Towards Argument Mining for Social Good: A Survey

Eva Maria Vecchi, Neele Falk, Iman Jundi and **Gabriella Lapesa**
Institute for Natural Language Processing
University of Stuttgart (Germany)
first[-middle].last@ims.uni-stuttgart.de

Abstract

This survey builds an interdisciplinary picture of Argument Mining (AM), with a strong focus on its potential to address issues related to Social and Political Science. More specifically, we focus on AM challenges related to its applications to social media and in the multilingual domain, and then proceed to the widely debated notion of argument quality. We propose a novel definition of argument quality which is integrated with that of deliberative quality from the Social Science literature. Under our definition, the quality of a contribution needs to be assessed at multiple levels: the contribution itself, its preceding context, and the consequential effect on the development of the upcoming discourse. The latter has not received the deserved attention within the community. We finally define an application of AM for Social Good: (semi-)*automatic moderation*, a highly integrative application which (a) represents a challenging testbed for the integrated notion of quality we advocate, (b) allows the empirical quantification of argument/deliberative quality to benefit from the developments in other NLP fields (i.e. hate speech detection, fact checking, debiasing), and (c) has a clearly beneficial potential at the level of its societal thanks to its real-world application (even if extremely ambitious).

1 Introduction

Considering Argument Mining (AM) for Social Good implies a strong conceptual shift: the discourse exchange is not to be interpreted as a competition to be won by the most persuasive contribution[1], but rather as a cooperative endeavor in which each individual contribution represents a move towards a shared goal. If argumentative discourse is cooperation, it is not to be taken for granted that the perfect debater, most often the primary objective in AM research, is necessarily also the best team player.

Building on this assumption, we review recent developments in the field of AM from the perspective of its application in socially relevant contexts. Our survey has a strong interdisciplinary perspective, putting the focus on the collaboration between NLP and the Social Sciences and, more specifically, in argumentation targeted at decision-making (*deliberation*). Deliberative discourse historically characterizes parliamentary debates; however, it pervades, more and more frequently, discussions in digital democracy forums and, beyond that, specific strands of discussions in "generalistic" social media. Looking at argumentation through the lens of deliberation has a 2-fold benefit. From a purely NLP perspective, the insights gained through modeling deliberative features can in turn be employed in applications targeting discourse in deliberative forums and social media more broadly, allowing systems to be more adaptable to real-world discourse settings. Social Sciences, in turn, can enormously benefit from the possibility of scaling up to a larger public with the support of NLP methods.

The novelty of this survey with respect to literature (Cabrio and Villata, 2018; Lawrence and Reed, 2019) is precisely in its interdisciplinary focus, which leads us to a novel formulation of the widely debated notion of argument quality (Wachsmuth et al., 2017a,b), which we put in direct comparison to Deliberative Quality (Bächtiger and Parkinson, 2019). The take-home message of this comparison is that the quality of a contribution to an argument cannot only be quantified in terms of its textual (linguistic/logical) properties and the relation to the preceding contributions (as commonly done

[1] In this paper, we use the term "contribution" to refer to a turn in a discourse exchange; more concretely a contribution is a textual unit in a discourse contex, e.g., a post in a forum, a tweet in a discussion thread; a speech in a parliamentary debate).

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, pages 1338–1352
August 1–6, 2021. ©2021 Association for Computational Linguistics

in argument quality), but also the relation to the "cooperation challenge" needs to be brought in the picture. In other words, a good contribution is one that ensures the discourse to unfold productively.[2]

We conclude the survey by defining the conceptual coordinates and the practical challenges of (semi-)*automatic moderation*, a highly integrative application of AM for Social Good which represents a natural testbed for the integrated definition of quality discussed above. We propose to implement moderation as a form of discourse optimization, and spell out the objective of such optimization – that is to say, the desiderata for an NLP-based moderator. We discuss the concrete challenges related to the tasks of an NLP moderator, and review existing work that, albeit not targeted at NLP moderation directly, can be brought in as part of a puzzle which is both ambitious and worthwhile to pursue.

2 Argument Mining

Argument(ation) Mining (AM) is a field encompassing varying tasks that deal with the automated analysis of arguments from natural language text. Habernal and Gurevych (2017) defines AM as "the general task of analyzing discourse on the pragmatics level and applying a certain argumentation theory to model and automatically analyze the data at hand". The progress in the field of NLP in recent years has also influenced this research area: automatic recognition and identification of arguments has been enabled in various domains and different models for the analysis and representation of argumentative structure have been developed. Furthermore, there is a growing research interest in other aspects of AM, such as argument quality.

2.1 Framework

Cabrio and Villata (2018) provide an elaborate overview of the AM framework in their data-driven analysis of the state of the art after five years of significant developments in the field of AM. Generally speaking, given a collection of natural language texts, the task at hand is implemented in two stages:

Argument extraction The system first identifies the documents which contain the argumentative structure and the specific textual spans in which argumentation is encoded. Once the textual boundaries are defined, subportions of the argumentative spans are assigned to a set of pre-established argument components (e.g. claims, premises, rebuttal, etc.). A variety of models were used for this including Näive Bayes (Moens et al., 2007), SVMs (Mochales and Moens, 2011), RNNs (Niculae et al., 2017; Eger et al., 2017), Pre-trained Language Models (Chakrabarty et al., 2019; Lugini and Litman, 2020), and other supervised–learning techniques (Ein-Dor et al., 2020).

Relation assignment The goal of the second stage is to model the relations between the argumentative spans identified in the first stage. These relations can exist between different arguments (support, attack) as well as within an argument (connecting the premises with the claim). Recent approaches to argumentative relation classification investigate for example relational models (Trautmann et al., 2020) or inject background knowledge by leveraging features from different knowledge bases (Kobbe et al., 2019). Detecting these relations is necessary to model the overall structure of the argumentation (discourse/debate). As this structure can be complex, the task is difficult, involving high-level knowledge representation and reasoning issues. After the relations are detected, the discourse structure can then be mapped to a graph representation, called argumentation graph, with the arguments as nodes and relations as edges. To simplify the problem, some approaches reduce the graph to a tree-structure representation (Peldszus and Stede, 2015; Stab and Gurevych, 2017). Different methods to generate the structure have been investigated, e.g. SVMs (Habernal and Gurevych, 2017; Niculae et al., 2017) or textual entailment (Cabrio and Villata, 2013; Cocarascu et al., 2020). Modeling the relations and argumentation flow within a debate is an important factor when defining the notion of argument quality, which will be presented in Section 3.

Consider the following example taken from an online debate about compulsory vaccinations[3] which demonstrates the framework quite clearly. Given a statement presenting background and context, participants are asked to discuss the question *"Does public health demand vaccinations?"* (**Claims** are in bold, and <u>premises</u> are underlined.)

