Comparing the Effectiveness of Different User Moderation Strategies for Deliberation on Twitter

Julian Dehne[a]* and Valentin Gold[b]

[a]Gesellschaft für Informatik, Berlin (julian.dehne@gi.de);

[b]Universität Göttingen

Comparing the Effectiveness of Different User Moderation Strategies for Deliberation on Twitter

This paper analyzes deliberative user moderation in social media. Using phrases to sample moderating statements from ordinary users a corpus of user moderation on Twitter was created. Using this corpus, the intended effect of user moderation on subsequent online discussions is analyzed on a large scale using computational methods. A small but systematic effect can be observed which was interpreted based on the intended function of the user's intervention and general concepts of deliberative quality.

Keywords: deliberation; moderation; social media; Twitter

Introduction

In the contemporary digital landscape, social media platforms are increasingly at risk of losing their capacity for productive deliberation, owing to the prevalence of automated trolls, the strategic dissemination of misinformation, and the pervasive influence of marketing interests, which often overshadow empirical scientific evidence and ethically grounded values. It is therefore crucial to identify and implement proactive strategies that bolster the quality of deliberation on social media platforms.

Extant research has explored various approaches to enhance the deliberative quality within these digital forums. These methods range from assigning responsibility to the social media platforms themselves or to the judicial legal system to approaches that penalize the users directly on the platform. Although each of these methods contributes to the enhancement of deliberative quality, it is essential to recognize that ultimately, the discourse is driven not by the platforms but by their users.

The dynamic of user engagement in upholding deliberative standards shows a dichotomy. On one hand, there is evidence indicating that users often assume responsibility proactively in moderating content. On the other hand, the efficacy of such user-driven interventions in improving deliberative quality appears contingent on specific conditions. One key variable is the scale of user engagement. While existing literature, such as the work of Friess et al. (2021), points to the positive impact of collective civic interventions — where groups of users are organized collectively — less is known about the influence of ordinary users moderating discussions on social media. The extent to which such singular interventions by users enhance the quality of deliberation remains an open question, warranting further empirical investigation.

In this paper, we conduct a large-scale analysis of Twitter/X[1] tweets with the question of how users react to attempts of moderation by other users and which moderation strategies work best for that purpose. First, based on previous empirical and theoretical studies, we derive a model for deliberative moderation. Then a phrase-based sampling approach is employed to identify instances of user moderation online and classify them into different strategies of deliberative moderation. Furthermore, the reaction patterns of different author compositions are analyzed as well as the impact the different strategies have in terms of their intended effect and on deliberative quality.

The assumption being investigated in this paper is that user moderation – although not primarily intended as a deliberative force – leads to patterns of reactions that can be categorized as deliberative or not depending on the intended function of the intervention and whether the outcome corresponds with that function.

Moderation and Deliberation

Moderation constitutes a specific type of deliberative communication: building on the previous turns, moderation adds another layer to the discussion. It shifts the focus away

[1] In the subsequent text, we use Twitter and tweets (and not X and posts). This is due to the fact that the data was gathered in 2023 before Twitter Inc. renamed their social media site to X. Also, before renaming, *posts* were named *tweets*.

from the actual content to the metalevel of communication. The link to the actual content might vary widely: While some moderators refer to the content of the previous turns directly, e.g., by integrating and combining lines of argumentation, others only focus on the metalevel, for instance by asking participants to be more friendly in their communication. These brief examples also demonstrate that the degree (and success) of metacommunication is highly dependent on the type of moderation – a perspective that we will elaborate further in a later section of this paper.

Researchers conclude that metacommunication (e.g., talking about the tone of discussion) engenders more metacommunication and is usually caused by an uncivil discourse (Han et al., 2018, p. 1). However, metacommunication does not significantly decrease incivility, but leads to more metacommunication, according to Molina and Jennings (2018, p. 56).

Moderating behavior is mainly researched as part of a bundle of reactions to uncivil behavior rather than as a proactive force. It is still uncertain whether moderation shares the same preconditions and effects as metacommunication in general, or if individual characteristics can be observed.

When it comes to previous research, targeted moderating interventions with metacommunicative aspects have been studied in several ways. These can be summarized by three questions: (a) Why do users take over an intervening role? (b) Which moderation approaches are they using? (c) How do participants react to moderation attempts? In the following sections, we briefly describe some of the studies that illustrate different directions of research that are relevant to our study. Our focus is mainly on the effects of moderation (question c), nevertheless, to answer this question, the other two preceding questions are also partly relevant. Even though Friess et al.

