AM4DSP: Argumentation Mining in Structured Decentralized Discussion Platforms for Deliberative Democracy

Sofiane Elguendouze[1] **Lucas Anastasiou**[2] **Erwan Hain**[1]
Elena Cabrio[1] **Anna De Liddo**[2] **Serena Villata**[1]
[1]Université Côte d'Azur, CNRS, Inria, I3S, France
[2]The Open University, Milton Keynes, United Kingdom
{name.lastname}@univ-cotedazur.fr,
{name.lastname}@open.ac.uk

Abstract

Argument(ation) mining (AM) is the automated process of identification and extraction of argumentative structures in natural language. This field has seen rapid advancements, offering powerful tools to analyze and interpret complex and large discourse in diverse domains (political debates, medical reports, etc.). In this paper we introduce an AM-boosted version of BCause, a large-scale deliberation platform. The system enables the extraction and analysis of arguments from online discussions in the context of deliberative democracy, which aims to enhance the understanding and accessibility of structured argumentation in large-scale deliberation processes.

1 Introduction

Deliberative democracy is a form of democracy in which citizens actively participate in public deliberation (Council of Europe, 2023), considering different perspectives and engaging in thoughtful discussion before decisions are made, which provides a mean to improve policy outcomes. This form of democracy aims to enhance policy outcomes by ensuring that public reasoning is central to the decision-making process (Elstub, 2018).

In today's digital age, technologies have expanded the potential reach of public deliberation. However, effectively scaling these deliberative practices while maintaining quality discourse remains a significant challenge (Klein, 2012). Collective deliberation can become a bottleneck in decision-making due to the complexity of processing large volumes of arguments and the diversity of viewpoints. This situation calls for innovative approaches to large-scale civic engagement.

In this context of large-scale deliberations, our work represents a systematic effort to address these challenges through an integrated socio-technical approach that combines Natural Language Processing (NLP), Computational Models of Arguments and Deliberative Democratic practices. More precisely, this paper presents the AM4DSP system, that combines the analytical power of argument mining (AM) methods and leverages BCause[1] strength in organising discussions in argumentative structures.

Effective deliberation platforms, such as BCause, are designed to transcend the limitations of chronologically organized discussions typically found in traditional forums (Anastasiou, 2023). To achieve this goal, this type of platforms incorporate features that organize discussions and structure arguments which significantly improve clarity, facilitate connections between ideas, and promote more informed contributions from participants (Rinner, 2006). Additionally, implementing argument-centric structures that allow users to directly reply to specific points helps maintain focus and enhances sense-making throughout the deliberation process (Irani et al., 2024).

Large-scale deliberations, which are characterized by multiple voices and perspectives, often lead to an overwhelming volume of textual information. This makes them a natural domain for the application of AM, which can automatically extract, analyze, and evaluate arguments from large civic discussions (Lawrence and Reed, 2019; Stede and Schneider, 2019). By leveraging NLP and deep learning methods, AM performs an automated detection of the arguments expressed in participants' statements, their structure and the interactions between them. This can improve decision and policy making by providing policy-makers with a clear view of the underlying positions and justifications presented by participants, feedback on the dynamics of the discussion, key areas of conflict etc. Moreover, AM helps contextualizing arguments by linking them to broader discussions, making it easier to track the evolution of ideas and identify trends in participants' preferences.

[1]https://bcause.app/

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 796–805

2 Related Work

2.1 Deliberation Platforms

Large-scale online deliberation platforms have evolved to address complex societal issues where traditional tools like email, forums and wikis fall short (Klein, 2015). Early systems such as MIT's Deliberatorium (Klein, 2011) and Consider.it (Kriplean et al., 2012) introduced approaches to support large-scale argumentation by reducing redundancy and encouraging clarity in exploring complex problems.

More recent platforms like DebateVis (South et al., 2020) and Kialo (Mei et al., 2024) incorporate visualization techniques to present argumentative structures intuitively, enabling participants to comprehend key discussion points and positioning along opinion spectrums (Shum et al., 2000).

However, these platforms face scaling challenges with large crowds. A balance must be maintained between structured argumentation and natural conversation flow to prevent platforms from becoming overly formal while still organizing argumentative structures effectively (Iandoli et al., 2009).

2.2 Argument(ation) Mining

Argument(ation) Mining (AM) (Cabrio and Villata, 2018; Vecchi et al., 2021) deals with the automated analysis of argumentation structures in written and oral texts (Lawrence and Reed, 2020) from various domains, such as legal cases, persuasive essays, scientific articles, user-generated content, and political debates. The ability of identifying argumentative components (e.g., Premises, Claims) and predicting their relations (e.g., Attack, Support) in these texts opens the door to cutting-edge tasks like fact-checking, counter-argumentation generation and argument quality assessment.