[2] The productive quality of a contribution can be defined in relation to Social Sciences literature (Steenbergen et al., 2003; Steiner et al., 2005), c.f. Section 3

[3] `http://debatepedia.idebate.org/en/index.php/Debate:_Compulsory_vaccination`

A_1: ***A vaccine is the best way to prevent an outbreak of a disease or to reduce its negative effects.*** *Vaccinated people become immune to a certain pathogen and do not develop a disease. Although there are occasionally side effects, these affect only a tiny number of people compared to the protection offered to the vast majority.*

A_2: ***Many vaccines have serious and sometimes deadly side effects****. With many vaccines the immunity is not lifelong. Sometimes the vaccines itself can cause a serious disease to develop as a side effect. If governments know that compulsory mass vaccination is likely to cause death or permanent disability in even a few cases, it is immoral for them to make it compulsory.*

Here, the argumentative text boundaries are first determined from the natural language discussion and the argument components (claims and premises) are extracted. Then, the relations between the two arguments are as follows: A_1 supports the argument while A_2 attacks it.

However, consider another example, extracted from an online debate platform *Kialo*[4]. Here, the participants' contribution and the structure mirror a more direct and conversational dynamic to argumentation.

A_1: ***Marvel Universe is better than DC Universe****.*

A_2: *Stan Lee's vision contains clarity and purpose, while DC is simply interested in churning entertainment to the masses.*

A_3: *Stan Lee no-longer has control over any of marvel, which can cloud the purpose of Marvel due to it being owned by Disney.*

A_4: *This is especially true due to his unfortunate passing.*

A_5: ***DC has been more apt to recycle parts of Intellectual Property****, they even made an entire movie using the ideas of the 1960's characters and comics.*

The seemingly simple example of an online exchange shows how a more conversational environment provides vaguer boundaries of argumentation structure and components. Each argument is more direct, not necessarily consisting of a *claim-premise* configuration, and the strength and productive quality of each argument is particularly relative to the context, each contribution affecting the argument differently either at a local or global level. Note, however, that the relations between arguments and claim are still relatively clear (e.g. A_2 supports while A_5 attacks the main claim in A_1; A_3 attacks A_2 directly; and A_4 closes any further discussion on A_3's premise).

Clearly, the environment and type of platform under consideration have a significant impact on a system's capacity to implement such a framework and on the degree of complexity found in the components and relations to extract, assign, and predict. Working in the realm of overtly argumentative text (such as persuasive essays (Stab and Gurevych, 2017)), while challenging of course, can be quite standardized. The language use is generally in line with natural language expectations and often standard (e.g. claim, premise and stance are clear), the structure and collective goal of the debate are rather controlled and topic-specific, and the collection of participants involved is often a closed or an easily-classified set (e.g. in parliamentary debates, news forums, etc.).

2.2 Scaling Up Argument Mining

In social media While overtly argumentative text, like those described above, represents the natural domain of application for AM, social media constitute a powerful source of large amounts of data (billions of words) despite facing particular challenges in AM.

Social media plays an increasingly significant role in modern political and social discourse, yet resources built for conducting AM on this type of data structure remain limited for clear reasons. These platforms inherently collect and spread a wide range of content, including personal opinions, facts, fake news, and additional information of interest to users. Distinguishing between personal opinion, fact, and fake news, for example, is not always straightforward, as seen in recent work on fake news detection (Kotonya and Toni, 2020). Further, the language used on such platforms is infamously chaotic and often non-standard in comparison to the language use in more structured environments, like parliamentary debates. The combination of these aspects introduces the unique challenge of implementing AM to particularly heterogeneous, poorly annotated data.

Recent work has aimed to tackle such challenges in social media. Dusmanu et al. (2017) apply a supervised classification approach to identify arguments on Twitter, focusing on the tasks of facts recognition and source identification. They study the feasibility of the approaches proposed to address these tasks on a set of tweets related to the Grexit and Brexitnews topics. Habernal and

[4] `https://www.kialo.com/explore/featured`

Gurevych (2017) provide an extensive analysis of the steps and the modeling strategies necessary to analyze social media data (e.g. forum posts) in terms of their argumentative structure, while Simpson and Gurevych (2018) tackle the issue of the scalability of AM algorithms.

Despite the rising attention and developments to AM in social media, one of the major challenges currently facing the field is the lack of consensus on how exactly to analyse argumentative user-generated texts such as online comments (Bauwelinck and Lefever, 2020). On the one hand, the amount of annotations available for the scale of this heterogeneous data remains limited. Recent work by Schaefer and Stede (2020), among others, have aimed to construct large Twitter corpora annotated for argument components, including argumentative spans within tweets. On the other hand, annotation guidelines are not necessarily clear, and the theoretical motivations underlying the proposed guidelines used to generate labelled corpora rarely include motivation for the use of a particular theoretical basis. Bauwelinck and Lefever (2020) introduce a pilot study and aim to provide a clear justification of the theories and definitions underlying the design of a set of guidelines.

The linguistic, structural, and logistic complexity and "openness" of such platforms clearly present unique challenges. However, being able to work well with argumentative text from social media and discussion forums is essential considering the continuously growing impact on the political and social framework of modern times.

Multilingual argument mining Multilinguality is an important area of research in NLP that has gained more attention recently because of the cross-lingual transfer potentials of Pre-trained Language Models (Devlin et al., 2019; Conneau et al., 2020) and because of the potentials for a societal impact at a global scale. The latter is particularly important when considering AM for Social Good since language should not be a barrier for participation if the goal is to allow any productive contribution.

Various recent studies have investigated multilinguality for AM. Eger et al. (2019) discuss a series of experiments on using machine translation and annotation projection for AM, specifically argument components extraction and classification in German, English, and Chinese. A similar approach to build training data in other languages using machine translation is done in Toledo-Ronen et al. (2020), which use a pre-trained multilingual BERT (Devlin et al., 2019) for modeling. This approach is shown to perform well for classifying argument stance and detecting evidence, but not for predicting argument quality scores. Multilingual stance detection in political social media text (Vamvas and Sennrich, 2020) is also investigated in Lai et al. (2020) using stylistic, structural, affective and contextual features from text and analysing the scenarios in which each of these features is effective.

Other work has also dealt with building non-English datasets (Lindahl, 2020; Bauwelinck and Lefever, 2020; Schaefer and Stede, 2020; Zotova et al., 2020), but there still seems to be a focus on Indo-European languages (and sometimes Chinese) with a lack of datasets and analysis extending to other languages. This is a general issue in NLP research that extends to performance bias in favor of standard dialects for example in English (Blodgett et al., 2016) and bias that could target certain user groups instead of protecting them as was shown for Hate Speech Detection (Davidson et al., 2019). This is an important limitation to address in AM as well for more inclusivity and towards a more positive societal impact.

3 Argument Quality: An Integrated Definition

The second stage in the framework of AM is defined as relation assignment (c.f. Section 2.1); a complex task that aims to predict the relations holding between the arguments defined in the first stage. Being able to model the relations between arguments and components within the structure, for example in argument graphs (Besnard and Hunter, 2014; Craven and Toni, 2016), allows us to actually work with the argumentative text in an application-based setting, understand the stance and context of arguments, and develop a story for the consequential impact of arguments on the discourse, among other things. Generally speaking, we can use this task as an approach to analyze *argument quality* (AQ).

