizer to convert the text into a matrix of token counts, followed by a term frequency-inverse document frequency (TF-IDF) transformation. The TF-IDF process evaluates the importance of a word in a body of text relative to the entire corpus. Class-based TF-IDF (c-TF-IDF)[7] assesses the importance of a word in a specific cluster, providing a more detailed understanding of the unique terms within a given cluster. To optimize the signal-to-noise ratio (SNR), we use a set of most common words in the English language, stop words, to better focus

on unique terms. We also add custom stop words from the discussion summary, as they are likely to appear in each topic without adding specific semantic value. For example, in the *american-assembly.bowling-green* dataset, we do not want the words “Bowling”, “Green”, “Warren”, or “County” to appear in any of the topics. Once we calculate the c-TF-IDF matrix, we use it to identify the most representative terms within each cluster.

### *Topic Representation*

We used Spacy to apply a part-of-speech filter to only include nouns and adjective-noun pairs in our representative keywords. We then employ Maximal Marginal Relevance (MMR) to select the most representative keywords for each topic while maintaining diversity. This step ensures the keywords accurately reflect the underlying topic, which is crucial for topic labeling and interpretation.

### *Topic Labeling*

We used a language model to generate topic labels based on the representative keywords and statements within each cluster. Specifically, we selected a quantized version of Mixtral due to its performance on reasoning tasks and used the simple instructional prompt to specify the generation task. Then we placed the representative keywords and a few example statements from each cluster into the user prompt and asked the model for a terse label and a brief heading. The generated labels are crucial for interpreting the clusters, generating actionable insights, and organizing them into an argument map.

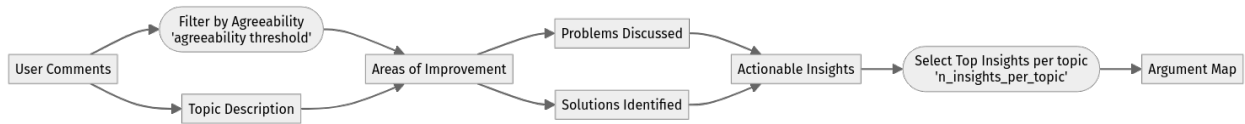


Figure 3.3: Argument generation pipeline that demonstrates how user-generated comments are used to synthesize actionable insights, which in turn are scored and sorted to generate argument maps.

### Argument Generation

This study employs a novel approach to generate actionable insights as depicted in Figure 3.3. We task a language model with synthesizing insights from clusters of semantically related statements. These insights, referred to as *arguments*, advocate for specific actions to address identified issues or implement solutions within the dataset’s context. To generate these arguments, we provide the language model with a descriptive representation of each topic and associated keywords identified during topic modeling. We also feed the model a set of filtered statements with a relatively high number of agree votes. We select from the same set of statements that were previously clustered based on their semantic similarity and placed into these unique topics. To ensure fair representation, we calculate an *agreeability* factor for each statement, defined as the ratio of agreement to the total of agreement and disagreement votes:

$$agreeability = \frac{agrees}{agrees + disagrees}$$

We consider statements based on various *agreeability* thresholds. The first is based on a raw threshold of over 0.5, which indicates that over half of the total votes received by the statement were agreeable, indicating a predominance of agreement over disagreement. The second selection

technique is based on a certain quantile of statements from each topic, selecting a fixed proportion of the statements compensating for any deviations in voting patterns across topics. For example, a specific topic about “Legalizing Marijuana” in one of our datasets was seen as controversial with lots of agree and disagree votes, reducing the *agreeability* of the statements. Trimming the bottom quantile at, say 0.2, allowed us to improve the quality of the arguments without discarding an excess of statements. The thresholds are adjustable, allowing for fine-tuning based on experimental needs. In general, this type of filtering ensures the language model focuses on popular opinion to generate insights and allows us to discard the least agreed-upon statements that presumably indicate unpopular opinions.

With these statements, we ask the model to produce a list of a specified number of areas of improvement. Then for each area, we ask the model to use the comments to identify the problems being discussed and the solutions identified in the discourse. We do this to ensure that the model does not generate its own content, but rather helps interpret the information already present in the public opinion dataset. Finally, we feed this information back to the language model to synthesize arguments.

### Argument-Statement Association

After generating coherent arguments urging immediate action, we aim to quantify the support behind each of these insights. We frame this as a classification task involving the categorization of each link between an LLM-generated argument and a human-generated comment as either SUPPORT, REFUTE, or UNRELATED. This requires a precise characterization of the semantic relationship between original statements and the newly formed arguments. To determine this, we experiment with several techniques with varying degrees of success. First, given a set of arguments from a topic, we use the language model to consider individual statements within a given

topic and select an argument that the statement most closely supports. This paradigm assumes that each statement is represented in one of the generated arguments and ignores the possibility of a statement supporting multiple arguments. Second, we use text embeddings to determine which statements are most closely aligned with each argument. Third, we present a pair of one argument and one statement at a time to a language model and have it classify the relationship from one of the following three labels: SUPPORTS, REFUTES, UNRELATED.

To calculate the argument acceptance score, we use voting data to determine which statements support or refute an argument and count the unique participants endorsing these statements. Several factors can skew our analysis and influence the balance and the interpretability and balance of data. These include chronological bias, where statements posted earlier tend to accumulate more votes, leading to a potential overestimation of their importance. The distribution of topics and arguments also plays a critical role; some topics may attract more attention and, subsequently, more statements and arguments, which can bias our analysis towards these more visible topics. Voter behavior introduces another layer of complexity, with some individuals participating more actively than others, potentially giving undue weight to the perspectives of a vocal minority. Additionally, the individual propensity to agree or disagree affects the overall picture of support, as it can dampen the nuances in the degree of agreement across different arguments. Since the argument acceptance score is used to surface the most agreed-upon insights and is expected to closely reflect public opinion, we apply normalization and analysis techniques to ensure that the insights derived reflect a balanced and comprehensive understanding of public sentiment. The task is to estimate the degree of consensus behind each of our generated arguments. While we could simply calculate the sum of votes cast on each statement towards calculating acceptance, this number would be heavily biased based on the factors discussed above. To account for these factors, we consider different techniques. Counting the number of participants who agree versus those who disagree with each argument would mitigate the influence of highly active individuals and would more accurately



estimate the consensus as opposed to the raw number of votes. We estimate an *acceptance* factor as an approximation of the ratio of people who would potentially agree with the argument given their vote on the statement. We also consider normalizing each statement’s weight by its *agreeability* that was defined earlier, thereby adjusting for the bias towards statements with a disproportionate number of votes. Another critical normalization strategy could involve considering the number of supporting statements for each argument, which helps filter out arguments that lack substantial backing, ensuring that the analysis focuses on well-supported viewpoints. However, this risks excluding fringe topics that might have been introduced late in the conversation and were not adequately covered. In this experiment, we apply a different combination of these techniques to provide a well-balanced assessment of the diverse opinions expressed in these debates, and qualitatively assess the arguments with the highest *acceptance* scores.

$$acceptance = \frac{\text{estimated number of voters that might agree}}{\text{number of voters considered in estimation}}$$

### Argument Mapping

The final part of the pipeline involves visually depicting the relationship between different entities on an argument map. This allows us to articulate and display the structure and interconnections of arguments and statements within a given context and enables a deeper understanding of the complex dynamics between human and AI-generated content. Visualization serves as a critical tool for dissecting and comprehending the relationship among various topics and uncovering underlying structures.

Argdown [29] is a simple syntax for analyzing complex argumentation structures, developed as a JavaScript library that parses the Markdown-inspired text documents and translates them into

visual argument maps. It is particularly suited for documenting, analyzing, and visualizing debates and discussions, making it a versatile tool for anyone involved in detailed argumentative analysis. We considered alternatives such as Graphviz and DebateGraph to construct visualizations. Argdown is a much simpler syntax that is built upon Markdown, has open-source tools and extensions, and provides excellent visualization with minimal configuration. Its integration with Pandoc is particularly beneficial for integration with several document formats. While Graphviz has open-source tools, it uses a more generalized language that would require additional layers of abstraction for argument mapping. DebateGraph is a web application that implements a very rich library of visualizations that would be highly suitable for more advanced use cases as our work evolves further.

We developed a reduced grammatical specification API for Argdown in Python, which aids in abstracting the syntax and produces a simple but valid argument map for our use case. To produce the maps, we iterate over our list of topics, and select the top three arguments within that topic. Then we select the three most highly voted comments that support each argument. Each argument points to the topic to which it belongs, and each statement points to the argument that it supports. This creates a three-layer hierarchy, where the topics calculated using c-TF-IDF are shown in the topmost layer, LLM-generated arguments are placed in the middle layer, and human-generated supporting statements are at the bottom. The links in the map are strictly for SUPPORT relationships since we treat REFUTE relationships as noise pending further evaluation. In the next chapter, Figure 4.13 shows example argument maps produced using this technique.

We use visual aids to clearly distinguish human-related content from that generated by these algorithms. The LLM-generated topic labels have a blue outline and a robot emoticon and the constructed arguments or insights have a solid blue background with the same robot emoticon, while the human-generated statements have a green outline and feature a regular smiling emoticon that represents a human contributor. Based on trial and error, we determined that the visualiza-

tion provides a reasonable information density when displaying three arguments for each topic and three statements from each argument. Since a statement can support multiple arguments within that topic, we end up with fewer than nine statements per topic. Larger numbers tend to yield maps that are very difficult to follow. Arguments are selected based on *acceptance*, which is the ratio of the estimated number of voters to the total voting population considered. Additionally, we calculate and select the same number of arguments from each cluster based on their position within that topic. This may result in the selection of certain arguments with lower *acceptance* overall depending on the voting patterns within the particular containing topic but ensures equitable representation of arguments from each topic. We select the statements for each argument based on the highest raw number of agreeable votes, ensuring that we are weighing in favor of statements that generated the most raw support for the given argument. For example, a statement with 1000 agree votes and 100 disagree votes would be favored over a statement with 10 agree and 0 disagree votes. Using the proportional *acceptance* variable works well for arguments since they tend to have larger agree/disagree numbers, while certain statements could have near zero votes, increasing the risk of skewing the data.

In the following chapter, we discuss the results of our experiments and analyze the generated insights and argument maps.

## CHAPTER 4: RESULTS

We examine a multi-class classification system for comment moderation and the use of Chain-of-Thought reasoning to identify irrelevant comments. Our analysis, using the BERTopic framework for topic modeling, effectively identified hierarchical clusters in several datasets. For argument generation, we employed a multistep reasoning process with LLMs to create unique and compelling actionable insights. We then correlated these insights with the voting data to identify the most agreeable positions. The argument mapping technique visually represented the relationships among user-generated comments and LLM-generated insights across various topics, along with an estimate of public acceptance for each position. This showcases the capabilities of LLMs in processing information and significantly increasing the signal-to-noise ratio of the data.

### Comment Moderation as a Multi-class Classification Task

The comment moderation task is crucial for maintaining the quality and relevance of discussions in online deliberation platforms like Polis. This task involves reviewing each statement submitted by users and deciding whether they are appropriate for publication on the platform or need further review. Given the complexity of human language and the variety of contexts in which comments are made, this task presents significant challenges.