(2021) analyze collective civic moderation (and not user moderation), they provide a rather coherent overview of the fields of study.

(a) When do users moderate?

Porten-Cheé et al. (2020) highlight that ordinary users have the authority to identify and label norm violations within the discourse. They assert that violations of political context norms, relation norms, information norms, process norms, and modality norms are perceived as particularly significant. For example, an information norm such as “don’t tell lies” is ranked higher than “be polite”. This hierarchy of norms is relevant when considering when ordinary users intervene. Previous research is mainly concerned with the violation of higher-level norms like incivility, hate and bigotry.

Also, Molina and Jennings (2018, pp. 57–58) state that when users “were made aware of the uncivil tone in a conversation, they were more likely to try to intervene in the situation and point out the incivility they witnessed”. Even though they do not find any evidence for incivility, Molina and Jennings (2018, p. 58) conclude that it “is worth investigating further”. We do so by extending the range of norm violations beyond incivility and also include other types in our analysis.

Awareness of norm violation is only one part. The other part is whether the digression can be traced back to some relevant experience for the users. The General Aggression Model (GAM), proposed by Allen et al. (2018), suggests that individuals develop specific knowledge structures related to aggression based on their experiences. Three specific knowledge structures particularly influence the interpretation of aggression in social situations: perceptual schemata (such as recognizing an uncivil comment), person schemata (beliefs about the comment sender), and behavioral scripts (information on how individuals should respond to a discussion comment). Drawing on this, Kluck and Krämer (2022, p. 2) conclude that, when individuals perceive comments from other users as uncivil, their response is influenced by behavioral patterns they have learned from previous experiences. From this follows that users are more likely to classify a comment as uncivil when they feel emotionally attacked and can relate to the target of the attack.

(b) Which strategies do users use?

Moderators in Twitch communities, as identified by Wohn (2019), can assume different roles. These roles include Helping hands, Justice Enforcers, Surveillance Units, and Conversationalists. Cai and Wohn (2019) identify five approaches that moderators may take to dealing with problematic behaviors: Educating, Sympathizing, Shaming, Humor, and Blocking. Kraut et al. (2012) outline five key difficulties that community leaders encounter: (1) Encouraging Contribution; (2) Encouraging Commitment; (3) Regulating Behavior; (4) Dealing with Newcomers; and (5) Starting New Communities. However, Seering (2020, p. 18) concludes that the effectiveness of moderation cannot be defined universally.

(c) What is the effect of user moderation?

Molina and Jennings (2018) hypothesize that comments condemning the incivility of other commenters would result in more civil comments and a decrease in uncivil comments. However, their findings did not yield significant results to support these hypotheses. Although they did not observe a significant impact of metacommunication on the overall civility level, which would have indicated the participants' ability to independently change the tone of a conversation without relying on moderators, they found evidence of metacommunication being practiced. This suggests that commenters positively responded to individuals attempting to intervene and denounce incivility.

Going beyond the previous approaches, Friess et al. (2021) analyze the effect of civic moderation concerning the variables *rationality, constructiveness, respect,* and *reciprocity*. Although all these are important for deliberation, they are only a subset targeting incivility, and thus missing other important process-oriented aspects like *understanding, sincerity, equality, and freedom* (Graham & Witschge, 2003), or justification (Monnoyer-Smith & Wojcik, 2012). It should also be noted that Friess et al. are looking at a specialized user group selected by their organizational goal of improving social media instead of ordinary users focused on in this study. In general, Friess et al. show that targeted counter-speech has a moderating effect on uncivil speech.

Types of Deliberative User Moderation

To measure the effect of user moderation, we first need to derive a framework for deliberative moderation. So far, moderation has only been used to explain the effects on deliberative quality, but we're not aware of any theoretical framework situating moderation within the theory of deliberative communication.

Theoretical Framework

Drawing on political philosophy and psychological research on moderation and mediation, a framework is developed that situates moderation between the deliberative function and its practice in social media. Moderation is then operationalized as a set of strategies and expressions of strategies so that it can be found and labeled in social media. Finally, a discourse-analytical framework is imported from qualitative social sciences to categorize the effect moderation has in terms of discourse component patterns. These can be interpreted as social scripts or simply as patterns of reaction to moderation.