However, within the AM community, an open question concerns the adequate definition and operationalization of the notion of AQ. Despite this, to move forward with the task of AQ analysis and to create large corpora with crowd-sourced annotations, some approaches rely on the relative assessment of quality: Given two arguments, which is more convincing? (Habernal and Gurevych, 2016; Toledo et al., 2019; Gretz et al., 2020)

Thus the natural way of quantifying the success of an argument is in terms of its persuasiveness. Indeed, plenty of previous work has explored the many factors which contribute to the persuasiveness of a message: the linguistic features employed by the authors (Persing and Ng, 2017), the semantic type of claims and premises (Hidey et al., 2017), the different sources of evidence produced to support an argument (Addawood and Bashir, 2016), the effects of the personality traits and prior beliefs on persuasiveness (Lukin et al., 2017; Durmus and Cardie, 2018; Al Khatib et al., 2020), the interaction with other participants (Ji et al., 2018; Egawa et al., 2020), the use of argument invention when debating about unknown topics (Bilu et al., 2019), the structure of the arguments (Li et al., 2020), and the effect of the style of the text in achieving persuasion (El Baff et al., 2020).

Persuasiveness is, however, not the only way to define whether an argument is good – at least not from a deliberation point of view. A good contribution to a debate is one which uncovers a previously unnoticed aspect of a problem, thus generating a perturbation in the discourse (controversies can be productive!). Or else, a good contribution is one that settles an issue, by stating the differences between opposing views and allowing the discourse to stabilize in a series of clusters (convergence on just one position is not necessarily a good outcome).

Most recent research projects (Wachsmuth et al., 2017b) aim to address the challenge of redefining the notion of AQ, away from persuasiveness and towards a more "situated" definition which has to do with the needs of argumentation in a real-world scenario. This new definition has been the basis for the creation of new corpora from different domains (Ng et al., 2020), where feature-based (Wachsmuth and Werner, 2020) and neural models were tested for automatic prediction (Lauscher et al., 2020). Other aspects of AQ have become the subject of AM research such as the relevance and impact of arguments (Durmus et al., 2019), the verifiability (Park and Cardie, 2018), local acceptability (Yang et al., 2019) and the best "deliberative move" (Al-Khatib et al., 2018).

We argue that this shift is necessary for two reasons: (1) Working with real-world applications of AM naturally forces us into the more heterogeneous realm of data structures, such as social media, in which language, structure, and content are less uniform and confined to the classic notion of logical debate; and (2) In order to encourage deliberation from an open audience of citizens, we need to redefine our concept of AQ and productive discourse such that there is equal worth and participation granted to each contributor of the argument.

Deliberative Quality We therefore propose adapting the definition of quality to integrate the abundant research on the topic from the field of Social Sciences. Here, the quality of a discourse has been investigated in the context of deliberation with the focus on *inclusivity*: how can the interplay of the different participants in the discourse lead to an optimal outcome for the collective? The focus here is not on the quality of the individual contributions. Instead, an overall quality of the discourse is determined by the fact that the individual quality dimensions are distributed among different contributions (e.g some participants do more rational reasoning, others share personal experiences). We would like to integrate those aspects that focus on inclusivity and cooperation.

Similar to Wachsmuth et al. (2017b), social scientists have developed a taxonomy, the discourse quality index (DQI), that describes the different desirable aspects of a discourse (Steenbergen et al., 2003). This taxonomy has been used to analyze the quality of deliberation in different contexts, ranging from more formal contexts, such as parliamentary debates (Steiner et al., 2005), to informal discussions in online forums (Trénel, 2004). Both implementations integrate logical coherence as one dimension, *cogency* in Wachsmuth et al. (2017b), *justification* in the DQI. Some aspects of inclusivity are also being touched upon in the rhetorical and dialectical dimension of Wachsmuth et al. (2017b), such as using appropriate language (*Appropriateness*) or whether an argument supports conflict resolution (*global relevance*). We concentrate on the following dimensions from the DQI, which particularly focus on the collaborative aspect of discourse.

- *Respect*: this dimension includes respectful tone, respect for other social groups/backgrounds, and openness towards other opinions.
- *Equality / Participation*: it is not desirable that some dominant participants make the bulk of contributions while many others remain passive. All participants should have equal opportunities to contribute and all topics, including those that

DQI (Steenbergen et al., 2003)	AQ (Wachsmuth et al., 2017b)	Description
Logical coherence	Local acceptability	Argument should be sound, rationally worthy
Justification level	Local sufficiency	(Enough) premises should support the claim
—	Local relevance	Premises should be suitable to support claim
Personal experiences	Emotional appeal	Argumentation should increase empathy
Emotional balance	Appropriateness	Suitable language and amount of emotions
—	Credibility	Is the participant credible? (e.g. an expert)
Topic relevance	Clarity	Use of clear and correct language, contribution on topic
—	Arrangement	Proper arrangement of premises and claim
Respect	Global acceptability	Other participants value / support contributions
Constructiveness	Global relevance	Argument contributes to the resolution of the issue
—	Global sufficiency	Possible counterarguments are rebutted
Equality	—	Discourse should not be dominated by few participants
Interactivity	—	Contributions are linked to other contributions

Table 1: Comparing Argument Quality and Discourse Quality

may only affect minorities, are equally relevant.

- *Interactivity*: beyond simply sharing opinions, acknowledging other viewpoints and interacting with other participants through listening and responding lead to new perspectives arising – compromises can emerge.
- *Testimoniality / Report of personal accounts*: sharing stories and personal narratives as an alternative form of communication can involve more people in the discourse, especially those who cannot identify themselves with rational argumentation. It can also make other participants aware of other perspectives as it generally increases empathy. Especially when traditional or universal norms need to be questioned, narratives are particularly well suited, as their ambiguity and vagueness creates room for interpretation. This is particularly important when new ideas or perspectives are introduced, since they cannot yet be rationally articulated.

Table 1 establishes a direct comparison between discourse quality dimensions of the DQI (Steenbergen et al., 2003; Steiner et al., 2005) and argument quality dimensions as defined in Wachsmuth et al. (2017b). Apart from the potential theoretical insights, the existing guidelines can be applied to annotate new or enrich existing corpora for AM. Despite the small size, the data already annotated based on the DQI can be made usable and extended for NLP. In addition, some of the quality dimensions can be further quantified or approximated using statistical methods. For example, interactivity or equality can be assessed with frequency-based methods, such as frequency of posts by distinct participants and response rate.

Summing up The overview of the definitions of AQ along with the discussion of the potential of the integration of Deliberative Quality features into an AM framework has one strong take-home message: The need for the scope of the investigation to go beyond (a) the persuasiveness of a an argumentative text (speeches, forum posts, tweets), and (b) their relation to the immediate preceding discourse. Instead, we pointed out the need to also assess the potential of the impact of that argumentative text on the upcoming discourse: this dimension of quality, inherently related to the interpretation of argumentation as a cooperation challenge, is currently lacking in current approaches to AQ.