In the context of this experiment, comment moderation is treated as a multi-class classification task, where each statement is classified into one of three categories: ACCEPT, REJECT, or UNSURE. This approach allows for a nuanced handling of comments over binary classification and does not force the model to make a decision when the context is unclear. Multi-class classification is particularly important in comment moderation, as it accommodates the diverse nature of comments

and allows the model to more precisely follow the provided instructions.

The experiment was structured into nine experimental configurations, each building upon the insights gained from the previous configuration. These configurations were methodically designed to test various aspects of the classification task, including the impact of example-based learning, the effect of deconstructing statements before classification, and the use of the Chain-of-Thought (CoT) reasoning techniques to simulate a reasoning process before making a decision.

To evaluate these experiments, we selected two datasets, *american-assembly.bowling-green* and *scoop-hivemind.biodiversity*, due to their high-quality statements and alignment with the Polis moderation guidelines. No pre-processing was applied to the data, ensuring that the raw statements were evaluated by the model. This decision was made to closely mimic the real-world scenario where moderators often have to make quick decisions based on the guidelines immediately available to them.

### *Evaluation Metrics*

The effectiveness of these experiments is evaluated using several key metrics that offer a distinct perspective on the model’s performance, focusing on its accuracy, F1 score, and the balance between cautious indecision and incorrect rejection.

#### *Accuracy*

Accuracy is crucial in classification tasks and reflects the model’s ability to align its predictions with the actual labels. We calculated accuracy as the number of correct predictions by the number of all predictions excluding UNSURE, effectively quantifying the match between the model’s predictions and the ground truth. The exclusion of UNSURE predictions from calculations en-

sures that the metric focuses on the model's performance in making definitive decisions. A higher accuracy indicates better model performance.

### *F1 Score*

The F1 score is a balanced measure that considers both the precision and recall of the classification process, particularly useful when the costs of false positives and false negatives differ significantly. It is the harmonic mean of precision and recall, providing a single metric that balances both the concerns of recovering all relevant instances, recall, and the correctness of the predictions, precision. The F1 score is especially useful when dealing with imbalanced datasets where positive cases, which in our case are the comments to be rejected, are significantly less common than the negative ones. It is calculated using the formula

$$F_1 = 2 \cdot \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2 \times TP}{2 \times TP + FP + FN}$$

A higher F1 score indicates better model performance concerning both false positives and false negatives, making it particularly useful when the costs of false positives are high, such as in comment moderation where incorrect rejections could suppress valid discussions. This metric ensures that both aspects are taken into account, providing a more holistic evaluation of the model's performance.

### *Unsure Rate*

The unsure rate quantifies the frequency with which the model opts for the UNSURE classification instead of making a decisive judgment. This metric is pivotal for understanding the model's

tendency towards caution in unclear situations. Opting for UNSURE over a potentially incorrect rejection or acceptance can be seen as a conservative strategy, aimed at minimizing the risk of erroneously moderating a comment. This approach prioritizes accuracy, precision and recall over decisiveness, leaning towards a preference for human moderation in ambiguous cases.

### *False Positive Rate*

The false positive rate is an essential metric for evaluating the model's propensity to incorrectly reject legitimate comments. This rate is inversely related to the desired outcome in spam detection scenarios; a lower false positive rate is indicative of a model that carefully preserves legitimate content while filtering out spam. The emphasis on minimizing false positives stems from the greater harm associated with erroneously rejecting a valid comment compared to the lesser impact of failing to block a spam statement. The model's design is inclined towards erring on the side of caution, favoring false negatives over false positives in its moderation decisions.

### *Overall Results*

The comparative analysis of various metrics across experimental configurations, as depicted in Figure 4.1, provides insights into the effectiveness of different techniques. The baseline three-class classification task shows reasonable accuracy but struggles with a high false positive rate. The incorporation of various techniques tends to increase the model's unsure rate, a desirable outcome for cautious moderation, while occasionally increasing F1. The experiments reveal that while seven-class classification tasks generally show a dip in accuracy and F1 score compared to the baseline, they offer increased explainability. Notably, the comment deconstruction technique markedly improves the F1 score and reduces false positives. Tables 4.1, 4.2, 4.3, 4.4 and 4.5 show the detailed results for each dataset across all configurations and metrics.

Config	Classes	Technique	Accuracy	F1	FPR	Unsure
1	3	Baseline	0.6916	0.2459	0.0528	0.0011
2	3	Examples	0.7032	0.2090	0.0215	0.0033
3	3	Second-Thought	0.6989	0.2012	0.0304	0.0290
4	3	Examples + Second-Thought	0.7015	0.1988	0.0231	0.0056
5	7	Baseline	0.7038	0.2421	0.0283	0.0089
6	7	Thought	0.6937	0.2370	0.0462	0.0379
7	7	Thought + Deconstruction	0.6958	0.2090	0.0371	0.0279
8	7	Deconstruction	0.7028	0.1957	0.0216	0.0123
9	3	Thought + Deconstruction	0.7012	0.2013	0.0424	0.0513

Table 4.1: Comment moderation results for the *american-assembly.bowling-green* dataset.

Config	Classes	Technique	Accuracy	F1	FPR	Unsure
1	3	Baseline	0.5538	0.3254	0.0844	0.0000
2	3	Examples	0.5162	0.1486	0.0395	0.0253
3	3	Second-Thought	0.5498	0.3069	0.0789	0.0158
4	3	Examples + Second-Thought	0.5148	0.1294	0.0267	0.0348
5	7	Baseline	0.5531	0.3015	0.0658	0.0158
6	7	Thought	0.5319	0.2747	0.1197	0.1076
7	7	Thought + Deconstruction	0.5464	0.3053	0.1156	0.0791
8	7	Deconstruction	0.5329	0.2283	0.0662	0.0380
9	3	Thought + Deconstruction	0.5207	0.2567	0.1361	0.0823

Table 4.2: Comment moderation results for the *scoop-hivemind.biodiversity* dataset.

Config	Classes	Technique	Accuracy	F1	FPR	Unsure
1	3	Baseline	0.6375	0.2564	0.0980	0.0000
2	3	Examples	0.6053	0.1667	0.0851	0.0500
3	3	Second-Thought	0.6364	0.1765	0.0800	0.0375
4	3	Examples + Second-Thought	0.6053	0.1667	0.0851	0.0500
5	7	Baseline	0.6234	0.2162	0.1020	0.0375
6	7	Thought	0.5600	0.2326	0.2292	0.0625
7	7	Thought + Deconstruction	0.5949	0.3043	0.2157	0.0125
8	7	Deconstruction	0.5789	0.1111	0.1600	0.0500
9	3	Thought + Deconstruction	0.5132	0.1778	0.2708	0.0500

Table 4.3: Comment moderation results for the *scoop-hivemind.freshwater* dataset.



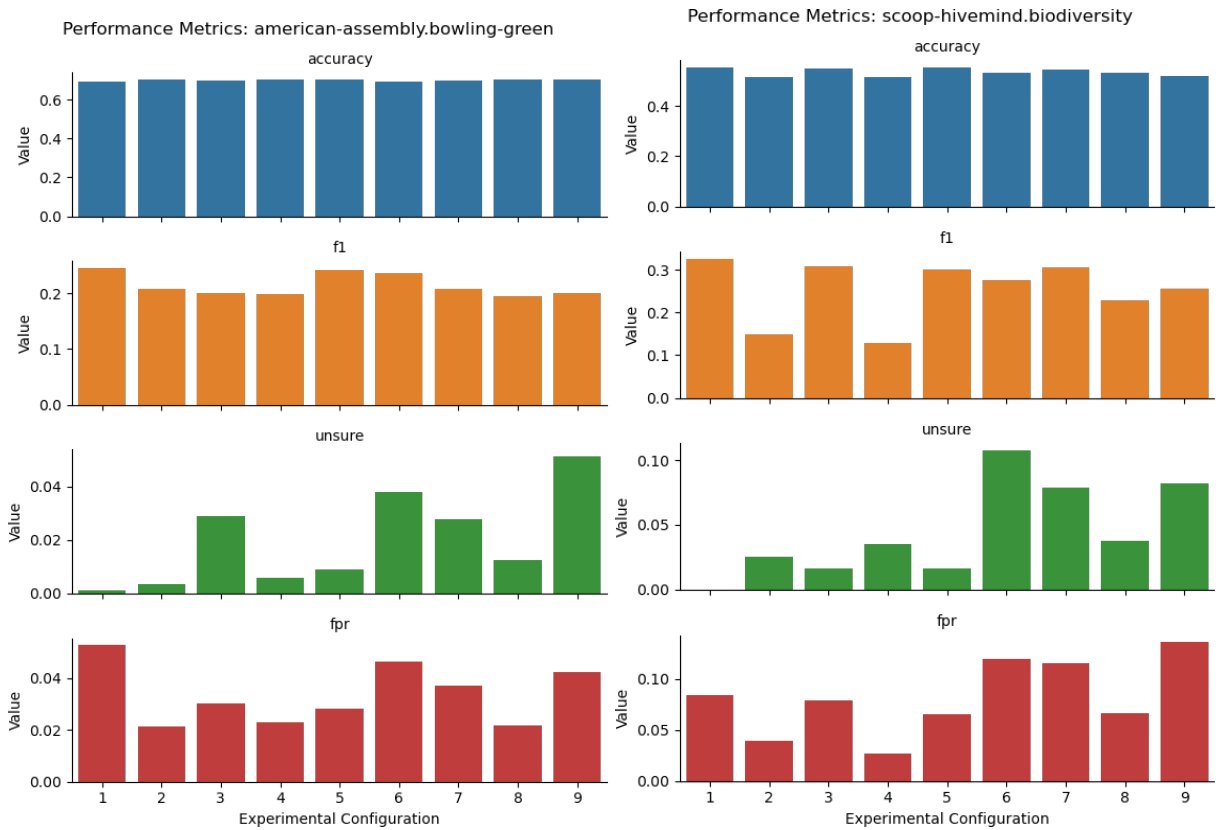


Figure 4.1: Performance of comment moderation techniques across different datasets. The left graph shows the results for the *american-assembly.bowling-green* dataset, while the right graph shows the results for the *scoop-hivemind.biodiversity* dataset. The x-axis represents the various experimental configurations, while the y-axis shows the accuracy, f1 score, unsure rate, and false positive rate.

### Baseline Three-class Classification

The baseline configuration shows a trade-off between accuracy and F1 score, with a notably low unsure rate but a substantial false positive rate. This serves as a crucial benchmark, indicating the model’s tendency to reject comments based on spam without adequately addressing the complexity of the comments.

Config	Classes	Technique	Accuracy	F1	FPR	Unsure
1	3	Baseline	0.6149	0.1231	0.0440	0.0000
2	3	Examples	0.5816	0.1449	0.0833	0.0473
3	3	Second-Thought	0.6014	0.1194	0.0659	0.0000
4	3	Examples + Second-Thought	0.5827	0.1212	0.0723	0.0608
5	7	Baseline	0.5850	0.1159	0.0889	0.0068
6	7	Thought	0.5423	0.2169	0.2273	0.0405
7	7	Thought + Deconstruction	0.5753	0.3111	0.2222	0.0135
8	7	Deconstruction	0.5405	0.1053	0.1648	0.0000
9	3	Thought + Deconstruction	0.5634	0.2439	0.1954	0.0405

Table 4.4: Comment moderation results for the *scoop-hivemind.taxes* dataset.