Moderation and deliberation are two concepts that can be closely linked. In the context of this study, social media is viewed from the perspective of how online communication may foster the exchange of views and arguments, and improve or worsen the social bonds that hold together democratic societies.

Moderation, as an essential component of deliberation, plays a vital role in fostering constructive and inclusive discussions. Deliberation, at its core, is the process of exchanging ideas, perspectives, and opinions to reach informed decisions. According to Stromer-Galley (2007, p. 3), deliberation is defined [. . .] as a process *whereby groups of people, often ordinary citizens, engage in reasoned opinion on a social or political issue to identify solutions to a common problem and evaluate those solutions.*

Moderation is understood beyond just deletion or flagging (Friess et al., 2021); it encompasses strategies used by participants to guide conversations constructively. This approach considers real-life deliberation, where speech acts cannot be erased, emphasizing comments that aim to rationalize discussions rather than censor them. In general, moderation is defined as a governance mechanism that structures community participation, fostering cooperation and preventing abuse, as per Grimmelmann (2015, p. 1). Edwards (2002) breaks down its function into three aspects: strategic, conditioning, and process. The strategic function sets the discussion's boundaries within the community's context, the conditioning function establishes rules and guidelines, and the process function involves managing the discussion by facilitating dialogue and mediating conflicts. Friess et al. (2021) distinguish between different types of moderators - professional, user, and ordinary users who occasionally moderate. Ordinary users' moderation blends with their general communication and personal agenda (Mutz, 2008), differing from user or professional moderators who enforce pre-set norms.

Five moderation functions for user moderation in social media were distilled from the literature. Figure *1* summarizes the moderation strategy framework.

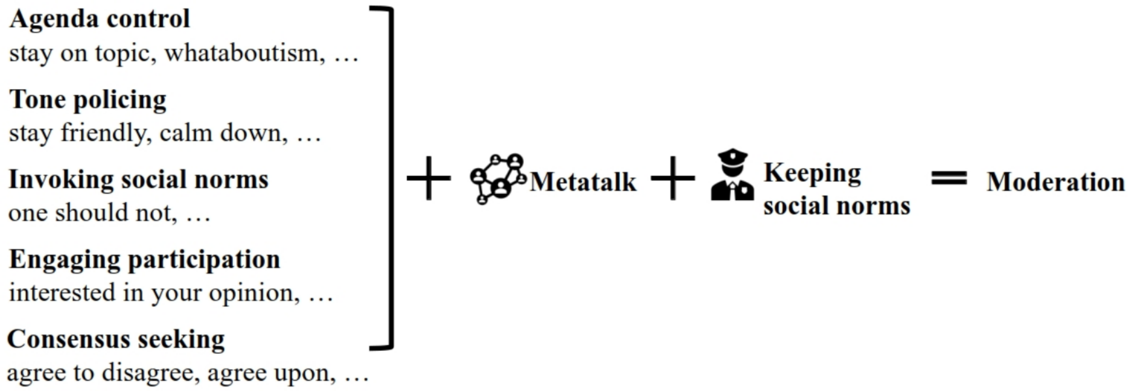

Figure 1: Deliberative Moderation Framework

The following gives a short intuition on why these moderation strategies in Figure 1 might be relevant to foster deliberative quality and also illuminates their basis in the literature. An in-depth discussion is outside the scope of this paper. On the one hand "Tone Policing" and "Invoking Social Norms" are popular aspects in current research dealing with the effect on the deliberative quality of various independent variables such as incivility, rationality, affective polarization, and emotions. On the other hand, "Agenda Control", "Engaging Participation", and "Consensus-Seeking support" derive directly from the ideals of deliberative democracy. These will be grounded in Habermasian theory directly.

Tone Policing aims to maintain a respectful and constructive atmosphere, encouraging understanding and compromise among participants. Keeping a cool head is commonly included in the literature to deliberation, with it acting both as a proxy of rationality and civility (Papacharissi, 2004). Although other lines of investigation exist that include anger (Kim, 2016) or moral indignation (Hwang et al., 2018), the political science literature has mainly focused on rationality. With the approach of starting with the intent (or function) of the moderation it is only logical to stick with the assumption of issue focus over emotions or relationships.