4 Grounding AQ in deliberation: moderation as a real-world application

Grounding AQ in a discourse perspective which quantifies "team-playing" and its impact on discourse dynamics is a clear challenge, both theoretically, in the Social Sciences and Argumentation Theory, and concretely, as the empirical quantification of discourse-grounded AQ will require large annotation efforts, real-time implementations, and thorough evaluation strategies. We propose to make a first step in tackling this challenge by mapping it into a concrete application: (semi-)automatic moderation implemented as a form of discourse *optimization*, or, as it is commonly referred to in the Social Sciences, *facilitation* (Kaner et al., 2007; Trénel, 2009).

To illustrate the dynamics of moderation, let us start from concrete examples from a deliberation platform, *RegulationRoom*. This discussion forum has been employed by public institutions to gather citizens contributions on discussions targeting very heterogeneous issues (more details can be found in Appendix). Let us consider the following example from a discussion on the distracted driving by commercial vehicle operators (e.g., truckers and bus drivers). The posts we selected (arrows indicate comment nesting) are from the discussion sub-thread: *Texting – what are the risks?*[5]

> ***User 1**: In 2004,... the driver failed to move out of the low-clearance lane while talking on a cellphone." This "accident" happened in 2004! He was TALKING on a CELL-PHONE!* ***IMO, "Turn Off Cell B/4 Driving!" should have become law long B/4 NOW!!*** *All these years have gone by, hundreds of LIVES have been lost, & our society is just NOW starting to work on this issue? AND we think we need to start with small steps like banning TEXTING (& sometimes in just commercial vehicles?)? [...]*
>
> → ***User 2**: A driver in California recently caused an accident because he spilled his coffee. Another driver almost wrecked because he was trying to light a cigarette. The bottom line is that ANY distraction while driving a car can cause an accident.* ***Where do we draw the line?*** *Also, there are millions of people out there who are completely capable of using their cell phone AND driving, at the same time.* ***Are we proposing that they should be punished, for the inabilities of others?*** *For people who spend much of their time in the car, this time might be their only chance to communicate with loved ones, do business, or make important calls.* ***If they are physically capable to use their phones safely while driving, why restrict their freedoms?***
>
> → → ***Moderator**: It's true that any distraction can cause an accident. The agency decided that texting was particularly unsafe, in part on the basis of the VTTI study that we reference lower on the page. Click the graphic to get a sense of the safety risks associated with different activities. A question: do you think that this rule imposes an undue burden on personal communication? What alternative restrictions on texting, if any, would you propose to impose on professional drivers?*

The example involves two users who clearly differ in their argumentation style and position. *User 1* has a clear position on the topic (claim in bold: not just texting, but all cellphone interactions should be banned), which she/he supports with personal reports (underlined text) an emotional tone, and a style which is typical of social media text. *User 2* replies, opening the post on a sarcastic note, which serves as the first premise to her/his (implicit) claim which is encoded in three rethorical questions (in bold): there should be no restrictions at all, because imposing them would be unfair. This is the case because (premises underlined): any distraction can cause an accident, some people are capable of using their phone while driving, people who spend lot of time in the car for professional reasons still need to communicate with loved ones. A moderator then joins the discussion to (a) provide a clarification as to why the focus is on texting and a link to further information on the matter, and (b) ask *User 2* to elaborate on the personal communication issue, and to propose alternatives. In the Appendix we report another example from the same topic and thread, where the user acts as a problematizer, challenging the scope and definition of the rule under discussion and the moderator acts as a "discourse traffic director", pointing out that the user should read and contribute to different threads in the discussion.

The guidelines for human moderators in RegulationRoom have been defined in advance in a 'moderator protocol' (eRulemaking Initiative et al., 2017) which reflect the moderator actions mentioned in the examples. In the protocol the moderator roles were divided into two main classes. Supervision functions include general moderator actions that do not necessarily target the specific content of the posts, e.g., greeting participants, monitoring compliance with netiquette (policing), or helping with technical difficulties. Substantive moderator functions aim to improve the quality of comments and promote fruitful discourse. As the examples above clearly show, this can both mean that the moderator encourages exchanges between discourse participants and participation in other posts (broadening the scope of the discussion), or helping users to improve the content of their posts (requests for clarification, focusing on one topic, substantive reasoning, sharing personal experiences).

RegulationRoom represents an excellent example of the beneficial role of the moderator in maintaining productive argumentation from participants. However, to the best of our knowledge, there is little to no NLP work targeting moderation modeling. Park et al. (2012) used data from RegulationRoom and conducted an annotation study to empirically categorize the types of moderator interventions specified in the moderator protocol. Classification experiments were conducted using SVM to predict the type of action a moderator would perform, given the previous comment. However this work is limited as it only focuses on two types of moderator interventions (broadening the scope of the discussion, improving argument quality) and as it does not predict whether the moderator should intervene, building on the assumption that a given comment has already been flagged as "in need for

[5] archive.regulationroom.org/texting/design-and-operation/index.html

moderation".

Besides the concrete example of RegulationRoom, moderation and discourse facilitation have been, and still are, a crucial topic in digital democracy.[6] The know-how of digital democracy experts is an invaluable starting point for the application of AM to moderation, as current research targets both the integration of digital solutions to facilitate online campaigns, and a critical reflection of the effects of such innovations on the deliberation outcomes.

Digital innovation supporting deliberation Argument maps (Walton, 2005) are widely employed to support online discussions, as an emerging optimization of the deliberation. Given a specific topic, for example possible reactions to climate change, users who wish to contribute to the discussion are requested to structure their contribution by producing an item in a conceptual map and optionally writing an accompanying post. Their contribution to the argument maps is often reviewed by a moderator. So in a sense, the argument map for a given deliberation process is the outcome of a process that comes both from below (the user) and above (the moderator).

Thanks to argument maps, the overall discourse picture can be overviewed and it is easier for the group of contributors to express support for one (or many) of the available options, without having to read a large number of long posts. An example of this approach is represented in *Deliberatorium*[7], an e-deliberation platform which has been extensively employed in many reference studies on the effect of digital innovation on deliberation (Klein, 2011). Another example of a digital deliberation platform which integrates argument maps and offers an option for moderation is COLAGREE (Yang et al., 2021; Ito, 2018). Among the studies testing the impact of such digital platforms on online deliberation, Spada et al. (2015) tests the effect of *Deliberatorium*'s argument maps on an online discussion among the supporters of the Italian Democratic party concerning the desired features of electoral law to be proposed by the party to the Parliament. This study compared the discussion of users employing Deliberatorium and a control group using a traditional forum format which was then encoded into argument maps. The comparison showed that the argument map modality did not discourage participation, and while it appeared to make users less creative (fewer new ideas as compared to the traditional forum), it also reduced the rate of claims without further discussion.