Config	Classes	Technique	Accuracy	F1	FPR	Unsure
1	3	Baseline	0.7030	0.1404	0.0588	0.0000
2	3	Examples	0.6933	0.1667	0.0769	0.0121
3	3	Second-Thought	0.7051	0.0000	0.0351	0.0545
4	3	Examples + Second-Thought	0.6987	0.0408	0.0526	0.0545
5	7	Baseline	0.7222	0.1509	0.0342	0.0182
6	7	Thought	0.6646	0.1587	0.1379	0.0424
7	7	Thought + Deconstruction	0.6519	0.1270	0.1316	0.0424
8	7	Deconstruction	0.6584	0.0678	0.1111	0.0242
9	3	Thought + Deconstruction	0.6522	0.1765	0.1538	0.0242

Table 4.5: Comment moderation results for the *scoop-hivemind.affordable-housing* dataset.

### *Three-class Classification with Examples*

Introducing examples to the classification task led to a decrease in overall accuracy and an increase in the unsure rate, suggesting that the model’s decision-making process might be unduly influenced by the content of the examples. This sensitivity highlights the challenge of using generic examples in training, which might not always align with the specific moderation needs of different datasets.

### *Three-class Classification with Second-Thought Reasoning*

Incorporating a “second-thought” reasoning process, where the model re-evaluates its initial reject decisions, shows promise in increasing the F1 score and reducing the false positive rate for certain datasets. This method appears to encourage a more cautious approach to moderation, favoring accuracy over hasty rejections.

### *Three-class Classification with Second-Thought Reasoning and Examples*

Combining examples with second-thought reasoning does not yield the anticipated improvements, indicating potential confusion induced by the examples. This combination results in lower accuracy and F1 score, underscoring the importance of dataset-specific example selection.

### *Seven-class Classification*

Expanding the classification task to encompass a broader range of labels yields results comparable to the baseline in terms of accuracy and F1 score. This approach enhances the model’s ability to discern among various reasons for rejection and highlights the potential of such a model assisting human moderators.

### *Seven-class Classification with Thought Statement*

Contrary to expectations, combining seven-class classification with explicit reasoning leads to an increase in UNSURE predictions and false positives with a marked decrease in F1 score. The accuracy somewhat increased due to the exclusion of UNSURE predictions from that calculation. This outcome suggests that the added complexity of articulating reasoning may detract from the

model's focus on the instructional prompt, impacting its moderation effectiveness.

#### *Seven-class Classification with Comment Deconstruction and Chain-of-Thought Reasoning*

Deconstructing comments to extract semantic content significantly improves the model's moderation capabilities, reducing false positives and enhancing the F1 score. This method underscores the potential of advanced semantic analysis in refining AI moderation strategies.

#### *Seven-class Classification with Comment Deconstruction*

Focusing solely on comment deconstruction without the added layer of chain-of-thought reasoning yields a balance between accuracy and a low false positive rate. This approach demonstrates the efficacy of semantic deconstruction in improving moderation outcomes. For the *american-assembly.bowling-green* dataset, this approach achieved the lowest FPR.

#### *Three-class Classification with Comment Deconstruction and Chain-of-Thought Reasoning*

The final experiment, integrating comment deconstruction with chain-of-thought reasoning within a three-class framework, strikes a balance between reducing false positives and maintaining an acceptable level of unsure decisions. While the accuracy may dip due to the increased unsure rate, this approach emphasizes the value of cautious and reasoned moderation decisions.

### *Comparative Analysis of Results*

The incorporation of techniques such as comment deconstruction and Chain-of-Thought (CoT) reasoning revealed notable differences in model performance across several metrics, including ac-

curacy, F1 score, unsure rate, and false positive rate. The results underscore a marked preference for the seven-class classification approach coupled with comment deconstruction. This configuration stands out for its high F1 score, low false positive rate, and enhanced reasoning capabilities, resulting in a more nuanced interpretation and processing of user-generated content. Despite the broad array of parameters tested, we were limited in exploring every potential combination due to the inherent complexity of configurations. Notably, the introduction of examples resulted in much lower consistency in results and adversely affected both accuracy and F1 score. Conversely, employing a “second-thought” reasoning approach generally increased the UNSURE rate while concurrently reducing the false positive rate and increasing the F1 score.

A pivotal insight from the analysis is the trade-off between the UNSURE rate and the false positive rate. Strategies that increased the model’s tendency to classify comments as UNSURE typically reduced false positives, aligning with a more cautious moderation stance. This approach favored erring on the side of caution, favoring further human review over the risk of incorrect automatic moderation decisions. In scenarios involving seven-class classification without explicit reasoning, we observed more certainty but a slightly increased false positive rate relative to the baseline. Incorporating simple thought-statement reasoning into the seven-class classification framework ironically resulted in the least favorable outcomes with the lowest accuracy and highest false positives. However, integrating comment deconstruction within this framework significantly reduced false positive rates, highlighting the usefulness of this technique in refining the moderation process.

The use of comment deconstruction with a simplified three-class classification strategy yielded the most favorable outcomes in terms of minimizing false positives and maximizing the F1 score. Throughout the experimental process, the model’s explanations for comment rejection appeared coherent, providing us with a deeper understanding of its decision-making process. These explanations were instrumental in refining the prompts used in the experiments, ensuring their generalizability across various datasets and language models. These experiments demonstrate the intricate

balance required to optimize accuracy, F1 score, and the unsure rate for AI-driven comment moderation. We propose the use of comment deconstruction with carefully crafted chain-of-thought reasoning mechanisms for use in future works, offering a reasonable combination of decision-making and explainability toward more transparent moderation methodologies.

### *Challenges in Comment Moderation*

The primary limitation was the apparent unsuitability of general-purpose language models for strict classification tasks, with their performance seemingly more adept in areas involving semantic extraction or summarization. Models that are specifically tuned for these classification tasks on appropriate datasets would yield better performance for the moderation of user-generated content.

The experimental setup revealed a significant sensitivity of models to the chosen prompts. The effectiveness of prompts varied across different tasks and models, indicating that a prompt optimized for one model might not perform well with another. This variability poses a challenge in experimental design, requiring careful consideration in prompt selection and optimization. Additionally, assessing the influence of specific prompts or examples on model performance emerged as a massive challenge due to the possible number of combinations, further complicating the experimental design process.

Due to constraints on time and resources, it was impractical to explore every conceivable combination of prompts, models, and other experimental variables. This limitation meant that potentially useful configurations and techniques remain unexplored and present an opportunity for additional research.

Another encountered challenge was the language models' inability to identify and exclude duplicate comments efficiently. A language model would need access to all previous comments, which

is typically done by placing all comments in the context window for evaluation. This requires a large context window and increases the memory footprint linearly with the number of comments, often causing the model to run out of GPU memory or overflow past its maximum context window length. Additionally, while operating with several hundred comments in the context window, the model struggled to follow instructions and showed significant degradation in performance. We mitigated this by placing one comment at a time in the context window, which worked very well for our experiments but made it impossible to detect duplicate comments. A more effective approach would involve clustering comments using their text embeddings and detecting semantically identical statements, which is a promising technique for implementing a moderation system.

Ultimately, automating comment moderation using language models involves a complex set of challenges including model selection, prompts design, crafting specific examples, limitations in experimental configurations, difficulties in balancing moderation metrics, and technical hurdles in managing large datasets and duplicate statements. Addressing these challenges will require targeted research, innovative approaches, and likely a reevaluation of the tools and techniques currently available in this field.

## Topic Modeling

To uncover the latent structures within the debate statements, our analysis focused on the use of the BERTopic framework to develop a pipeline leveraging transformer embeddings combined with class-based Term Frequency-Inverse Document Frequency (TF-IDF) statistics to cluster and interpret the statements effectively. In the following subsections, we discuss our findings and choices in detail.

### *Embedding and Preprocessing*

We selected the *WhereIsAI/UAE-Large-V1* [16] embedding model based on its robust performance on the HuggingFace Massive Text Embedding Benchmark particularly for clustering tasks while maintaining an extremely low memory footprint. The choice to calculate embeddings at the statement level, rather than the sentence level, is aimed at capturing the comprehensive semantics including potentially multiple sentences within a single statement. We used mean pooling in the transformer model, in contrast to max pooling, to ensure that the resultant embeddings accurately reflected the average semantic content across the entire statement, thereby ensuring the reliability of clustering in the subsequent steps.

### *Dimensionality Reduction*

We used the Uniform Manifold Approximation and Projection (UMAP) [18] algorithm to manage the complexity of the high-dimensional embedding vectors and facilitate the clustering of the statements. UMAP was chosen over Principal Component Analysis (PCA) and other dimensionality reduction techniques due to its superior capability in preserving both the local and global structure of high-dimensional data. Unlike PCA, which linearly transforms the data focusing primarily on preserving global structure and maximizing variance, UMAP is capable of capturing the nonlinear relationships inherent in the data, making it particularly effective for text embeddings derived from transformer models. By translating these embeddings into a more manageable sub-100 dimensional vector space, we were able to effectively identify high-density clusters. We determined optimal UMAP hyperparameters were determined through extensive experimentation, focusing on four parameters shown in Table 4.6. The objective was to strike a delicate balance between preserving local detail and the overarching global data structure, ultimately aiming to uncover the hidden relationships between similar statements.



Parameter	Considered Values	Selected Value
n_neighbors	2, 4, 8, 16, 32, 64	8
min_dist	0.0, 0.1, 0.25, 0.5, 0.75, 1.0	0.0
n_components	32, 64, 128, 256, 512, 1024	32
metric	euclidean, manhattan, cosine	euclidean

Table 4.6: UMAP parameters tested along with their considered and optimal values.

**n\_neighbors** is arguably the most important parameter that determines the size of the local neighborhood around each point, dictating the balance between the emphasis on local versus global structures within the data. A smaller *n\_neighbors* value encourages the algorithm to prioritize local data structure, fostering the formation of tighter, more defined clusters based on the immediate neighborhood of points. Conversely, a larger value places greater weight on the global data structure, potentially blending local clusters into broader groupings and emphasizing the overarching layout of the data points over individual cluster distinctions.

Figure 4.2 shows the effect of various values on UMAP’s output. At *n\_neighbors* = 2 and *n\_neighbors* = 4, the graphs fail to reveal any meaningful clusters. We select *n\_neighbors* = 8 as the optimal value when the graph begins to take a more meaningful shape. This value strikes a balance, ensuring that neither excessively granular nor overly broad clusters dominate the analysis.

**min\_dist** controls the minimum distance between points in the low-dimensional embedding space, regulating the tightness of clustering. Figure 4.3 shows the effect of various values on UMAP’s output. A lower *min\_dist* value permits closer clustering, capturing the intricate local structure of the data, which is particularly beneficial for detailed clustering tasks. In contrast, higher values enforce a more dispersed distribution of points, generating more visually appealing but often losing essential information in the data. We selected *min\_dist* = 0 to allow for the best possible clustering. This setting enables UMAP to accurately represent the close relationships between semantically similar statements without undue dispersion.