Invoking Social Norms involves reminding participants of community guidelines to foster a harmonious environment (Cialdini & Goldstein, 2004).

Agenda Control focuses on organizing the spread of topics within the discussion to improve consistency and understanding (Graham & Witschge, 2003).

Agenda Control, Engaging Participation, and **Consensus-seeking support**: The strategies “Engaging Participation” and “Consensus-seeking support” were derived from the Habermasian ideal of deliberation: *A conversation is a discourse if and only if a rationally motivated consensus could be achieved assuming that the arguments can be put forward as often as necessary and for as long as necessary (adapted and translated by the authors)* (Habermas, 1981, p. 71).

Consensus-seeking strategies involve summarizing key points and promoting nuanced views to reach a shared understanding and a *motivated consensus*. In contrast to the concept of the public sphere, this communicative idea fits perfectly with the conversations held on Twitter as they do not impose any restrictions as to how often arguments are presented. Contrary to live debates they are not dependent on the physical conditions of humans becoming fatigued and can stretch time indefinitely assuming the platform stays online and the attention does not waver. The first part of the definition introduces consensus as a criterion. The following citation links the consensus building to the single participants, displaying Habermas’ efforts to integrate systems theory with symbolic interactionism and Durkheim’s theory of solidarity: *A discoursive consensus depends at the same time on the yes and the no of each single participants, and it depends on surpassing egocentrism.* (translated by the authors) (Habermas, 2009, p. 19). Engaging participation broadens the discussion by incorporating diverse perspectives, especially important in social media where attention mechanisms might limit exposure.

In summary, a process-oriented Habermasian concept of deliberation is used in this paper, which assumes that some sort of rational discussion is possible and that the goal of deliberative moderation is to move the discussion closer to the ideal of a calm, focused, argument-based, and consensus-oriented discussion. The framework covers low-key interventions like invoking basic norms of politeness up to higher-level functions that are more likely associated with an academic panel discussion. The advantage of this approach is that it does not require a definition of deliberative quality which has been strongly criticized (Steenbergen et al., 2003; Bächtiger & Parkinson, 2019, p. 50).

Empirical Framework

Building upon our theoretical framework, moderation is operationalized by a set of phrases. Arriving at a set of phrases is, however, rather challenging. In principle, we could translate the task to a machine learning exercise: go through a randomly selected number of tweets and label these accordingly – whether they contain a moderation statement or not. Then, we could extract those phrases that define moderation. However, due to vast amounts of available social media data, finding moderation statements is extremely rare and this inductive approach is prone to fail.

A second approach builds upon a priori knowledge about when conversations will most likely be moderated. This knowledge can be wrapped into computational procedures. For instance, if moderation is more likely to be seen when incivility increases, sentiment analysis might be used to decrease the number of conversations and tweets to look at. To succeed, this computational approach requires translating each moderation strategy into a (statistical) function. Even though this might work for some strategies; others are hard to proxy.

A third approach – our applied approach – derives from a dictionary of phrases that are most likely used when moderators intervene in a conversation. For each of the five moderation strategies and based on the predefined theoretical grounds of moderation, we employ a set of terms and phrases to determine moderation. Based on this deductive approach, we have deduced around 70 moderation phrases in both English and German. Still, the most important challenge is contextual ambiguity: A phrase that is used to moderate can also be used in a different context – the meaning changes. As an example, let's think about the phrase "calm down": If the phrase is used in a sentence like "Please calm down. No need to get nasty.", it serves the moderation function. If, however, the sentence reads "The public needs to calm down.", the meaning of the phrase differs. There is one obvious solution to solve contextual ambiguity: we could just derive more specific phrases, e.g. "Please calm down". But in this case, we would miss the sentence "Calm down! No need to get nasty." Hence, for our set of phrases, we need to balance generalizability and specificity.

On the one hand, the phrases should precisely define moderation – which speaks towards deriving phrases that are specific to certain tweets. On the other hand, the phrases should show general applicability. If we apply too general phrases, we would end with many false positives; in the opposite case of too specific phrases, we miss many moderation tweets (false negatives). In other words: the search query is a function of the length of the phrases.