Yet, the need for trained moderators tends to be a significant bottleneck (both in terms of time and of costs) in digital deliberation. Moreover, empirical research on the effect of moderation on deliberation has uncovered the risks of biased moderation. For example, the experiment in Spada and Vreeland (2013) tests the extent to which moderators can influence participants' behavior by expressing their views during the moderation process.

4.1 NLP-Supported Moderation: desiderata and challenges

NLP-supported moderation represents a clear solution to the bottleneck problem affecting facilitation in digital democracy. Automatic tools can take over some of the tasks that human moderators typically perform when monitoring online discussions. For example, in Social Sciences, one of the most discussed issues in crowd-scale deliberation is "flaming", i.e., aggressive and disrespectful communicative behavior (Lampe et al., 2014). Here, moderators could benefit from hate-speech and trolling detection methods in NLP.

NLP methods to support deliberative decision-making have already been applied for the real-time visualisation of argument maps (El-Assady et al., 2017). Deliberation in real-time applications has the clear potential of structured arguments extraction from the news media (Daxenberger and Gurevych, 2020), the identification of the argumentative structure in deliberative contexts (Liebeck et al., 2016), as well as automatic argument summarization (Lawrence et al., 2017).

Beyond the real-time support to users (and moderators) provided by the methods described above, further tasks specific to AM which are part of the role of a human or (semi-)automoated moderator include: detecting fallacies (Habernal et al., 2018b), reasoning and common-sense (Habernal et al., 2018a), relevance estimation (Potthast et al., 2019). In addition, detecting and highlighting parts of an argument that are a good target for attacks (Jo et al., 2020a) can help the moderator to motivate more participation and argumentation from opposing sides of a discussion. Another important source is the detection of implicitly asserted prepositions

[6] See Dahlberg (2011) for an outline of positions in deliberative democracy.

[7] http://deliberatorium.mit.edu

(Jo et al., 2020b) which has a counterpart in the framing detection task (Card et al., 2015; Akyürek et al., 2020), as framing is a manipulation strategy which highlights specific aspects of an issue under discussion to promote certain interpretations.

Further NLP tasks which can play a crucial role in ensuring a healthy interaction are, for example, Hate Speech Detection (Warner and Hirschberg, 2012; Waseem and Hovy, 2016; Schmidt and Wiegand, 2017), Fact Checking (Vlachos and Riedel, 2014; Kotonya and Toni, 2020), Facts recognition and source identification (Dusmanu et al., 2017).

How to represent discourse? Thus far, we have discussed the main ingredients of a rich NLP-informed approach to deliberative discourse. These components, together with the deliberation-augmented definition of AQ sketched in section 3 are the features that the NLP moderator takes as an input. One question remains open: How to represent the argumentative discourse within a contribution (e.g. a forum post) and across contributions (e.g. an entire online deliberation campaign)? We can approach also this question from an interdisciplinary perspective. Reference work in political science aims at modeling the mechanisms of political discourse in forms of discourse networks, as defined in Leifeld (2017). A discourse network is a bipartite graph, containing two classes of nodes: actors (e.g. Angela Merkel; the left-wing party; etc.) and claims (e.g. housing opportunities should be established for refugees); Edges between actors and claims indicate the support or opposition of a certain actor to a specific claim. Discourse coalitions (Hajer, 1993) and argumentative clusters are the projection of the affiliation network on the actor and claim sides of the network (Leifeld and Haunss, 2012; Haunss et al., 2013). Recent NLP research has targeted integration machine learning in the discourse network analysis workflow (Padó et al., 2019; Haunss et al., 2020). Crucially for AM, discourse networks can integrate claims and actors with a third class of nodes, the frame nodes, which encode the reason put forward by an actor to support or reject a claim. This type of representation is perfectly compatible with a graph-based approach on argument representation which has already been established as to be preferred to a tree-structure representation both empirically (Niculae et al., 2017) and theoretically (Afantenos and Asher, 2014).

Moderation can thus be modeled as optimization of specific quantitative properties of the discourse network: participant inclusion, can be enforced by ensuring that the contributions of peripheric actor nodes receive the deserved salience; argument mapping and summarization can be modeled by identifying "hot" sub-graphs in the network; the impact of a contribution (the grounded notion of AQ we have been advocating thus far) can be quantified as the perturbation introduced in the network, with its long term effects on convergence or polarization.

Who moderates the (NLP) moderators? The problem of biased moderation obviously relates to the issue of bias in NLP (Blodgett et al., 2020; Caliskan et al., 2017; Bolukbasi et al., 2016; Spliethöver and Wachsmuth, 2020) and it has a clear implication in the application of NLP methods to moderation. For example, we would not want our NLP models to infer a negative impact on AQ from cues which just reveal that the user belongs to certain groups. This is a real risk when quality is equated to "success", in turn quantified in terms of likes, replies, retweets. The public of a forum may be sensitive to such cues, but the moderator should be unbiased with respect to them. Another source of bias is the degree of literacy of a contribution: while users who express themselves poorly are likely to be less popular with the forum public, their contributions may still be a very good move in the "cooperation challenge" – one that moderators (NLP or humans, online or in-person) have to ensure will not be left unexploited.

5 Conclusion

While there are clear social drawbacks to working with data and approaches to AM that limit the participation of the argumentation/deliberation, opening the floodgates to unregulated, evenly weighted contribution of all arguments also presents a dilemma. We present an interdisciplinary formulation of the notion of argument quality, which is more apt to work with heterogeneous data and platforms, such as discussion forums and social media. With the goal of ensuring a productive development of the discourse, we propose NLP-supported moderation to facilitate argumentation and deliberation in digital democracy.

Acknowledgments

We acknowledge funding by the Bundesministeriumfur Bildung und Forschung (BMBF) through the project E-DELIB (*Powering up E-deliberation: towards AI-supported moderation*).

References

Assel A. Addawood and Masooda N. Bashir. 2016. What is your evidence? A study of controversial topics on social media. In *Proceedings of the 3rd workshop on Argument Mining*, pages 1–11, Berlin, Germany.

Stergos Afantenos and Nicholas Asher. 2014. Counter-argumentation and discourse: A case study. *CEUR Workshop Proceedings*, 1341.

Afra Feyza Akyürek, Lei Guo, Randa Elanwar, Prakash Ishwar, Margrit Betke, and Derry Tanti Wijaya. 2020. Multi-label and multilingual news framing analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8614–8624, Online. Association for Computational Linguistics.

Khalid Al Khatib, Michael Völske, Shahbaz Syed, Nikolay Kolyada, and Benno Stein. 2020. Exploiting personal characteristics of debaters for predicting persuasiveness. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7067–7072, Online. Association for Computational Linguistics.

Khalid Al-Khatib, Henning Wachsmuth, Kevin Lang, Jakob Herpel, Matthias Hagen, and Benno Stein. 2018. Modeling deliberative argumentation strategies on Wikipedia. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2545–2555, Melbourne, Australia. Association for Computational Linguistics.