2.2.1 Argumentation Mining Models

Numerous approaches have been proposed for AM, typically relying on hand-crafted syntactic and lexical features (Stab and Gurevych, 2014b), pre-trained language models (Agarwal et al., 2022) or both (Cocarascu et al., 2020). Most recent works leverage supervised approaches with autoencoding transformer architectures like BERT (Devlin et al., 2019), due to their capabilities in understanding the surrounding context and long-term dependencies of arguments. (Habernal et al., 2024) applied argument mining to decisions from the European Court of Human Rights (ECHR). They built models based on pre-trained BERT and RoBERTa (Liu et al., 2019), and achieved comparable results.

2.2.2 Argumentation Mining Data

Since the beginning of the AM field, political debates and user generated content on debate platforms and social media have been considered as promising scenarios to automatically extract and analyse arguments. For the first, we can mention the dataset of political debates from the US presidential elections called USElecDeb60To16 (Haddadan et al., 2019) spanning several decades, where numerous candidates were engaged in multiple discussions. For the user generated content, online debate platforms such as Kialo[2], idebate.org[3] and the subreddit ChangeMyView[4], provide access to publicly-available participatory conversations annotated with argumentative relations (Agarwal et al., 2022; Mezza et al., 2024).

3 BCause Deliberation Platform

BCause is a structured discussion platform designed for large-scale online deliberations that addresses limitations in traditional social media platforms for supporting quality discourse. Unlike conventional social media that often leads to polarization and divisive interactions, BCause incorporates specialized design elements to promote cohesive discussion and reduce group bias (Anastasiou and De Liddo, 2023).

BCause structures deliberation through a layered argumentative approach following the IBIS (Issue-Based Information System) paradigm (Kunz and Rittel, 1970). At its core, the platform organizes discussions around clearly defined debate topics (the issue), positions (opinions or possible solutions to the topic), and arguments (statements that either support "pro" or oppose "con" the parent position). This structure aims to improve the signal-to-noise ratio and provides logical organization to discussions, see Figure 1.

The platform visualizes argumentative structures through two primary methods: (1) A time-ordered timeline providing chronological context. (2) An argument tree displaying the logical relationship between positions and supporting/opposing arguments (evidences). The IBIS model implemented in BCause organizes content hierarchically around three primary elements: Debate topics (issues to

[2]https://www.kialo.com/
[3]https://idebate.net/
[4]https://www.reddit.com/r/changemyview/

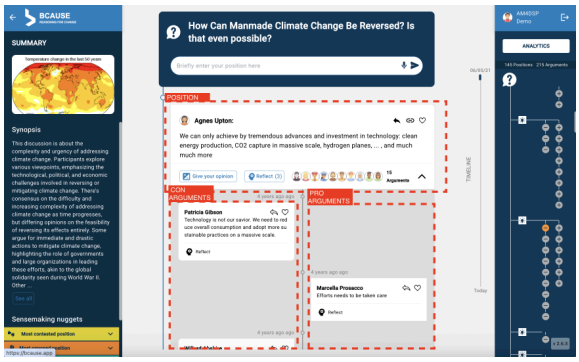

Figure 1: BCause discussion interface with IBIS components annotated

be discussed), Positions (opinions or possible solutions to the topic) and Arguments/Evidences (statements that either support "pro" or oppose "con" a position). This structure intentionally simplifies the argumentative process to make it accessible to non-expert users while still capturing the essential elements of reasoned deliberation. It does however introduce some interesting interplay with classic argumentation models that we explore in the rest of this paper. When the deliberation process is hybrid (involving both face-to-face and online discussions) BCause can provide the online space for continuing a live event, by offering the opportunity for discussion moderators and admins to upload transcript (natural dialogic text among the participants) that occurred during the live event.

4 AM4DSP System Overview

AM4DSP is a boosted version of the BCause deliberation platform, designed to automatically process argumentation structures and relations. These additional functionalities enable the system to perform a deeper analysis of deliberative discussions, extending beyond the standard structure of BCause by integrating AM models at different stages. Figure 2 illustrates the complete architecture of our system.

4.1 System's Main Facilities

The argumentation models integrated into BCause open to a novel range of functionalities, going beyond the traditional setup of discussions on BCause. The main key features are:

- **Discussion-level (statement-level) argumentation,** to automatically analyze user contributions and classify each statement as a position, a supporting or attacking evidence, aligning with BCause's native structure. This would allow to automatically classify and integrate in BCause the arguments put forward by the participants without relying on manual labels.
- **Component-level argumentation,** to break down each statement into its argumentative components (claims and premises) and link them both within and across statements. It offers deeper structural insights and finer-grained analysis for interpreting complex statements where users often blend their main point with supporting evidence.
- **Cross-statement analysis,** to examine the connections between arguments from different participants. It allows identifying broader argumentative structures that span across the full discussion, offering a more comprehensive view of the dialogue and its underlying arguments, that cannot be currently captured by BCause.
- **Transcript analysis,** to analyze longer and more natural dialogues from live events, such as those generated from audio recordings of deliberative events.