To arrive at a balanced set of phrases, we labeled moderation instances based on a judgment made by three annotators. Due to vast amounts of data – we just cannot manually label thousands (or even millions) of potential moderation candidates – we chose to manually label 1.500 tweets. In the first step, we downloaded conversations that included one of the 70 moderation phrases. In the second step, the three annotators had to judge whether a given tweet entails moderation. To provide some context, the annotators were given the root tweet, the two preceding tweets, the post containing the queried phrase as well as the two following tweets. Overall, the inter-annotator agreement varies widely. While for some terms and phrases, no agreement could be made whether this tweet contains moderation, others show a high inter-annotator agreement of about 0.95 (Krippendorff, 2004).

The disagreement is mainly due to difficulties in interpreting the degree of metatalk (see Figure *1*). We then query those phrases at a large scale that show both a high inter-annotator agreement as well as a high precision for moderation. Using this procedure, we end up with 7 phrases belonging to the strategies of tone policing, engaging participation, and agenda control. The 3 German phrases are: "entspann dich" (*calm down*), "näher erläutern" (*elaborate further*), and "bleib beim thema" (*stick to topic*); the English phrases are: "back to topic", "back to the topic", "stick to topic", and "can you elaborate". Even though we aim for an even distribution of English and German phrases for each of the three moderation strategies, not all of the translated phrases show the same precision in the other language. For instance, while the phrase "calm down" marks moderation in German tweets, we could not achieve high inter-annotator agreement for English tweets.

Research Questions and Hypotheses

Based on the previous research and the specified theoretical grounds, we derive some specific questions that summarize our expectations. Our main research question revolves around the structure of reactions to the attempts of user moderation. Hence:

RQ: Does user moderation improve deliberative quality?

As described earlier, we are dealing with social media and user responses towards intentions to shape communication through moderation means. One of the many responses is just to not respond but drop out of the communication; we cannot take responses for granted. On the other hand, users might also express their (dis)agreement with the intervention and continue the running communication. Moreover, other users might take a stance in favor or against the intervention. Our first hypothesis focuses on the response patterns and aims to extract and analyze the response patterns:

H1: User moderation produces different answer patterns than not moderated conversations.

In particular, we derive three specific hypotheses that summarize our expectations. The first hypothesis addresses the general question of whether the conversation ends or continues; the second and third hypotheses focus on the users of the communication:

H1a: User-moderated conversations are longer than unmoderated conversations.

H1b: Users that are addressed by the moderation are more likely to react in the following conversation.

H1c: Moderation engages the participation of new users.

We are not only interested in the response patterns but also the deliberative quality of the communication. This expectation is summarized in Hypothesis 2.

H2: Moderation improves deliberative communication.

We also derive more specific hypotheses – for each of the three moderation strategies. For instance, if the moderation strategy was focused on maintaining topic control, it becomes imperative to analyze whether the communication got back on track (reduce the variance of the topics). The same is true for the other two types of moderation: tone policing and elaboration support. Regarding H2, our specific expectations are

H2a: Emotion Control improves the tone of the discussion.

H2b: Elaboration Support leads to more elaboration.

H2c: Agenda Control leads the communication back to the original topic.

While the hypotheses formulate our general expectations, one important aspect is still missing: Our expectations regarding the effect size. On the one hand, and in comparison to previous research, we expect to see weaker effects of moderation. In most of our moderation cases, we are not dealing with hate speech. It is only single users who try to do their best in supporting and maintaining a certain degree of deliberative quality in communication. And we expect most users to just not care. In comparison to hate speech, the quality of communication is already rather good, so there is only a small need to intervene (and no need to delete tweets or ban users for legal reasons).

On the other hand, the effect of an intervention might only be seen directly after the intervention took place. We don't expect an intervention to influence communication in the long run. With more tweets posted, the effect should lose its relevance. The same is true for the number of tweets before the intervention: we don't expect users to be aware of the complete chain of tweets. With chains of 100 or even more tweets, users only jump into the discussion at a certain tweet, not willing to read through all of the previous tweets. By design, it is also rather cumbersome to click through all the previous tweets on Twitter. In other words, we assume moderation to mainly be a local event. Hence, effect sizes should differ when the analysis is restricted to only a subset of the communication.

When it comes to defining the size of the subset, i.e. the number of tweets to analyze, we restrict ourselves to a window of five tweets: two tweets before the intervention, the moderation, and two tweets after the intervention. Even though the first tweet might already include ample reasons to be moderated, for instance when hate speech is seen, we are mainly interested in analyzing proper discussions. Only after the first (constructive) tweet, something might go wrong. For instance, the discussion gets heated – which in turn motivates a user to intervene. Ideally, the user that is addressed by the intervention shows some reaction, with the second tweet after the intervention being on track again.