Andre Bächtiger and John Parkinson. 2019. *Towards a New Deliberative Quality*. Oxford University Press, Cambridge, MA, USA.

Nina Bauwelinck and Els Lefever. 2020. Annotating topics, stance, argumentativeness and claims in Dutch social media comments: A pilot study. In *Proceedings of the 7th Workshop on Argument Mining*, pages 8–18, Online. Association for Computational Linguistics.

Philippe Besnard and Anthony Hunter. 2014. Constructing argument graphs with deductive arguments: a tutorial. *Argument & Computation*, 5(1):5–30.

Yonatan Bilu, Ariel Gera, Daniel Hershcovich, Benjamin Sznajder, Dan Lahav, Guy Moshkowich, Anael Malet, Assaf Gavron, and Noam Slonim. 2019. Argument invention from first principles. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1013–1026.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Su Lin Blodgett, Lisa Green, and Brendan O'Connor. 2016. Demographic dialectal variation in social media: A case study of African-American English. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1119–1130, Austin, Texas. Association for Computational Linguistics.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 4356–4364, Red Hook, NY, USA. Curran Associates Inc.

Elena Cabrio and Serena Villata. 2013. A natural language bipolar argumentation approach to support users in online debate interactions. *Argument & Computation*, 4(3):209–230.

Elena Cabrio and Serena Villata. 2018. Five years of argument mining: a data-driven analysis. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 5427–5433. International Joint Conferences on Artificial Intelligence Organization.

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Dallas Card, Amber E. Boydstun, Justin H. Gross, Philip Resnik, and Noah A. Smith. 2015. The media frames corpus: Annotations of frames across issues. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 438–444, Beijing, China. Association for Computational Linguistics.

Tuhin Chakrabarty, Christopher Hidey, Smaranda Muresan, Kathy McKeown, and Alyssa Hwang. 2019. AMPERSAND: Argument mining for PERSuAsive oNline discussions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2933–2943, Hong Kong, China. Association for Computational Linguistics.

Oana Cocarascu, Elena Cabrio, Serena Villata, and Francesca Toni. 2020. Dataset independent baselines for relation prediction in argument mining. In *Computational Models of Argument - Proceedings of COMMA 2020, Perugia, Italy, September 4-11, 2020*, volume 326 of *Frontiers in Artificial Intelligence and Applications*, pages 45–52. IOS Press.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Robert Craven and Francesca Toni. 2016. Argument graphs and assumption-based argumentation. *Artificial Intelligence*, 233:1–59.

Lincoln Dahlberg. 2011. Re-constructing digital democracy: An outline of four 'positions'. *New media & society*, 13(6):855–872.

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial bias in hate speech and abusive language detection datasets. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 25–35, Florence, Italy. Association for Computational Linguistics.

Johannes Daxenberger and Irina Gurevych. 2020. Arguments as social good: Good arguments in times of crisis. In *Proocedings of the AAAI Fall 2020 Symposium on AI for Social Good*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186, Minneapolis, Minnesota.

Esin Durmus and Claire Cardie. 2018. Exploring the role of prior beliefs for argument persuasion. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1035–1045, New Orleans, Louisiana. Association for Computational Linguistics.

Esin Durmus, Faisal Ladhak, and Claire Cardie. 2019. The role of pragmatic and discourse context in determining argument impact. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5668–5678, Hong Kong, China. Association for Computational Linguistics.

Mihai Dusmanu, Elena Cabrio, and Serena Villata. 2017. Argument mining on twitter: Arguments, facts and sources. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2317–2322.

Ryo Egawa, Gaku Morio, and Katsuhide Fujita. 2020. Corpus for modeling user interactions in online persuasive discussions. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1135–1141, Marseille, France. European Language Resources Association.

Steffen Eger, Johannes Daxenberger, and Iryna Gurevych. 2017. Neural end-to-end learning for computational argumentation mining. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11–22, Vancouver, Canada. Association for Computational Linguistics.

Steffen Eger, Johannes Daxenberger, Christian Stab, and Iryna Gurevych. 2019. Cross-lingual argumentation mining: Machine translation (and a bit of projection) is all you need! In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 831–844.

Liat Ein-Dor, Eyal Shnarch, Lena Dankin, Alon Halfon, Benjamin Sznajder, Ariel Gera, Carlos Alzate, Martin Gleize, Leshem Choshen, Yufang Hou, et al. 2020. Corpus wide argument mining—a working solution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7683–7691.

Mennatallah El-Assady, Annette Hautli-Janisz, Valentin Gold, Miriam Butt, Katharina Holzinger, and Daniel Keim. 2017. Interactive visual analysis of transcribed multi-party discourse. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics-System Demonstrations*, pages 49–54, Vancouver, Canada.

Roxanne El Baff, Henning Wachsmuth, Khalid Al Khatib, and Benno Stein. 2020. Analyzing the Persuasive Effect of Style in News Editorial Argumentation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3154–3160, Online. Association for Computational Linguistics.

Shai Gretz, Roni Friedman, Edo Cohen-Karlik, Assaf Toledo, Dan Lahav, Ranit Aharonov, and Noam Slonim. 2020. A large-scale dataset for argument quality ranking: Construction and analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7805–7813.

Ivan Habernal and Iryna Gurevych. 2016. Which argument is more convincing? analyzing and predicting convincingness of web arguments using bidirectional LSTM. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1589–1599, Berlin, Germany. Association for Computational Linguistics.

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

Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. 2018a. The argument reasoning comprehension task: Identification and reconstruction of implicit warrants. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1930–1940, New Orleans, Louisiana. Association for Computational Linguistics.

Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. 2018b. Before name-calling: Dynamics and triggers of ad hominem fallacies in web argumentation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 386–396, New Orleans, Louisiana. Association for Computational Linguistics.

Maarten A Hajer. 1993. Discourse Coalitions and the Institutionalization of Practice: The Case of Acid Rain in Britain. In *The Argumentative Turn in Policy Analysis and Planning*, pages 43–76. Duke University Press.

Sebastian Haunss, Matthias Dietz, and Frank Nullmeier. 2013. Der Ausstieg aus der Atomenergie. Diskursnetzwerkanalyse als Beitrag zur Erklärung einer radikalen Politikwende. *Zeitschrift für Diskursforschung*, 1(3):288–316.

Sebastian Haunss, Jonas Kuhn, Sebastian Pado, Andre Blessing, Nico Blokker, Erenay Dayanik, and Gabriella Lapesa. 2020. Integrating manual and automatic annotation for the creation of discourse network data sets. *Politics and Governance*, 8(2).

Christopher Hidey, Elena Musi, Alyssa Hwang, Smaranda Muresan, and Kathleen McKeown. 2017. Analyzing the semantic types of claims and premises in an online persuasive forum. In *Proceedings of the 4th workshop on Argument Mining*, pages 11–21.

Cornell eRulemaking Initiative et al. 2017. Ceri (cornell e-rulemaking) moderator protocol.