While prior user evaluations have been conducted on the core BCause platform, assessing its discussion structure, aesthetics, impact on sensemaking and quality of discussion, these tests did not involve the AM functionalities showcased here. The complete AM4DSP system, including its analytical interface, is scheduled for user testing in real-world use cases as part of the upcoming ORBIS project[5] pilots, which will provide crucial feedback on its usability and deliberation effectiveness.

5 Experimental Setting

In this Section, we present the annotated datasets collected to train the AM models, the models used in the experiments as well as the obtained results.

5.1 Datasets

As discussed earlier in Section 3, the BCause platform structures discourse at the statement level, where each statement (i.e., an argumentative sentence articulated by a participant) is either a position (i.e., a stance taken on a particular argumentative topic) or an evidence that either supports or attacks a given position. We use a sample of structured data from BCause to train the AM models

[5] https://orbis-project.eu/

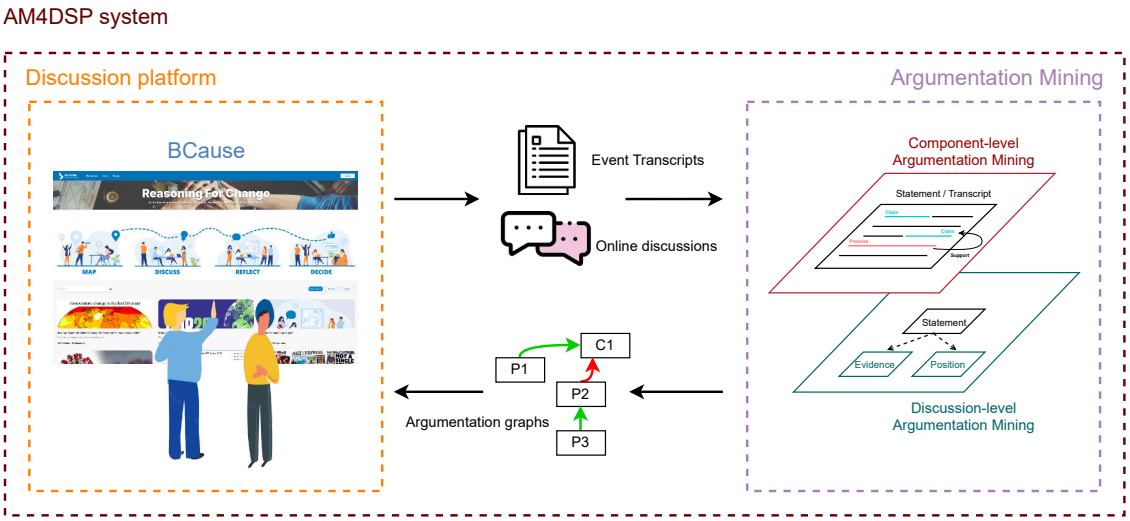

Figure 2: AM4DSP performs argumentation analysis at both the discussion-level and component-level for structured discussions, and provides component-level analysis for transcripts.

targeting what we name *discussion-level argumentation* (Sec. 5.2). In addition to the data from BCause, we have also used Touche23-ValueEval (Mirzakhmedova et al., 2023), a dataset derived from discussions publicly available on internet, mapping the relations *against* and *in support of* with *con* and *pro*, respectively.

Given that we do not have annotated data from BCause to test the *component level* task, to build our classifiers for argumentative components detection and relations detection we used standard datasets as the annotated transcripts of the televised political debates in the US presidential campaigns (1960-2016) (Haddadan et al., 2019), Persuasive Essays (Stab and Gurevych, 2014a) and the dataset presented in (Habernal and Gurevych, 2017).

Tables 1, 2 and 3 provide statistics on the datasets used in our experimental settings (before the train/dev/test splitting).