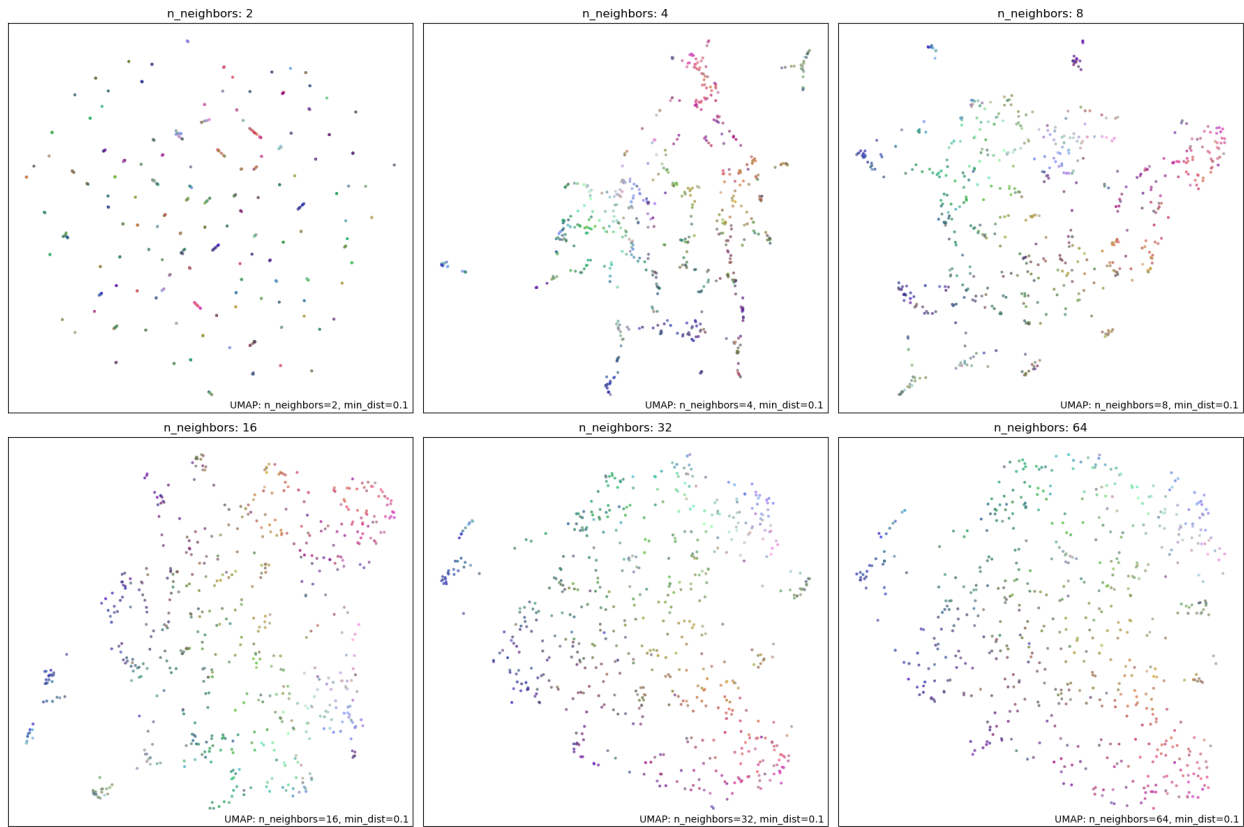


Figure 4.2: UMAP  $n\_neighbors$  parameter. Smaller neighborhoods capture finer details while sacrificing global structure. Large values overly emphasize global structure but lose the finer details. Optimum values of 4, 8, and 16 show a reasonably balanced spread of data points.

**n\_components:** Reducing the dimensionality of the dataset is a critical step to manage complexity, and  $n\_components$  parameter directly controls that. A lower value may oversimplify the data structure, while an excessively high value can retain unnecessary complexity, hindering the identification of meaningful patterns. UMAP documentation recommends a value between 10 and 100 to be determined through experimentation. In contrast, the BERTopic library defaults to a value of 5 if not explicitly set.

In our analysis, the effect of  $n\_components$  on the resulting reduced vector is difficult to visualize,

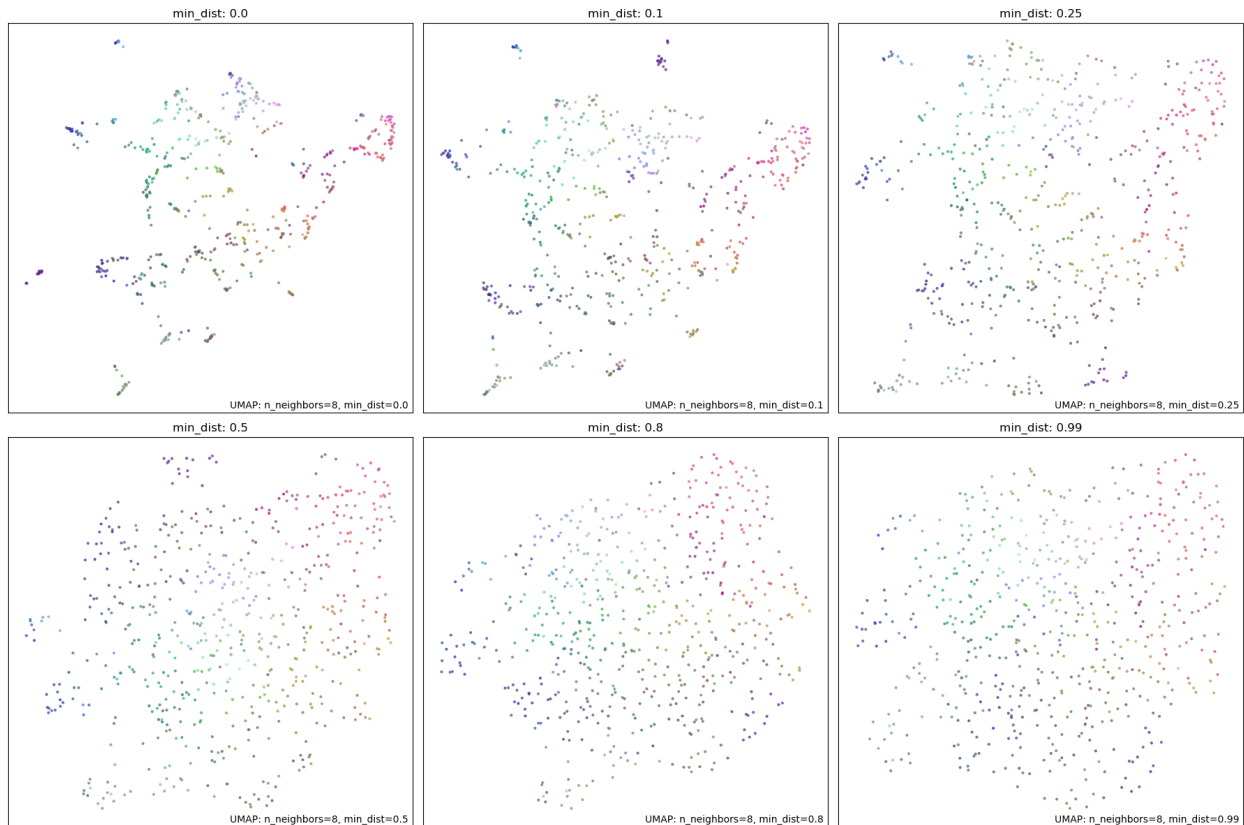


Figure 4.3: UMAP *min\_dist* parameter. These figures show the impact of varying minimum distance between each projected data point. Higher values result in visually appealing charts but destroy the structural information that enables clustering. Lower values are better in the context of clustering data points based on similarity with 0.0 yielding the best results.

since all of our visualizations are in 2D. To reliably inspect the effect of *n\_components*, we first reduced the vector to the selected value of *n\_components* and then down to two dimensions for visualization. Figure 4.4 shows the result of various values. We ultimately selected *n\_components* = 32 as the lowest value that still appeared to retain the semantic structure of the data.

**metric** influences how the algorithm interprets the distances between points, which in turn affects the shape and structure of the resulting clusters. Most common metrics include *euclidean*, *cosine*, and *manhattan* distances, each with its strengths depending on the nature of the dataset. The Eu-

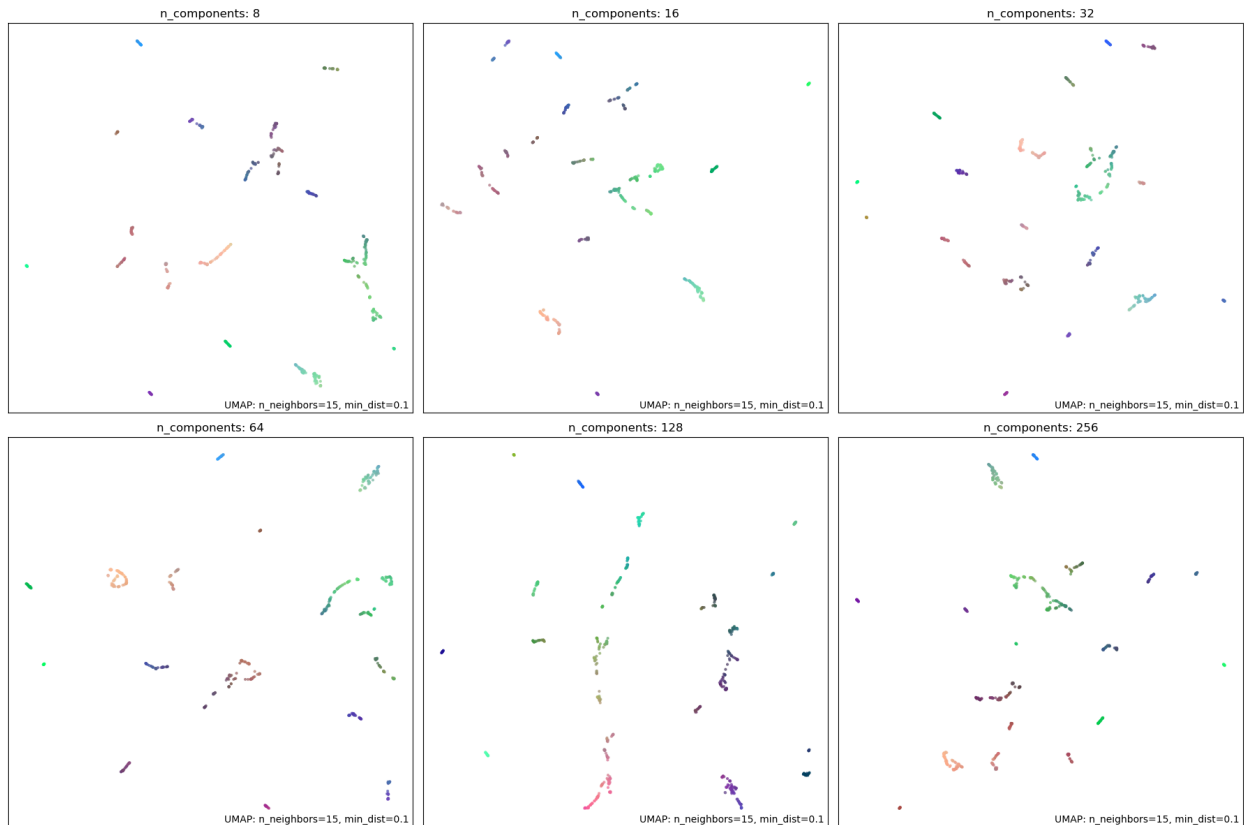


Figure 4.4: UMAP  $n\_components$  parameter. These graphs show a 2D projection of embedding vectors that had already been reduced to  $n\_components$  dimensions. While we run these vectors through UMAP twice, we expect that the changes made to the global structure of these embeddings from the first run will be appropriately captured in the final 2D projection. The colors of individual data points are still computed using PCA on the original embeddings, maintaining relevance to the global structure of the data.