Takayuki Ito. 2018. Towards agent-based large-scale decision support system: The effect of facilitator. In *Proceedings of the 51st Hawaii International Conference on System Sciences*, pages 351–360.

Lu Ji, Zhongyu Wei, Xiangkun Hu, Yang Liu, Qi Zhang, and Xuanjing Huang. 2018. Incorporating argument-level interactions for persuasion comments evaluation using co-attention model. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3703–3714, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Yohan Jo, Seojin Bang, Emaad Manzoor, Eduard Hovy, and Chris Reed. 2020a. Detecting attackable sentences in arguments. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1–23, Online. Association for Computational Linguistics.

Yohan Jo, Jacky Visser, Chris Reed, and Eduard Hovy. 2020b. Extracting implicitly asserted propositions in argumentation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 24–38, Online. Association for Computational Linguistics.

Sam Kaner, Lenny Lind, Catherine Tolid, Sarah Fisk, and Duane Berger. 2007. *Facilitator's guide to participatory decision-making*. John Wiley & Sons/Jossey-Bass, San Francisco.

Mark Klein. 2011. The MIT Deliberatorium: Enabling large-scale deliberation about complex systemic problems. In *2011 International Conference on Collaboration Technologies and Systems (CTS)*.

Jonathan Kobbe, Juri Opitz, Maria Becker, Ioana Hulpus, Heiner Stuckenschmidt, and Anette Frank. 2019. Exploiting background knowledge for argumentative relation classification. In *2nd Conference on Language, Data and Knowledge, LDK 2019, May 20-23, 2019, Leipzig, Germany*, volume 70 of *OASICS*, pages 8:1–8:14. Schloss Dagstuhl - Leibniz-Zentrum für Informatik.

Neema Kotonya and Francesca Toni. 2020. Explainable automated fact-checking: A survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5430–5443.

Mirko Lai, Alessandra Teresa Cignarella, Delia Irazú Hernández Farías, Cristina Bosco, Viviana Patti, and Paolo Rosso. 2020. Multilingual stance detection in social media political debates. *Computer Speech & Language*, 63:101075.

Cliff Lampe, Paul Zube, Jusil Lee, Chul Hyun Park, and Erik Johnston. 2014. Crowdsourcing civility: A natural experiment examining the effects of distributed moderation in online forums. *Government Information Quarterly*, 31(2):317 – 326.

Anne Lauscher, Lily Ng, Courtney Napoles, and Joel Tetreault. 2020. Rhetoric, logic, and dialectic: Advancing theory-based argument quality assessment in natural language processing. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4563–4574, Barcelona, Spain (Online). International Committee on Computational Linguistics.

John Lawrence, Joonsuk Park, Katarzyna Budzynska, Claire Cardie, Barbara Konat, and Chris Reed. 2017. Using argumentative structure to interpret debates in online deliberative democracy and erulemaking. *ACM Trans. Internet Technol.*, 17(3 - 25).

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

Philip Leifeld. 2017. Discourse network analysis. policy debates as dynamic networks. In Jennifer N. Victor, Mark N. Lubell, and Alexander H. Montgomery, editors, *The Oxford Handbook of Political Networks*, chapter 12, pages 301–325. Oxford University Press, Oxford.

Philip Leifeld and Sebastian Haunss. 2012. Political Discourse Networks and the Conflict over Software Patents in Europe. *European Journal of Political Research*, 51(3):382–409.

Jialu Li, Esin Durmus, and Claire Cardie. 2020. Exploring the role of argument structure in online debate persuasion. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8905–8912, Online. Association for Computational Linguistics.

Matthias Liebeck, Katharina Esau, and Stefan Conrad. 2016. What to do with an airport? mining arguments in the german online participation project tempelhofer feld. In *Proceedings of the 3rd workshop on Argument Mining*, pages 144–153, Berlin, Germany.

Anna Lindahl. 2020. Annotating argumentation in swedish social media. In *Proceedings of the 7th Workshop on Argument Mining*, pages 100–105, Barcelona, Spain (Online). Association for Computational Linguistics.

Luca Lugini and Diane Litman. 2020. Contextual argument component classification for class discussions. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1475–1480, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Stephanie Lukin, Pranav Arand, Marilyn Walker, and Steve Whittaker. 2017. Argument strength is in the eye of the beholder: Audience effects in persuasion. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 742–753, Valencia, Spain.

Raquel Mochales and Marie-Francine Moens. 2011. Argumentation mining. *Artificial Intelligence and Law*, 19(1):1–22.

Marie-Francine Moens, Erik Boiy, Raquel Mochales Palau, and Chris Reed. 2007. Automatic detection of arguments in legal texts. In *Proceedings of the 11th international conference on Artificial intelligence and law*, pages 225–230.

Lily Ng, Anne Lauscher, Joel Tetreault, and Courtney Napoles. 2020. Creating a domain-diverse corpus for theory-based argument quality assessment. In *Proceedings of the 7th Workshop on Argument Mining*, pages 117–126, Online. Association for Computational Linguistics.

Vlad Niculae, Joonsuk Park, and Claire Cardie. 2017. Argument mining with structured SVMs and RNNs. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 985–995, Vancouver, Canada. Association for Computational Linguistics.

Sebastian Padó, André Blessing, Nico Blokker, Erenay Dayanik, Sebastian Haunss, and Jonas Kuhn. 2019. Who sides with whom? towards computational construction of discourse networks for political debates. In *Proceedings of ACL*, Florence, Italy.

Joonsuk Park and Claire Cardie. 2018. A corpus of eRulemaking user comments for measuring evaluability of arguments. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Joonsuk Park, Sally Klingel, Claire Cardie, Mary J. Newhart, Cynthia Farina, and J. Vallbé. 2012. Facilitative moderation for online participation in erulemaking. In *dg.o '12*.

Andreas Peldszus and Manfred Stede. 2015. Joint prediction in MST-style discourse parsing for argumentation mining. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 938–948, Lisbon, Portugal. Association for Computational Linguistics.

Isaac Persing and Vincent Ng. 2017. Why can't you convince me? Modeling weaknesses in unpersuasive arguments. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17)*, pages 4082–4088.

Martin Potthast, Lukas Gienapp, Florian Euchner, Nick Heilenkötter, Nico Weidmann, Henning Wachsmuth, Benno Stein, and Matthias Hagen. 2019. Argument search: Assessing argument relevance. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'19, page 1117–1120, New York, NY, USA. Association for Computing Machinery.

Robin Schaefer and Manfred Stede. 2020. Annotation and detection of arguments in tweets. In *Proceedings of the 7th Workshop on Argument Mining*, pages 53–58, Online. Association for Computational Linguistics.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain. Association for Computational Linguistics.

Edwin Simpson and Iryna Gurevych. 2018. Finding convincing arguments using scalable bayesian preference learning. *Transactions of the Association for Computational Linguistics*, 6.

Paolo Spada, Mark Klein, Raffaele Calabretta, Luca Iandoli, and Ivana Quinto. 2015. A first step toward scaling-up deliberation: Optimizing large group e-deliberation using argument maps. In *American Political Science Association (APSA), 110th Annual Meeting. Politics after the Digital Revolution*, Washington DC.