Dataset	Statement classification		Relation classification	
	Position	Evidence	Attack	Support
BCause	3.9k	5k	2.9	2.2k
Touche23-ValueEval	0.5k	8.7k	3.9k	4.7k

Table 1: Data for *discussion-level argumentation*

Dataset	Argument Components		Argument Relations		
	Claim	Premise	Attack	Support	NoRel
USElecDeb60To16	29k	26k	2.8k	19.8k	23k
Persuasive Essays	2257	3832			
Habernal Gurevych 2017	195	538	-	-	-

Table 2: Data for the *component-level argumentation*

Dataset	O	B-Premise	I-Premise	B-Claim	I-Claim
USElecDeb60To16	566492	26055	350079	29624	338941
Persuasive Essays	35946	2257	29828	3832	59652
Habernal Gurevych 2017	61414	195	3491	538	20566

Table 3: Detailed statistics on datasets used for building *component-level argumentation* models following the BIO-tagging scheme

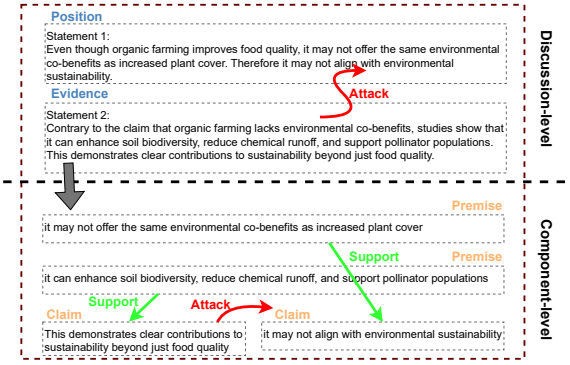

Figure 3: Example showcasing both *discussion-level* and *component-level* argumentation tasks.

5.2 AM Models

Since discussions in BCause adhere to a statement-centric structure, this required an adaptation of the standard AM pipeline to align with the structural requirements of the platform. Two AM steps, namely *discussion-level* and *component-level*, are therefore performed, as exemplified in Figure 3. The first step is performed at the statement level, and its goal is to first classify statements into positions and evidences, and then to identify the relations between

them (support or attack). Importantly, the goal is not to link the evidences to their corresponding parent positions as this is already encoded within BCause's structured data, but rather to uncover additional relationships that can emerge between statements on a broader, discussion-wide scale.

On top of this, the second step concerns a granular analysis performed at the component level following the standard AM pipeline, where argumentative components (premises and claims) within each statement are identified and their relationships within and across statements are determined. As for event transcripts, which are typically longer and more unstructured than the discussions on BCause, only the component argumentation task is applied.

We trained a set of AM models[6] on the datasets described in Section 5.1. We tested BERT-based architectures (RoBERTa and DeBERTa (He et al., 2021)), as well as autoregressive large language models, in particular GPT-2 XL version (1.5B) (Radford et al., 2019) and OPT-1.3B (Zhang et al., 2022). The rationale behind selecting these models despite the availability of more recent and powerful paid alternatives (e.g. GPT-4o) is grounded in the constraints of the ORBIS project. We aim to rely exclusively on open-source and freely available models to ensure long-term sustainability and avoid ongoing operational costs after the project's conclusion.

5.3 Experimental Settings and Results

5.3.1 Discussion-level Argumentation

Statement Classification. We trained our models on a combination of BCause and Touche datasets. The target classes are: Position and Evidence. The training was conducted using a learning rate of 2e-5, a batch size of 32, and a maximum sequence length of 128 tokens, across 5 epochs.

Statement Relation Classification. The target classes are Support and Attack. We use the same dataset and configuration described above, and increase the maximum sequence length to 256.

Table 4 shows the results (f1-macro scores) of our models finetuned on different datasets and evaluated on a test partition composed of 500 new statements obtained from the latest data crawls from BCause. Deberta-v3 model trained on merged data from BCause and Touche achieves the best score for both statement and relation classification.

[6] https://github.com/orbis-marianne/orbis-am-models

Train dataset	Statement classification	Relation classification
Bcause+Touche (deberta-v3)	**0.89**	**0.90**
Bcause+Touche (roberta)	0.83	0.34
Bcause (deberta-v3)	0.71	0.55
Touche (deberta-v3)	0.53	0.65

Table 4: Results for *discussion-level* argumentation

5.3.2 Component-level Argumentation

Component Detection. For BERT-based models, the task is framed as a sequence tagging problem using the standard BIO-tagging scheme. We fine-tune the DeBERTa-v3 model for token classification over 10 epochs with a learning rate of 1e-4 and a maximum sequence length of 64 tokens. For OPT and GPT models, we redesign the task as text generation problem to align with their decoder-only architecture. We convert BIO labels into tagged sequences (e.g., <premise>...</premise>, <claim>...</claim>) and prepare prompts consisting of an instruction describing the task followed by the input sequence in plain text. Figure 4 shows an example of the prompt-response pair used to fine-tune the LLM models. The expected output is a replication of the input plain text with appropriate argument tags inserted.