clidean metric measures the straight-line distance between points and is well-suited to preserving the geometric structure of the data. In contrast, the *cosine* metric assesses the angle between vectors, emphasizing directional similarity over magnitude, which can be particularly useful where orientation in the vector space signifies semantic similarity. Figure 4.5 shows the effect of various metrics on UMAP's output. All three metrics are valid for our data and yield similar results. However, the Euclidean metric takes the magnitude of the vectors into account, which is not desirable

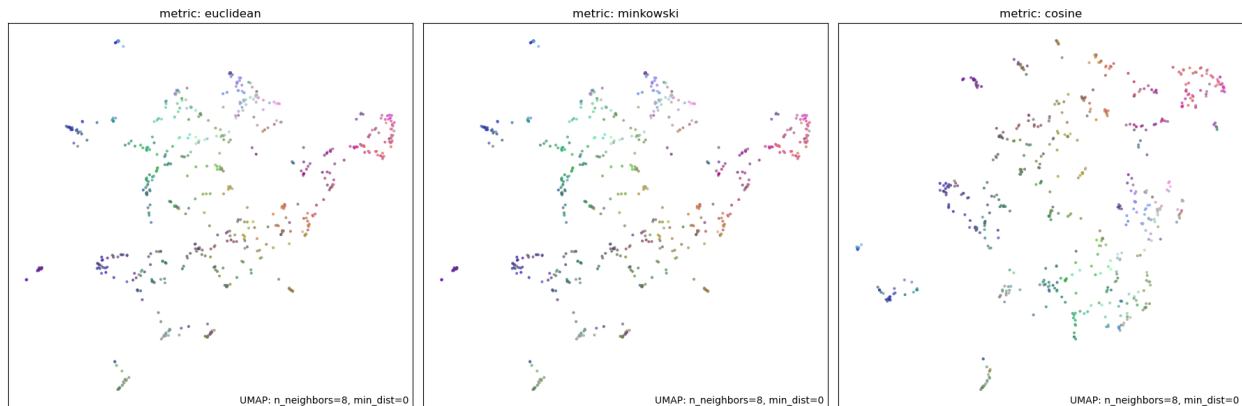


Figure 4.5: UMAP *metrics* parameter. All metrics appear to yield similar results. We select *cosine* to ensure that vectors are compared based on their semantic similarity but not magnitude.

for calculating semantic similarity. We select the *cosine* metric as the optimal choice for our data.

Additionally, enabling *densmap* significantly enhanced the clustering results by incorporating density estimates as a regularizer in the optimization of the low-dimensional representation, aiding HDBSCAN in more robust cluster identification. This was determined from the UMAP documentation.

With these parameters, we generate a final projection of our dataset. In Figure 4.6, the first image shows the 2D projection of the embeddings resulting from the UMAP model. The second figure shows a network connectivity graph that highlights the distance between various points in a 2D space. The edge thickness represents the distance between the points. This graph shows a high-density core and several outliers, which is consistent with a scale-free network and is ideal for clustering.

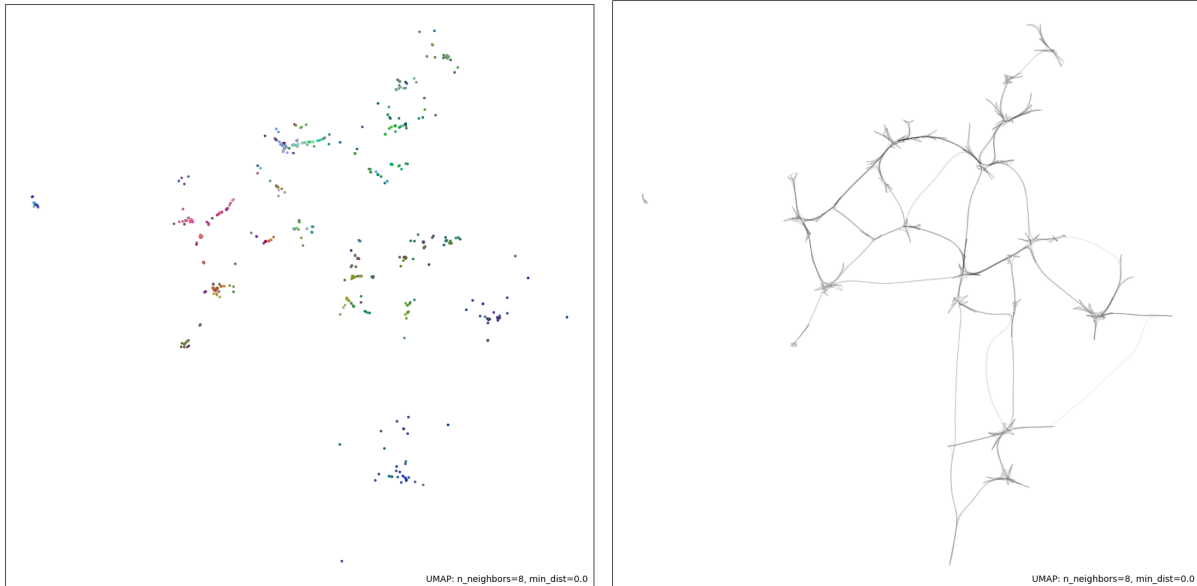


Figure 4.6: Final UMAP Projections of Statement Embeddings for the *american-assembly* dataset. The first one shows the data points in groups somewhat consistent with their colors, which shows a balance between global structure and local relationships. The second image is a network connectivity graph that shows the distance between various clusters. A high-density core with several offshoots is consistent with a scale-free network and is considered a strong candidate for clustering. A hairball-type projection would have implied a poorly projected set of vectors.

### *Clustering*

The HDBSCAN algorithm was instrumental in identifying high-density regions in our dataset that corresponded to topics and areas of interest. However, some data points in the resulting topic assignment, as illustrated in Figure 4.7, did not belong to any cluster and were labeled as outliers. We refrained from dismissing these outliers as noise since our statements had been moderated and screened for spam or irrelevant content in previous steps. To ensure that all statements were fully represented in our subsequent analyses, we assigned these data points to the nearest topic based on the probability distribution calculated by BERTopic. If any outliers still remain, we assign them based on embedding similarity to the topic centroids. Interestingly, this outlier assignment changed

each topic's representation and shifted the centroids due to the addition of new data points. This iterative process ensured that no data was overlooked and that all statements were considered in our further analyses.

Visual representations of the data clustering process are illustrated in Figure 4.7, depicting the embeddings in a two-dimensional space. Each topic is represented by a different color while the outliers are initially represented in gray. The second graph shows the reassignment of these outliers to the nearest topics.

Figure 4.8 further shows the distribution of statements across identified topics, both before and after the outlier assignment process. Initially, a significant number of statements are assigned to  $topic = -1$  representing the outliers. Upon redistribution, each topic's statement count increases somewhat evenly as expected. The distribution of statements across clusters is equitable and well-balanced.

Using BERTopic, we also structured the topics into a coherent hierarchy based on similarities illustrated in Figure 4.9. This hierarchical arrangement serves as a precursor to the development of an argument map, facilitating a logically connected representation of the topics.

In this set of experiments, UMAP and HDBSCAN algorithms effectively identified dense clusters in the data, which were then labeled and represented by a set of keywords through our comprehensive topic modeling pipeline. The outlier analysis revealed several statements unassigned to any topic, which were subsequently allocated to the most similar topics based on their embeddings, ensuring no data was overlooked in our analysis. This systematic approach underscores the interconnectivity between distinct yet related areas of discussion and enables us to derive actionable insights with a comprehensive understanding of the underlying thematic structures.



Figure 4.7: Distribution of statements colored by topic projected in a 2D plane using PCA. The upper image shows the statement distribution before the outlier assignment while the lower one is updated after the assignment process. While the gray data points disappear in the process, each topic is inherently changed and represented with a different set of labels that include the newly assigned statements.



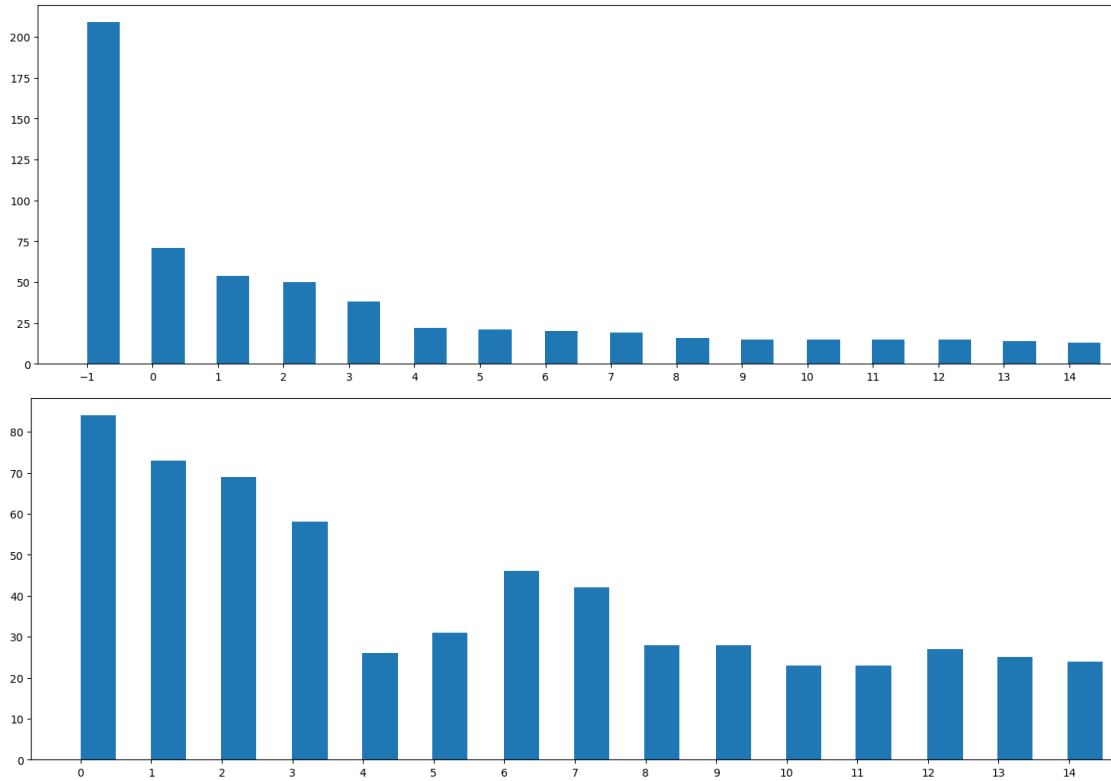


Figure 4.8: The top graph shows the distribution of statements across various topics immediately upon clustering. The first bar represents a large proportion of statements that do not belong to a topic and are considered outliers. The bottom graph shows a relatively even topic distribution after assigning each outlier to a topic based on its embedding proximity to the topic centroids.

### Argument Generation

Number of arguments in each topic is based on the number of statements it contains. To ensure a good balance of comments, we set the number of arguments,  $n_{args} = \lceil 2 * \ln n_{comments} \rceil$ . This formula sufficiently keeps the number of generated arguments low and makes it less sensitive to the number of comments. Most of our topics had 4 - 70 statements, which resulted in 3 - 9 arguments. A balanced number of arguments allows us to limit the amount of output for each topic while ensuring equitable representation.