Paolo Spada and James Raymond Vreeland. 2013. Who moderates the moderators? The effect of non-neutral moderators in deliberative decision making. *Journal of Public Deliberation*, 9(2,3).

Maximilian Spliethöver and Henning Wachsmuth. 2020. Argument from old man's view: Assessing social bias in argumentation. In *Proceedings of the 7th Workshop on Argument Mining*, pages 76–87, Online. Association for Computational Linguistics.

Christian Stab and Iryna Gurevych. 2017. Parsing argumentation structures in persuasive essays. *Computational Linguistics*, 43(3):619–659.

M. Steenbergen, Andre Baechtiger, Markus Spörndli, and J. Steiner. 2003. Measuring political deliberation: A discourse quality index. *Comparative European Politics*, 1:21–48.

Jürg Steiner, André Bächtiger, Markus Spörndli, and Marco R. Steenbergen. 2005. *Deliberative Politics in Action: Analyzing Parliamentary Discourse*. Theories of Institutional Design. Cambridge University Press.

Assaf Toledo, Shai Gretz, Edo Cohen-Karlik, Roni Friedman, Elad Venezian, Dan Lahav, Michal Jacovi, Ranit Aharonov, and Noam Slonim. 2019. Automatic argument quality assessment - new datasets and methods. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5625–5635, Hong Kong, China. Association for Computational Linguistics.

Orith Toledo-Ronen, Matan Orbach, Yonatan Bilu, Artem Spector, and Noam Slonim. 2020. Multilingual argument mining: Datasets and analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 303–317, Online. Association for Computational Linguistics.

Dietrich Trautmann, Michael Fromm, Volker Tresp, Thomas Seidl, and Hinrich Schütze. 2020. Relational and fine-grained argument mining: The lmu munich project remlav within the dfg priority program ratio "robust argumentation machines". *Datenbank-Spektrum*, 20.

Mathias Trénel. 2004. Measuring the deliberativeness of online discussions. coding scheme 2.4. *report, Berlin: Social Science Research Centrex*.

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

Jannis Vamvas and Rico Sennrich. 2020. X-stance: A multilingual multi-target dataset for stance detection. In *5th Swiss Text Analytics Conference (SwissText) & 16th Conference on Natural Language Processing (KONVENS)*. CEUR Workshop Proceedings.

Andreas Vlachos and Sebastian Riedel. 2014. Fact checking: Task definition and dataset construction. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 18–22, Baltimore, MD, USA. Association for Computational Linguistics.

Henning Wachsmuth, Nona Naderi, Ivan Habernal, Yufang Hou, Graeme Hirst, Iryna Gurevych, and Benno Stein. 2017a. Argumentation quality assessment: Theory vs. practice. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 250–255, Vancouver, Canada. Association for Computational Linguistics.

Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein. 2017b. Computational argumentation quality assessment in natural language. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 176–187, Valencia, Spain. Association for Computational Linguistics.

Henning Wachsmuth and Till Werner. 2020. Intrinsic quality assessment of arguments. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6739–6745, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Douglas Walton. 2005. *Argumentation Methods for Artificial Intelligence in Law*. Springer Science & Business Media.

William Warner and Julia Hirschberg. 2012. Detecting hate speech on the world wide web. In *Proceedings of the Second Workshop on Language in Social Media*, pages 19–26, Montréal, Canada. Association for Computational Linguistics.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California. Association for Computational Linguistics.

Chunsheng Yang, Wen Gu, Takayuki Ito, and Xiaohua Yang. 2021. Machine learning-based consensus decision-making support for crowd-scale deliberation. *Applied Intelligence*.

Wonsuk Yang, Seungwon Yoon, Ada Carpenter, and Jong Park. 2019. Nonsense!: Quality control via two-step reason selection for annotating local acceptability and related attributes in news editorials. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2954–2963, Hong Kong, China. Association for Computational Linguistics.

Elena Zotova, Rodrigo Agerri, Manuel Nuñez, and German Rigau. 2020. Multilingual stance detection in tweets: The Catalonia independence corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1368–1375, Marseille, France. European Language Resources Association.

A Appendix

E-rulemaking & RegulationRoom

E-rulemaking is a type of (e-)deliberation process which originated in the United States. Its goal is to use digital innovations to increase participation and transparency in the decision-making process of the Federal Government. More concretely, given a new regulation to be written (or the need to significantly update an existing one), a government agency directly involves the citizens in the discussion of specific aspects of that rule, sharing relevant data with the citizens and committing to incorporate the output of their deliberation in the final rule. A crucial role is obviously paid by the E-rulemaking ”provider”, who sets up the infrastructure both practically (e.g., creating websites and portals for citizens to participate) and qualitatively (by monitoring the discussion and creating summaries to be submitted to the agency).

RegulationRoom is a deliberation platform designed by the Cornell eRulemaking Initiative (CeRI) to support various large scale e-deliberation, hosted by the Legal Information Institute (LII) at the Cornell Law School, has been employed by public institutions to gather citizens contributions on rules targeting very heterogeneous issues, such as airline passengers rights, home mortgage consumer protection, distracted driving by commercial motor vehicles, among others.

The example provided in the paper and the additional example in this appendix are an excerpt from the distracted driving discussion, which is publicly available at http://archive.regulationroom.org/texting/index.html.

Before proceeding to the additional example, we elaborate on the deliberation context from which the examples are extracted.

The Federal Motor Carrier Safety Administration had been planning new federal regulations to address distracted driving by truckers, and the examples show a discussion about a specific subtopic: What are the risks of texting while driving? Examples of other subtopics for the same discussion are: What counts as texting? Which drivers are covered? What penalties should caught drivers receive? How will any law enforcement entity know when a driver is texting?

The discussion took place in April 2010. Original posts are time-stamped and organized in discussion threads; we anonymized the user names.

Additional moderation example

User 3*: I don’t dispute the distraction factor. 10 Minutes on any highway in the country should offer enough proof for all but the most obtuse.* ***What I object to is the singling out of any particular group of drivers as the focus of another un-enforceable law*** *(or, shall we say, really only enforceable after the fact).*

Truckers already face a huge pile of regulations that apply only to them, and not to other drivers on the road. In most cases, these regulations are at least tangentally appropriate given the nature of the vehicle driven. In this case, however, the activity in question is one engaged in by drivers off all classes of vehicle. ***It seems to me to be more appropriate for the regulation or non-regulation to come at the state level, and cover ALL vehicle operators.***

→ ***Moderator****: Thanks for your thoughtful comments. For more information about why FMCSA has proposed to imposes regulations against commercial drivers, please see our next post called “Which Drivers are Covered.” After reading through this material, let the community know if your opinion has changed.*

As to your comment about enforcement, you’ve identified one of the most difficult questions about this proposed regulation. Feel free to continue to discuss this question in the post called “Who & How of Enforcement.”