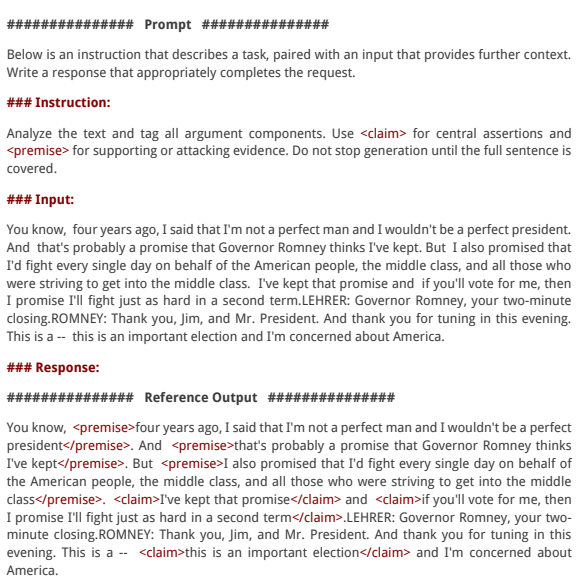

Figure 4: Full prompt and generation example with OPT

To reduce creativity during generation, we apply a low temperature (0.01) and a narrow nucleus sampling threshold (top-p = 0.1), constraining the model to deterministic behavior. Input sequences are chunked into sub-sequences of up to 1024 tokens to fit the models' context limits. Both models are fine-tuned over 10 epochs, with the best checkpoint (based on macro F1 on the validation set)

used for testing. The models are designed to identify three target classes: Premises, Claims, and Non-argumentative tokens. All datasets follow an 80/10/10 split. Table 5 shows the F1 macro results for argument component detection with various models trained on various datasets and evaluated with multiple test configurations (cross testing).

Model	ED	PE	HG	Merged
USElecDeb60To16 (deberta-v3)	0.47	0.47	0.33	0.47
USElecDeb60To16 (roberta)	0.47	0.45	0.30	0.47
Persuasive Essays (deberta-v3)	0.27	0.71	0.24	0.32
Persuasive Essays (roberta)	0.28	0.69	0.41	0.33
Habernal Gurevych 2017 (deberta-v3)	0.18	0.15	0.32	0.18
Habernal Gurevych 2017 (roberta)	0.21	0.22	0.36	0.22
Merged (deberta-v3)	**0.47**	**0.90**	**0.71**	**0.49**
Merged (roberta)	0.46	0.83	0.58	0.48
Merged (fine-tuned OPT-1.3B)	-	-	-	**0.77**
Merged (fine-tuned GPT-2-1.5B)	-	-	-	**0.77**

Table 5: Cross evaluation scores for argument component detection (ED=USElecDeb60To16; PE=Persuasive Essays; HG=Habernal Gurevych 2017)

Table 6 provides a detailed view on the component detection results following various configurations (using different datasets and models). The results are provided in terms of the F1 macro scores obtained on the test partitions of the datasets used for model training.

Component Relation Classification. We fine-tuned the DeBERTa-v3 base model as a sequence classifier, to classify relations into three distinct classes: Support, Attack, and NoRelation. We used a learning rate of 1e-4, with input sequences capped at 64 tokens. The training process spans 10 epochs. Results on USElecDeb60To16 dataset (the only one among the ones we selected, that is annotated with relations) are shown in Table 7.

As expected, algorithms perform best when trained and tested on the same data, and generalize poorly across datasets due to differences in writing style, topic, and argumentation structure. BERT-based models trained on the merged dataset outperform all single-dataset models across all test sets with substantial improvements, especially on PE and HG. This suggests that combining datasets introduces greater variety and richer argumentation patterns, which helps the model generalize better across domains. DeBERTa-v3 consistently outperforms RoBERTa across nearly all configurations, reflecting its stronger representation capacity, especially for this token classification task. The fine-tuned OPT and GPT models achieve the highest F1 score on the merged test set (0.77), showing that generative LLMs can be effectively repurposed for token classification as they may better capture long-range dependencies and context when properly fine-tuned with task-specific formatting. All these experiments allowed us to select the best models to be included in AM4DSP.

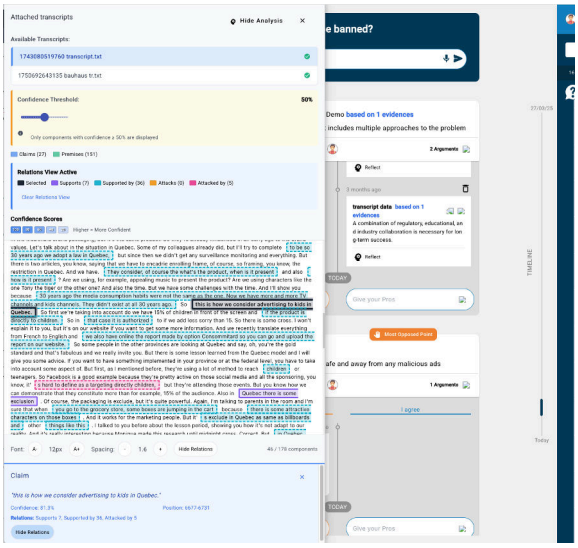

Figure 5: Argument mining applied on attached transcript as rendered in BCause. The selected argumentative component is highlighted in black and the related components in distinct colours: purple for supporting, cyan for supported by, red for attacked by components.