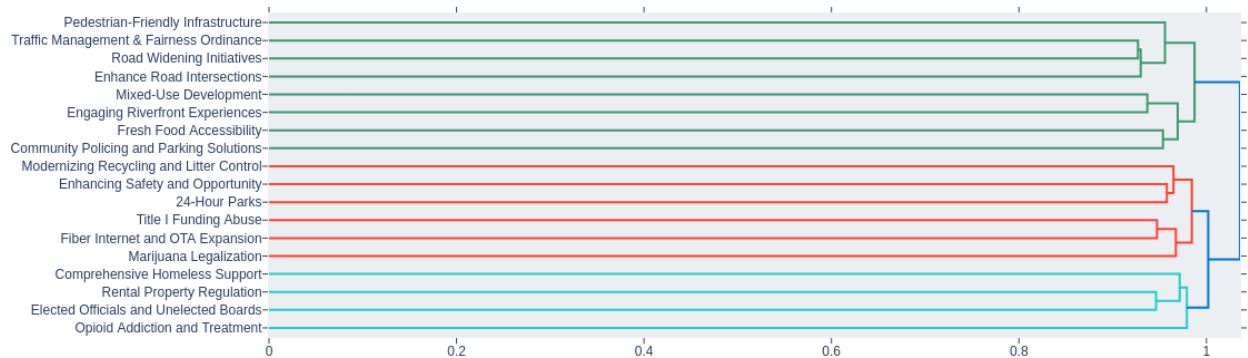


Figure 4.9: Topic Hierarchy dendrogram depicting inter-topic distance among centroids of statement clusters. Topics in proximity have common themes and can be safely merged into a larger topic. This hierarchical map can be used to organize the argument map in more nuanced layers and aid in further analysis.

Since argument generation involves semantic interpretation and extraction, we expected better results with chain-of-thought techniques and tried several variations. We placed the task and statements in the LLM’s context and asked it for one argument at a time, which caused the model to repeat itself often. To minimize duplicate arguments, we asked the model for a list of  $n_{args}$  most important areas of improvement in the first message. This enabled the model to respond at once and increased the uniqueness of each identified area. Then we prompted the model with one area of improvement at a time and asked it to identify related problems, propose actionable solutions from the statements, and then make a compelling argument in one sentence urging the need for action. This technique yielded a significantly better result. Since good prompts are extremely difficult to quantify, we resort to a qualitative assessment of the results here. Figure 4.10 shows our prompts and the generated output.

user	List the 8 most important areas of improvements from these statements, each on a new line.
assistant	<ul style="list-style-type: none"> <li>- Cooperation between WKU and Bowling Green for community involvement and internships</li> <li>- Accessible mentorship programs for young adults and teenagers</li> <li>- Community enrichment classes and activities, including woodworking, welding, gardening, and home beautification</li> <li>- More after-school activity centers for youth with extended hours of operation</li> <li>- Increased number of concerts held in Bowling Green</li> <li>- Better incorporation of university resources in community, including Kentucky Museum</li> <li>- Public adult swim facilities for year-round physical use</li> <li>- More activities and places to go for teenagers</li> </ul>
user	AREA OF IMPROVEMENT: Cooperation between WKU and Bowling Green for community involvement and internships
assistant	<p>PROBLEMS IDENTIFIED: Lack of cooperation and partnership between WKU and Bowling Green, limited opportunities for community involvement and internships for students</p> <p>ACTIONABLE SOLUTIONS: Strengthening mutually beneficial partnerships between WKU and Bowling Green, increasing cooperation for community involvement and internships</p> <p>ARGUMENT: Strengthening WKU-Bowling Green partnerships will create more community involvement and internship opportunities, benefiting both students and the local community.</p> <p>ARGUMENT LABEL: Boosting student opportunities and community involvement</p>
user	AREA OF IMPROVEMENT: Community enrichment classes and activities, including woodworking, welding, gardening, and home beautification
assistant	<p>PROBLEMS IDENTIFIED: Limited community enrichment classes and activities in Bowling Green, particularly in woodworking, welding, gardening, and home beautification</p> <p>ACTIONABLE SOLUTIONS: Establishing and promoting community enrichment classes and activities that cater to various interests and hobbies</p> <p>ARGUMENT: Introducing diverse community enrichment classes and activities will cater to various interests, enhancing the overall quality of life in Bowling Green.</p> <p>ARGUMENT LABEL: Diversifying community enrichment</p>
user	AREA OF IMPROVEMENT: More after-school activity centers for youth with extended hours of operation
assistant	<p>PROBLEMS IDENTIFIED: Insufficient after-school activity centers for youth with limited hours of operation in Bowling Green</p> <p>ACTIONABLE SOLUTIONS: Developing and expanding after-school activity centers with extended hours of operation</p> <p>ARGUMENT: Providing more after-school activity centers with extended hours will offer youth safe, engaging, and educational spaces beyond traditional school hours.</p> <p>ARGUMENT LABEL: Expanding youth opportunities</p>

Figure 4.10: The set of instructions and prompts used to generate arguments from comments using the LLM. Text highlighted in green is produced by the language model, while everything else is inserted into the LLM context by us programmatically.

## Argument-Statement Association

Associating arguments with individual statements is a crucial part of our pipeline since it enables us to measure the support behind each argument and prioritize the ones with the highest levels of consensus. We experimented with three techniques to deduce this association.

1. Present all arguments from a topic to the LLM at once and ask it to classify each statement to the most semantically related argument.
2. For each argument, use text embeddings to find the most closely aligned statements.
3. Frame the problem as a three-class classification task (SUPPORTS, REFUTES, UNRE-

LATED) and present each possible argument-statement pair within a topic to the LLM one at a time.

The first experiment was simple but produced very unreliable results. We asked the LLM to output the index of the most closely related argument, and the preliminary results almost appeared random. We also asked the LLM whether the statement would SUPPORT or REFUTE the argument, or whether it was UNRELATED. This paradigm assumes that each statement is represented by exactly one of the generated arguments and ignores the possibility of it supporting multiple arguments. The results were very unreliable and the chain-of-thought reasoning did not appear to help. We believe it was caused by too much-unrelated text in the model’s context window that was severely impacting its reasoning abilities.

The second technique reliably produced a list of semantically related argument-statement pairs, but it did not indicate whether the two had a SUPPORT or REFUTE relationship, or even if one logically followed the other.

The third technique, shown in Figure 4.11, yielded the best results. We specified in the prompt that the model should only indicate SUPPORT if a person who agrees with the statement will definitely support the argument. We also had the model produce a thought statement before making a decision, allowing it to deliberate on how such a person would respond. Applying this process to every possible argument-statement combination is very computationally expensive, and while we minimized the number of combinations by limiting those combinations to within each topic, it was still the most significant bottleneck in our pipeline. However, with an Nvidia H100 GPU, we were able to process all Polis datasets through our proposed pipeline in under 24 hours, so the computational cost did not become a bottleneck. We recommend a combination of second and third techniques to further improve the results.

You will be presented a statement and an argument. Statement is a user-generated comment from a discussion. Argument is an actionable solution.

TASK: Determine whether the statement supports, refutes, or is unrelated to the argument.  
 SUPPORT: The argument is consistent with the statement. A person who agrees with the statement will definitely support the argument.  
 REFUTE: The argument goes against the statement. A person who agrees with the statement will definitely with the argument.  
 UNRELATED: The statement and argument are not directly related. Implementing the argument will not directly address the underlying issue.

instruction

---

OUTPUT FORMAT  
 THOUGHT: Deliberate on how strongly a person who agrees with the statement will support the argument.  
 RELATIONSHIP: One of the following: SUPPORT, REFUTE, UNRELATED  
 REASON: Provide a reason for your choice.

---

user ARGUMENT 6: Active Community  
 Enhancing indoor sports facilities would encourage physical activity, promote teamwork, and cater to diverse athletic interests.

---

user STATEMENT 891: After school activity centers for youth with later hours of operation.

---

assistant THOUGHT: A person who agrees with the statement would likely support the argument, as later hours of operation for activity centers would provide more opportunities for youth to engage in physical activity and potentially participate in indoor sports.  
 RELATIONSHIP: SUPPORT  
 REASON: Later hours of operation for activity centers align with the argument's goal of enhancing indoor sports facilities and promoting physical activity.

---

user STATEMENT 857: BG's Kummer Little gym is open limited hours for indoor track. On other hand, County has new gyms that are open lots of hrs. Inequitable.

---

assistant THOUGHT: A person who agrees with the statement may refute the argument, as they might perceive the unequal access to indoor sports facilities as an issue that needs to be addressed before enhancing existing facilities.  
 RELATIONSHIP: REFUTE  
 REASON: The argument focuses on improving existing facilities, while the statement highlights the inequity between different facilities, which might be seen as a more pressing concern.

---

user STATEMENT 693: Bowling Green needs more accessible programs for mentorships of young adults and teenagers.

---

assistant THOUGHT: A person who agrees with the statement might find the argument unrelated, as the enhancement of indoor sports facilities does not directly address the need for mentorship programs for young adults and teenagers.  
 RELATIONSHIP: UNRELATED

Figure 4.11: The set of instructions and prompts used to characterize the relationship between arguments and comments using the LLM. Text highlighted in green is produced by the language model, while everything else is inserted into the LLM context by us programmatically.

Dataset	TOTAL	SUPPORT	REFUTE
american-assembly.bowling-green	4361	1625	17
scoop-hivemind.affordable-housing	503	327	9
scoop-hivemind.biodiversity	657	539	3
scoop-hivemind.freshwater	188	145	0
scoop-hivemind.taxes	401	201	7
scoop-hivemind.ubi	278	199	6

Table 4.7: The number of argument-statement pairs listed by each dataset, and the number of pairs that are labeled as SUPPORT or REFUTE. The model tends to classify very few links as REFUTE.

Figure 4.12 shows the distribution of argument-comment mappings with a SUPPORT relationship. The distribution of generated argument-statement mappings across topics was generally proportional to the number of statements in each topic but skewed in favor of topics with more agreeable votes. This is because the less agreeable topics, such as the discussion around legalizing mar-

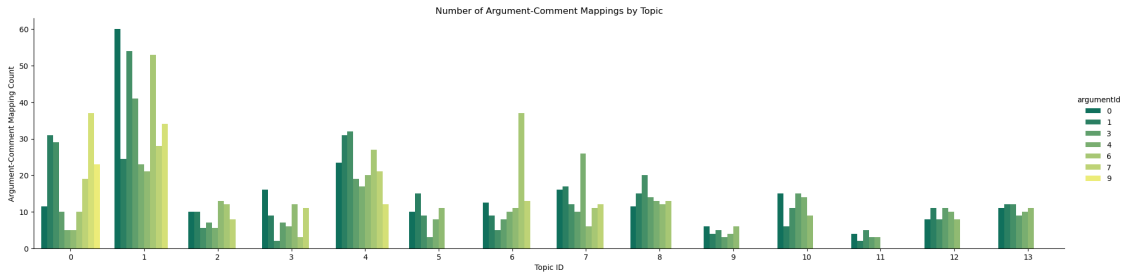


Figure 4.12: Distribution of argument-comment mappings with a SUPPORT relationship. The x-axis shows individual arguments clustered by topic while the y-axis indicates the number of SUPPORT links between that argument and corresponding statements. The tallest bars represent arguments with the most supporting statements but not necessarily the most agreed-upon positions. The initial topics have more statements and thus a larger number of arguments. This distribution skews towards the initial and less controversial topics that are more heavily represented through statements and votes.

ijuana in Bowling Green, tend to have a lesser proportion of agreeable votes, representing less popular opinions. This appears to work towards the goal of surfacing the arguments with the highest consensus without losing representation of minority statements, so we consider this a success. In measuring *acceptability* for individual arguments, we focus on metrics that account for this distribution, and thus we use a ratio instead of a raw count of votes.