6 AM4DSP Integration in BCause

The AM system is served through a REST API[7] facilitating integration with online deliberation platforms, namely BCause. The code is a Django application, using Django-Rest-Framework to build the REST API. The deployment of the API is heavily dependent on docker and docker-compose. The Django application runs with a Gunicorn server, behind a Nginx SSL proxy that is configured within the docker-compose file.

The system follows a multi-step workflow: (1) Pre-processing, where raw data (statements and transcripts) is cleaned and formatted for analysis, and (2) Argument Mining, where pre-processed data is passed to the AM service for a multi-stage analysis (discussion-level, then component-level). Once identified, argument components and their relationships are returned to BCause in the form of argumentation graphs allowing for interactive visualisation (Figures 5, 6).

[7]API: https://orbis.i3s.unice.fr/api/docs/, source code: https://github.com/orbis-marianne/orbis-argument-mining-tool

Model	B-C	I-C	B-P	I-P	O	Acc	F1-Macro	F1-Avg	Support
USElecDeb60To16 (deverta-v3)	0.39	0.38	0.47	0.44	0.66	0.57	0.47	0.6	194452
USElecDeb60To16 (roberta)	0.39	0.39	0.46	0.46	0.65	0.57	0.47	0.59	194941
Persuasive Essay (deberta-v3)	0.6	0.74	0.58	0.79	0.82	0.75	0.71	0.75	15072
Persuasive Essay (roberta)	0.59	0.68	0.59	0.78	0.19	0.74	0.69	0.74	15241
Habernal Gurevych 2017 (deberta-v3)	0	0	0.4	0.34	0.86	0.76	0.32	0.73	1945
Habernal Gurevych 2017 (roberta)	0	0	0.56	0.41	0.82	0.71	0.36	0.72	1950
Merged (deberta-v3)	0.41	0.43	0.46	0.5	0.63	0.56	0.49	0.58	209865
Merged (roberta)	0.4	0.4	0.46	0.49	0.66	0.57	0.48	0.60	210518
Merged (OPT-1.3B)	0.75	0.73	0.71	0.73	0.92	0.80	0.77	0.80	85288
Merged (GPT-2)	0.76	0.73	0.72	0.72	0.92	0.80	0.77	0.79	89455

Table 6: Detailed f1-scores for argument component detection with various models

Dataset	Relation classification
USElecDeb60To16 (deberta-v3)	**0.69**
USElecDeb60To16 (roberta)	0.61

Table 7: Results obtained for relation classification on the *component-level* argumentation.

Discussion-level argumentation in turn generates an aggregated key argument graph representing the discussion's logical structure. This graph is accessible through BCause's analytics view and unveils hidden argumentative relations, reduces misinterpretations of argument polarity (supporting vs. opposing), and identifies duplicated arguments. Users can explore a synthesised representation of complex issues based on the main arguments while maintaining the ability to seamlessly navigate back to the source data for detailed context.

7 AM workflow in AM4DSP

To demonstrate how the system[8] supports argument mining, we designed a typical workflow that begins with creating a discussion space. In order to populate the discussion with some initial positions and pro/con arguments, users can upload a transcript (e.g., from a public debate or video) through the Import Transcript tool, then automatically extract argument components and visualize them in a preliminary argument tree. At this stage, further user contributions can be added, enabling a mix of automatically extracted and user-supplied arguments.

Once the data are available, the platform provides interactive tools for finer-grained analysis. The transcript viewer highlights argumentative components (premises and claims) directly in the text, with a confidence slider to control the granularity of detection. Selecting a highlighted claim reveals its related premises (support/attack). For a broader overview, the platform generates an Argument Network Analysis, where claims and premises appear as nodes in a force-directed graph, with edges indicating support or attack relations. Users can toggle views to inspect arguments originating from posts, transcripts, or both in combination.

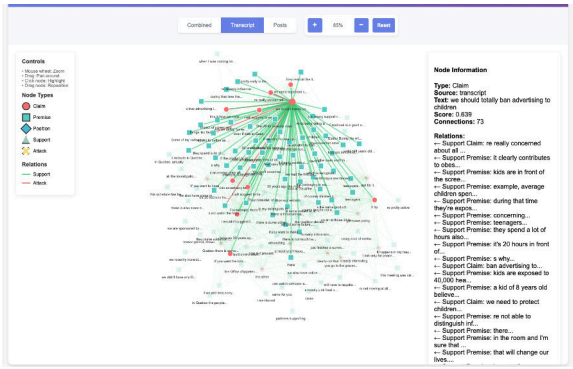

Figure 6: The network visualization of argumentative discourse in BCause analytics, showing interconnected nodes representing claims (red circles), premises (blue squares), and their relationships via support (green lines) and attack (red lines).

8 Conclusion

In this paper, we introduced AM4DSP, an enhanced version of the BCause deliberation platform, augmented with AM capabilities. This publicly available open-access system extends the basic discussion schema by incorporating a dual-level argumentative analysis: a higher-level discussion (or statement) analysis and a fine-grained component-level argumentation. AM4DSP also provides broader insights by identifying cross-statement relations, enabling the mapping of argumentative connections across multiple statements and contributions from different participants. The argumentation mining framework can be integrated into other platforms beyond BCause via a REST API.

[8] A screencast video demonstrating the system can be found at the following link https://youtu.be/c4uMSdTnm5k; the live demo can be accessed at `https://bcause.app/discussions/-OMMdBg09OoAPTFhifkR`

Limitations

This work demonstrates the potential of providing an advanced argument mining service directly integrated into the UX of existing deliberation systems. However, the value and effectiveness of the argument extraction from the perspective of human evaluation and the impact this might have on the quality of human deliberation have yet to be evaluated. This also depends on factors, such as the UX of the chosen deliberation system and, as we have pointed out, the data structure used. Future research should be aimed at conducting user studies in which these aspects can be considered more thoroughly. A key limitation of our current system is the reliance on large-scale, high-quality annotated datasets for building accurate Argument Mining models, particularly at the component level, where data scarcity is a significant challenge. Unlike other NLP tasks, Argument Mining requires domain-specific expertise for annotation, making the process costly and labor-intensive. Recent advancements in large language models (LLMs) offer potential solutions through techniques like in-context learning and few-shot learning, where models such as Mistral and Llama can generalize tasks with minimal data. However, despite their potential, these models are often impractical for everywhere deployment due to their size, computational demands, and concerns over data privacy (when relying on external LLM APIs). Addressing these challenges will require continued research into lightweight, domain-adapted LLMs that balance performance, efficiency, and usability in practical AM scenarios.

Ethical considerations

The integration of Argument Mining (AM) into BCause opens valuable opportunities for structured analysis of deliberative discussions, but it also presents some ethical risks. Automated analysis could be misused to manipulate or bias discussions, for example by selectively highlighting arguments to favor certain positions, misclassifying user contributions to downplay dissent, or profiling participants based on their expressed opinions. These risks are particularly concerning in politically sensitive or high-stakes decision-making contexts, where the neutrality and transparency of the system are crucial.

To mitigate such risks, several safeguards are/should be implemented. First, the models and datasets used in AM are openly documented to ensure transparency about their training, limitations, and potential biases. Second, user privacy is preserved by anonymizing data, securely storing it, and never transfer it outside trusted environments. The final system should remain a support tool rather than a decision-maker: human oversight is essential to validate outputs and contextualize results. Finally, mechanisms for accountability should be put in place, such as audit trails and explainable AI components, to allow stakeholders to understand and contest the system's reasoning when necessary.

Acknowledgments

This work has been supported by the European Project ORBIS "Augmenting participation, co-creation, trust and trasparency in Deliberative Democracy at all scales" under the Horizon Europe Programme (Grant Agreement No. 101094765). This work has also been partially supported by the French government, through the 3IA Cote d'Azur investments in the project managed by the National Research Agency (ANR) with the reference number ANR-23-IACL-0001, and by UKRI under the UK Government's Horizon Europe Guarantee scheme (Reference Number: 10048874).

References

Vibhor Agarwal, Sagar Joglekar, Anthony P. Young, and Nishanth Sastry. 2022. Graphnli: A graph-based natural language inference model for polarity prediction in online debates. In *Proceedings of the ACM Web Conference 2022*, WWW '22, page 2729–2737, New York, NY, USA. Association for Computing Machinery.

Lucas Anastasiou. 2023. Computational argumentation approaches to improve sensemaking and evidence-based reasoning in online deliberation systems. Unpublished.

Lucas Anastasiou and Anna De Liddo. 2023. BCause: Reducing group bias and promoting cohesive discussion in online deliberation processes through a simple and engaging online deliberation tool. In *Proceedings of the First Workshop on Social Influence in Conversations (SICon 2023)*, pages 39–49, Toronto, Canada. Association for Computational Linguistics.