Table 4.7 shows the argument-statement link distribution. For larger datasets, the majority of the pairs are classified as UNRELATED, which is favorable since we expect the model to be highly selective. Smaller datasets tend to have fewer overlapping topics with an overwhelming number of SUPPORT relationships. A very small fraction of statements appears to refute any of the arguments. A closer investigation of those specific links labeled as REFUTE shows that often the argument and comment are tangentially related but do not advocate for or against the same thing, and the model should have labeled these links as UNRELATED but appears confused due to the assertions made in either the statement or the argument. This classification step can be refined with additional reasoning prompts but is beyond the scope of this study. Instead, we treat

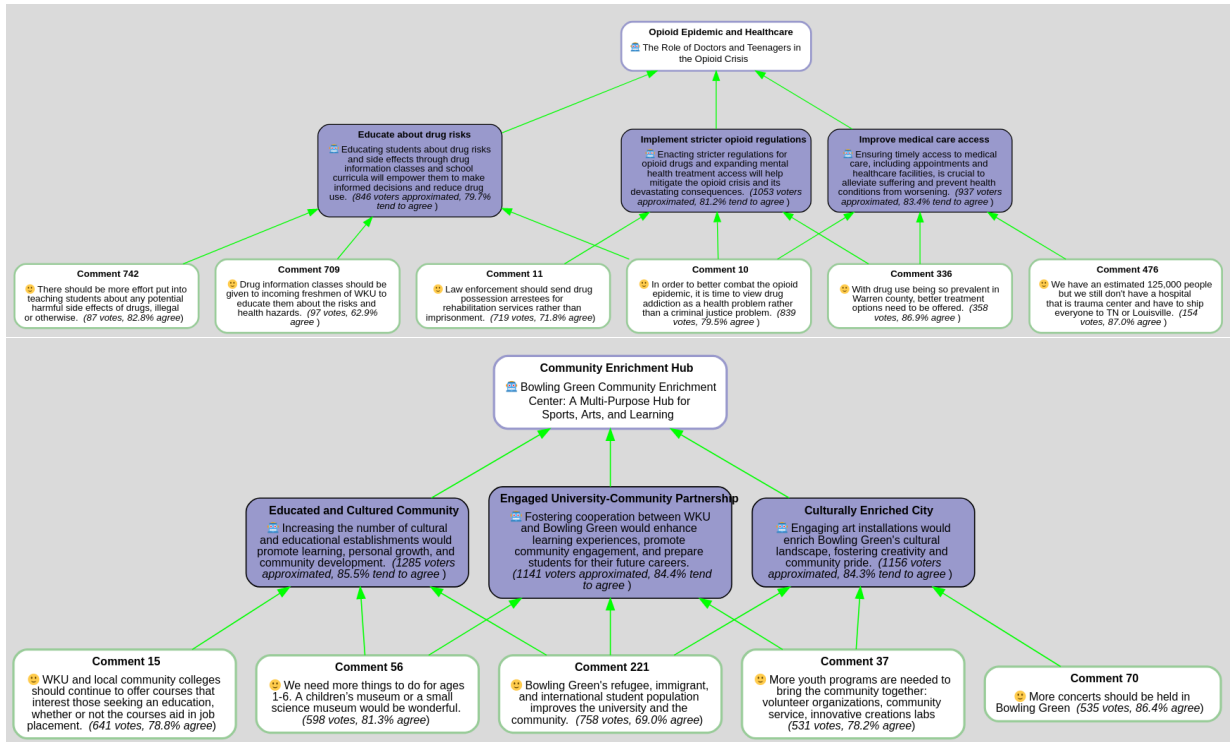


Figure 4.13: Argument maps generated using *american-assembly.bowling-green* dataset, covering two different topics. The upper map addresses healthcare and the opioid epidemic, while the lower focuses on community enrichment programs. The middle row in each map contains LLM-generated insights, while the lowest row has human-authored statements that support one or more of the arguments. We use Polis voting data to estimate the support for each generated argument.

the REFUTE relationships as noise in our data, ignore them, and instead focus on the SUPPORT relationships.

### Argument Mapping

Figure 4.13 shows argument maps generated using the *american-assembly.bowling-green* dataset. The first one addresses healthcare and the opioid crisis in the city of Bowling Green, while the second focuses on community enrichment programs. The top row shows the topics calculated

earlier using the c-TF-IDF algorithm, while the LLM-generated arguments are placed in the middle row along with a robot icon to clearly indicate the origin. Human-generated support statements are placed in the bottom row along with a smiley that denotes human authorship. The links are green indicating SUPPORT relationships, and REFUTE would have been annotated with the red color.



## CHAPTER 5: CONCLUSION

### Summary of Findings

The core problem underpinning this thesis is the challenge of harnessing LLMs to process and distill vast troves of unstructured information, incorporating diverse perspectives, and identifying actionable insights that represent public opinion and can drive policy decisions in their jurisdiction. The first challenge is in moderating the user-generated content to be displayed to other participants in real-time during a live event. As identified by Iandoli et al. [9], this daunting task requires a significant commitment from the hosting organization and takes up precious human hours that can instead be focused elsewhere. The Second challenge arises from the unstructured nature of these user-generated statements, making it very difficult to comprehend the direction of the debate and overall public opinion. The third issue is in parsing and distilling this information into meaningful recommendations that can be acted upon. Using the proposed pipeline, we demonstrate a set of simple steps to perform semantic extraction and reasoning on the posted comments. While the error rate of the classification task remained high, deconstructing the comments provided us with context around them and provided a way to measure semantic value added to the discussion, which can be used to automatically approve or flag statements for further human review. This technique presents an opportunity to develop a comprehensive solution to assist human moderators.

While the LLMs provide us with reasoning capabilities, they are often overpowered tools with significant memory and computation requirements. However, text embeddings derived from these transformer models are numerical vectors that capture the semantics of each word or sentence in a high-dimensional space. These embeddings are inexpensive to calculate and can be easily stored for later analysis. We incorporated embeddings into several steps of our pipeline; topic modeling was performed on these pre-computed vectors to determine the semantic similarity of various

statements. Using comment clustering, we establish a hierarchy of semantically linked topics, facilitating an understanding of thematic structures and the generation of actionable insights. The process of identifying the link between arguments and statements can be significantly optimized by the use of embeddings, and more importantly, this would enable us to make connections across topics.

In the process of designing and executing this pipeline, we identified dangers and pitfalls of using LLMs without rigorous analysis. Our algorithms could easily silence minority viewpoints and favor opinions entered early into the discussion. To avoid these dangers, we carefully consider how each of the decisions would impact the outcome of each step. Second, we often ask an LLM to explain its thought process and then store it for later analysis. This ensures that we adjust for any possible biases introduced by the algorithm. The generated argument maps visually represent the relationships between topics and insights, and highlight popular opinions.

### Theoretical Implications

This thesis addresses the challenge of moderating and structuring online debates, which are susceptible to low signal-to-noise ratios, unsystematic coverage, and polarization among participants, as outlined by Klein [14]. By leveraging LLMs for content moderation and argument structure extraction, the thesis lays the foundation for more coherent and productive online deliberation spaces. This is crucial for tackling wicked problems, which, according to [31], require a collective optimization approach to be reliably addressed. The capabilities of LLMs facilitate automation and reduce human cognitive load in performing relatively simple tasks. This enables us to design and architect platforms that augment human capabilities, improve the quality of human interactions, and ensure that the deliberation remains focused, inclusive, and representative of diverse viewpoints.

## Practical Implications

We demonstrate our novel approaches and highlight several strengths of LLMs through various sub-tasks. In content moderation, we identify an opportunity for human-AI teaming by allowing an LLM-based pipeline to review and automatically approve most of the content, while significantly reducing the number of statements that require human intervention, improving the interactivity and responsiveness of the platform. The incorporation of transformer embeddings into the pipeline speeds up the processes that do not require LLM inference. Semantic extraction further enables us to assess the relevance of user-generated content to the discussion at hand. Through the systematic processing and subsequent organization of user-generated content into coherent structures, LLMs are shown to be effective in distilling vast amounts of unstructured opinion data into actionable insights that inform leadership decisions. Since our work is completely modular and published under free and open-source software license, any component can be independently re-used, improved upon, and incorporated into existing pipelines. We identify opportunities for further development that enable iterative improvement of these capabilities.

By leveraging the advancements highlighted in this research, policymakers and leaders can more effectively filter and understand public sentiments, ensuring discussions remain relevant and constructive, reducing the risk of misinformation while preserving diverse viewpoints. This ability to quickly grasp the public consensus on various issues, identify emerging trends, and respond to public concerns in a timely manner ensures that decision-making is informed by a comprehensive understanding of public needs and preferences. Furthermore, the research underscores the importance of ethical considerations in the use of AI for public discourse analysis, serving as a guideline for policymakers and platform developers to implement AI solutions responsibly, ensuring fairness and inclusivity in public engagements. Through effective public engagement, a broader range of voices is heard and considered in the policymaking process. Additionally, the generation of

argument maps is useful for policymakers and leaders in understanding the complexity of public opinion, identifying points of agreement and contention, and communicating policy decisions back to the public in an accessible format. Aligned with the concept of General Collective Intelligence (GCI)[31], the research suggests that LLMs can enhance the general problem-solving ability of groups, tapping into a more effective method for collective decision-making, where the insights generated from public opinion data lead to more resilient and widely accepted policies.

### Limitations and Challenges

During the course of our research, we encountered several limitations that underscore the complexity of this pipeline. The Computational Democracy Project published a few datasets from years ago that were crucial in developing our technique, which can benefit from more and newer data to accurately model and understand contemporary public sentiment. Furthermore, implementation of Polis content moderation guidelines is left to the discretion of the hosting organization, which often resulted in a lack of standardized moderation practices across datasets, allowing for the inclusion of comments irrelevant to the discussion topic.

The interpretability of language models poses another significant hurdle, with models showing high sensitivity to prompt changes, leading to inconsistent results that are challenging to measure and control in an experimental setting. These unique prompting techniques require trial-and-error testing but do not transfer well to another language model. For example, the prompt that we developed on Mistral 7B and Mixtral 8x7B instruct models did not work well at all on the Llama 13b chat model. Varying model types and sizes would have resulted in a much broader scope of analysis, but considering that these models are frequently updated, any findings related to the currently available LLMs would have become outdated within months. We focused on laying a foundation for a generalized technique that would continue to stay relevant for a long time.

Another issue is with the limited context window length of the models that commonly persists. Though the newer models are capable of handling much larger contexts, we found that their reasoning capabilities declined sharply with increased context size, necessitating careful management of context sizes within our pipeline. We overcame this by using the guidance framework that allowed us to break away from the traditional LLM prompting-response loop and instead closely control the LLM context and generations. However, the problem of duplicate detection in comment moderation could not be solved efficiently while maintaining high accuracy. This task proved extremely difficult for LLMs as their context memory footprint expands linearly with the number of comments, significantly degrading the model’s reasoning abilities with the entire dataset of comments in memory. We propose addressing it using text embeddings instead.

The semantic extraction step of the pipeline carries the risk of generating false positives and negatives that are very difficult to detect and filter. This could be mitigated by either employing multiple models in parallel and comparing their outcomes, or using different models sequentially to have one model check the other’s output. Either of these techniques would significantly increase computational costs but improve the accuracy of decisions.

These limitations further emphasize the challenges in applying LLMs to real-world policymaking and public opinion analysis. The variation in comment moderation standards across discussions and organizations is difficult to model and implement, especially considering that relevance to a given topic is often subjective. Topic modeling is a critical component of our pipeline that also proved to be a complex task requiring manual parameter tuning for optimal results. While we employed a simplified approach by heuristically estimating parameters, a more precise approach would require a manual understanding of the underlying dataset. Semantic extraction and pairing of arguments and statements require a more rigorous verification process to ensure the soundness of the resulting estimation of support for each generated insight. These challenges highlight the need for ongoing development and refinement of these tools to generate more timely and valuable

insights.

### Future Research Directions

We plan to focus our ongoing efforts on three major areas to further develop our methodologies for semantic extraction and reasoning, exploring connections across different topics, and expanding our approaches to additional platforms. Each of these areas aims to deepen our analysis and adapt our strategies to a broader range of deliberative environments, ensuring more robust and versatile applications of our findings.

#### *Semantic Extraction and Reasoning in Earlier Stages*

In our analysis, we treat each comment posted as either highlighting a problem or suggesting a solution, drawing inspiration from the IBIS methodology where a tree comprises issues, positions responding to those issues, and arguments supporting or challenging these positions. Polis, by design, prevents direct responses to other comments, rendering each comment as an isolated statement without any discussion around it. The voting data from Polis, while invaluable for gauging public opinion, does not align seamlessly with the IBIS framework. Future work will involve extracting issues, claims, and positions from these individual comments, meticulously documenting the connection between these nodes, and constructing an argument tree, which will be substantiated with the available voting data. This approach promises a more comprehensive analysis of the discussions by enabling deeper insights that connect ideas across disparate comments and topics.

### *Connections Across Topics*

In our methodology, the identification and characterization of connections between statements and arguments is restricted to their designated topics. This process is computationally intensive, and therefore we limit the search space to optimize the complexity of this task while maximizing the relevance of the results. This approach, although efficient, potentially overlooks the broader benefits of the exploration of these semantic connections across the entire dataset. Expanding our analytical framework to identifying and understanding connections across the entire dataset, especially when combined with more precise characterization of statements as issues, claims, or positions, will significantly increase the value derived from our analysis.

### *More Platforms*

We intend to generalize our proposed techniques to other major discussion platforms such as Kialo, which is a web-based platform designed for structured debates and critical thinking, primarily designed for educational use. It is particularly relevant for fostering constructive and logical discussions on a wide range of topics, from social issues to academic subjects. Kialo's structured format is its most distinctive feature, which sets it apart from typical online forums or comment sections. It promotes rational discourse by encouraging users to focus on the merits of each argument. An earlier study [2] identified that the users on such platforms tend to perceive that they are responding to ideas and not people, reducing the likelihood of discussions devolving into personal attacks or emotional responses. While Polis is designed for large-scale public consultations and employs data analytics for large-scale engagement and sentiment analysis, Kialo focuses on structured, critical discussions on specific topics, making it ideal for educational and detailed deliberative processes. Kialo organizes debates in a tree-like structure, where the main topic branches into pros and cons,

which are then further evaluated for their impact. This structure enables users to explore complex topics in a detailed and organized manner, promoting clarity and depth in discussions, as evidenced in the MIT Collaboratorium experiments [14]. Since we spend a considerable amount of effort in organizing comments in a structure, having a pre-built tree will significantly improve our pipeline and more importantly, help us to generalize our technique to a broader set of platforms.

### Final Contemplation

The integration of LLMs into democratic deliberations and policy development carries tremendous potential in harnessing collective intelligence for the betterment of society. The newer language models promise a significant boost in reasoning capabilities that would not have been possible a year ago. This is an unprecedented opportunity to deepen democratic engagement and tailor policy responses while enabling the public to engage in deep dialogue in a historically divided opinion landscape. This alignment between public sentiment and policy action would not only enhance the legitimacy of democratic institutions but also ensure that governance is more responsive, informed, and inclusive. Ultimately, we hope that this research will result in tools that better assist policymakers and leaders in making efficient and equitable decisions that improve consensus and strengthen democratic institutions.



## LIST OF REFERENCES

- [1] Rania Albalawi, Tet Hin Yeap, and Morad Benyoucef. Using Topic Modeling Methods for Short-Text Data: A Comparative Analysis. *Frontiers in Artificial Intelligence*, 3:42, 2020.
- [2] Simon Buckingham Shum, Maarten Sierhuis, Jack Park, and Matthew Brown. Software Agents in Support of Human Argument Mapping. In *Proceedings of the 2010 conference on Computational Models of Argument: Proceedings of COMMA 2010*, pages 123–134, NLD, August 2010. IOS Press.
- [3] E. Jeffrey Conklin. *Dialogue mapping: building shared understanding of wicked problems*. Wiley, Chichester, England ; Hoboken, NJ, 2006. OCLC: ocm60491946.
- [4] Jeff Conklin and Michael L. Begeman. gIBIS: a hypertext tool for exploratory policy discussion. *ACM Transactions on Information Systems*, 6(4):303–331, October 1988.
- [5] Shiyao Ding and Takayuki Ito. Self-Agreement: A Framework for Fine-tuning Language Models to Find Agreement among Diverse Opinions, May 2023. arXiv:2305.11460 [cs].
- [6] James S. Fishkin. *When the People Speak: Deliberative Democracy and Public Consultation*. Oxford University Press, August 2011.
- [7] Maarten Grootendorst. BERTopic: Neural topic modeling with a class-based TF-IDF procedure, March 2022. arXiv:2203.05794 [cs].
- [8] Rafik Hadfi and Takayuki Ito. Augmented Democratic Deliberation: Can Conversational Agents Boost Deliberation in Social Media? In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems, AAMAS '22*, pages 1794–1798, Richland, SC, May 2022. International Foundation for Autonomous Agents and Multiagent Systems.

- [9] Luca Iandoli, Ivana Quinto, Paolo Spada, Mark Klein, and Raffaele Calabretta. Supporting argumentation in online political debate: Evidence from an experiment of collective deliberation. *New Media & Society*, 20(4):1320–1341, April 2018. Publisher: SAGE Publications.
- [10] Takayuki Ito, Rafik Hadfi, and Shota Suzuki. An Agent that Facilitates Crowd Discussion. *Group Decision and Negotiation*, 31(3):621–647, June 2022.
- [11] Paul A. Kirschner, Simon J. Buckingham Shum, Chad S. Carr, Dan Diaper, and Colston Sanger, editors. *Visualizing Argumentation: Software Tools for Collaborative and Educational Sense-Making*. Computer Supported Cooperative Work. Springer, London, 2003.
- [12] M. Klein. How to Harvest Collective Wisdom on Complex Problems: An Introduction to the MIT Deliberatorium, 2012.
- [13] Mark Klein. Crowd-Scale Deliberation For Complex Problems: A Progress Report, March 2022.
- [14] Mark Klein and Luca Iandoli. Supporting Collaborative Deliberation Using a Large-Scale Argumentation System: The Mit Collaboratorium, February 2008.
- [15] W. Kunz and H.W.J. Rittel. *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, 1970.
- [16] Xianming Li and Jing Li. Angle-optimized Text Embeddings, November 2023. arXiv:2309.12871 [cs].
- [17] Thomas W. Malone. How Can Human-Computer “Superminds” Develop Business Strategies? In Jordi Canals and Franz Heukamp, editors, *The Future of Management in an AI World*, pages 165–183. Springer International Publishing, Cham, 2020.

- [18] Leland McInnes, John Healy, and James Melville. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction, September 2020. arXiv:1802.03426 [cs, stat].
- [19] Colin Megill, Elizabeth Barry, and Christopher Small. "Coherent Mode" for the World's Public Square, November 2022. arXiv:2211.12571 [cs].
- [20] Jeff Z. Pan, Simon Razniewski, Jan-Christoph Kalo, Sneha Singhanian, Jiaoyan Chen, Stefan Dietze, Hajira Jabeen, Janna Omeliyanenko, Wen Zhang, Matteo Lissandrini, Russa Biswas, Gerard de Melo, Angela Bonifati, Edlira Vakaj, Mauro Dragoni, and Damien Graux. Large Language Models and Knowledge Graphs: Opportunities and Challenges, August 2023. arXiv:2308.06374 [cs].
- [21] Jack Park. Toward the development of a real-time expert system. In *1986 Rochester FORTH Conference Proceedings, Rochester NY*, pages 23–33, 1986.
- [22] Jack Park. Topic Mapping: A View of the Road Ahead. In Lutz Maicher and Jack Park, editors, *Charting the Topic Maps Research and Applications Landscape*, Lecture Notes in Computer Science, pages 1–13, Berlin, Heidelberg, 2006. Springer.
- [23] Jack Park. Knowledge Gardening as Knowledge Federation. In *Proceedings of the First International Workshop on Knowledge Federation*, volume 552, Dubrovnik, Croatia, 2008. CEUR Workshop Proceedings.
- [24] Errikos Pitsos. Kialo: Empowering Reason, January 2023.
- [25] S. B. Shum. Cohere: Towards Web 2.0 Argumentation. In *Proceedings of the 2008 conference on Computational Models of Argument: Proceedings of COMMA 2008*, pages 97–108, NLD, June 2008. IOS Press.

- [26] S. B. Shum, A. Selvin, M. Sierhuis, J. Conklin, Charles B. Haley, and B. Nuseibeh. Hypermedia Support for Argumentation-Based Rationale. In *Rationale Management in Software Engineering*, pages 111–132. Springer, Berlin, Heidelberg, 2006.
- [27] Christopher Small. Polis: Scaling Deliberation by Mapping High Dimensional Opinion Spaces. *RECERCA. Revista de Pensament i Anàlisi*, July 2021.
- [28] Christopher T. Small, Ivan Vendrov, Esin Durmus, Hadjar Homaei, Elizabeth Barry, Julien Cornebise, Ted Suzman, Deep Ganguli, and Colin Megill. Opportunities and Risks of LLMs for Scalable Deliberation with Polis, June 2023. arXiv:2306.11932 [cs].
- [29] Christian Voigt. Argdown and the Stacked Masonry Layout: Two User Interfaces for Non-Expert Users. In *Computational Models of Argument*, pages 483–484. IOS Press, 2014.
- [30] Taylor Webb, Keith J. Holyoak, and Hongjing Lu. Emergent analogical reasoning in large language models. *Nature Human Behaviour*, 7(9):1526–1541, July 2023.
- [31] Andy E. Williams. Are wicked problems a lack of general collective intelligence? *AI & SOCIETY*, 38(1):343–348, February 2023.
- [32] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing Reasoning and Acting in Language Models, March 2023. arXiv:2210.03629 [cs].