Elena Cabrio and Serena Villata. 2018. Five years of argument mining: A data-driven analysis. In *IJCAI*, volume 18, pages 5427–5433.

Oana Cocarascu, Elena Cabrio, Serena Villata, and Francesca Toni. 2020. Dataset independent baselines for relation prediction in argument mining. In *Computational Models of Argument*, pages 45–52. IOS Press.

Council of Europe. 2023. Report on deliberative democracy. Accessed: 2023-02-25.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Stephen Elstub. 2018. Deliberative and participatory democracy. *The Oxford handbook of deliberative democracy*, pages 187–202.

Ivan Habernal, Daniel Faber, Nicola Recchia, Sebastian Bretthauer, Iryna Gurevych, Indra Spiecker genannt Döhmann, and Christoph Burchard. 2024. Mining legal arguments in court decisions. *Artificial Intelligence and Law*, 32(3):1–38.

Ivan Habernal and Iryna Gurevych. 2017. Argumentation mining in user-generated web discourse. *Computational Linguistics*, 43(1):125–179.

Shohreh Haddadan, Elena Cabrio, and Serena Villata. 2019. Disputool – a tool for the argumentative analysis of political debates. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 6524–6526. International Joint Conferences on Artificial Intelligence Organization.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. *Preprint*, arXiv:2006.03654.

Luca Iandoli, Mark Klein, and Giuseppe Zollo. 2009. Enabling on-line deliberation and collective decision-making through large-scale argumentation: A new approach to the design of an internet-based mass collaboration platform. *Int. J. Decis. Support Syst. Technol.*, 1:69–92.

Arman Irani, Michalis Faloutsos, and Kevin Esterling. 2024. Argusense: Argument-centric analysis of online discourse. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, pages 663–675.

Mark Klein. 2011. The mit deliberatorium: Enabling large-scale deliberation about complex systemic problems. *2011 International Conference on Collaboration Technologies and Systems (CTS)*, pages 161–161.

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

Mark Klein. 2015. A critical review of crowd-scale online deliberation technologies.

Travis Kriplean, Jonathan Morgan, Deen Freelon, Alan Borning, and Lance Bennett. 2012. Supporting reflective public thought with considerit. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*, CSCW '12, page 265–274, New York, NY, USA. Association for Computing Machinery.

W. Kunz and H. W. Rittel. 1970. *Issues as elements of information systems (Vol. 131, p. 14)*. Institute of Urban and Regional Development, University of California, Berkeley, CA.

John Lawrence and Chris Reed. 2019. Argument mining: A survey. *Computational Linguistics*, 45(4):765–818.

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

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Bing Mei, Peipei Xiong, and Hongyu Xu. 2024. Kialo edu. *RELC Journal*, page 00336882231226156.

Stefano Mezza, Wayne Wobcke, and Alan Blair. 2024. Exploiting dialogue acts and context to identify argumentative relations in online debates. In *Proceedings of the 11th Workshop on Argument Mining (ArgMining 2024)*, pages 36–45, Bangkok, Thailand. Association for Computational Linguistics.

Nailia Mirzakhmedova, Johannes Kiesel, Milad Alshomary, Maximilian Heinrich, Nicolas Handke, Xiaoni Cai, Barriere Valentin, Doratossadat Dastgheib, Omid Ghahroodi, Mohammad Ali Sadraei, Ehsaneddin Asgari, Lea Kawaletz, Henning Wachsmuth, and Benno Stein. 2023. The touché23-valueeval dataset for identifying human values behind arguments. *Preprint*, arXiv:2301.13771.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI*, 1(8).

Claus Rinner. 2006. Argumentation mapping in collaborative spatial decision making. In *Collaborative geographic information systems*, pages 85–102. IGI Global.

Simon Buckingham Shum, Albert M. Selvin, and White Plains. 2000. Structuring discourse for collective interpretation.

Laura South, Michail Schwab, Nick Beauchamp, Lu Wang, John P. Wihbey, and Michelle A. Borkin. 2020. Debatevis: Visualizing political debates for non-expert users. *2020 IEEE Visualization Conference (VIS)*, pages 241–245.

Christian Stab and Iryna Gurevych. 2014a. Annotating argument components and relations in persuasive essays. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1501–1510, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Christian Stab and Iryna Gurevych. 2014b. Identifying argumentative discourse structures in persuasive essays. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 46–56.

Manfred Stede and Jodi Schneider. 2019. *Argumentation Mining*. Springer International Publishing.

Eva Maria Vecchi, Neele Falk, Iman Jundi, and Gabriella Lapesa. 2021. Towards argument mining for social good: A survey. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1338–1352, Online. Association for Computational Linguistics.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, and 1 others. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.