In summary, we expect moderation to affect the participation of users, the deliberative communicative quality, while taking different windows of communication into consideration. When it comes to measuring the effects, these three types of expectations need to be combined.

Data and Methods

The hypotheses are tested on a sample of Twitter data. Before introducing the applied method, we discuss both the sampling procedure, the extracted deliberative features, the units of analysis, and the derived variables. Please note the distinction between features and variables. While features are applied to every tweet, these features are further summarized into our dependent and independent variables based on different units of analysis. To arrive at these units, we introduce the analogy to an experiment: Moderation is seen as a treatment influencing communicative quality.

Data Sampling

Our sampling strategy involves four steps. In the first step, based on the derived set of phrases, we sampled almost 140.000 tweets in English (127.782 tweets) and German

(11.776 tweets) that include one of the 7 phrases. The tweets were posted between February 2022 and February 2023, so we expect to see some variation over time. The distribution of phrases is rather unequal – both between moderation categories as well as between languages: with about 110.000 tweets, the vast majority of English tweets fall into the moderation category “engaging participation”. The most seen phrase in German belongs to the category “tone policing”.

In the second step, for a random sample of the downloaded tweets, we retrieved the full conversation. This results in a dataset of 39.371 unique conversations with a total sum of 4.298.811 tweets. But not all users directly reply to the preceding tweet. Some only refer to the conversation by using hashtags or @-mentions. Hence, we see about 10.000 conversations with only one tweet. Please also note that in some cases, the preceding tweet(s) got deleted, starting a conversation with the second (or third) tweet. To respond to these challenges, we only keep conversations that include more than five and less than 500 tweets. The data is distilled to 25.070 unique conversations with 4.126.948 tweets.

Even though the conversations are given a unique identifier by Twitter, they do differ from our understanding of a real-world conversation. Twitter conversations can be seen as a (large) network with many (smaller) discussions. Starting with the first tweet (the root of the conversations), many discussions might come into existence. Users taking part in one branch of the conversation might not be aware of the other branches. This is due to the character of Twitter: Conversations are defined by direct responses to the preceding tweet.

To extract the response patterns, we introduce the concept of conversation paths or conversation flows. These are chains of replies that start with a root post (the first post with the opening statement of a conversation) and end with a leaf (the last contribution to a conversation). Using these structures as the unit of analysis comes close to our understanding of a discussion and also facilitates the analysis process[2]. It maps the complicated network structure to quasi-linear conversations that can be interpreted analogously to a real-life discussion.

Additionally, since social media sites delete tweets and posts that are against their policies and are also required to delete content due to legal requests, some of the trees contain missing references. For instance, if a tweet gets deleted, the corresponding reply tree is broken. Hence, each of the conversation trees was validated. In the common case of deletion, the orphaned branches were attached to the root post. As per definition, the root post does not include moderation. For this reason, this procedure does not lead to semantic errors.

Similar conversation modeling describes online conversations as polyadic conversations (Magnani et al., 2012), reply-graphs (Cogan et al., 2012; Joglekar et al., 2020; Nishi et al., 2016), or implicit thread structures (Wang et al., 2021). This also allows the interpretation of a moderating post as an intervention: the result of the intervention can be computed as the sum of the effect on all conversation paths that include the moderating post. Moreover, most annotating tools are not equipped to deal with tree structures. However, if transformed into conversation flows, these can be analyzed qualitatively like any other transcribed group discussions.

Yet, we did not analyze all of the extracted paths. In the third sampling step, we took a random sample of the conversations to run through the analysis pipeline. The main reason to further decrease the amount of data is limited computing capacities: as we describe in the next section, some of our features are rather computationally

[2] To reconstruct conversational threads, we rely on the Python library *[hidden]* (*[url hidden]*).

intensive to generate. The fourth and final sampling step focuses on specific requirements that the conversation paths need to follow: our sample is based on English and German phrases (and their tweets). In some instances, however, a conversation path includes tweets of different languages. These paths were excluded from our analyses. Additionally, as described earlier, we have defined some properties for the moderation tweet: in each path, the moderation must take place earliest at the third tweet. After the moderation, we expect to see at least two subsequent tweets. These properties come close to real-world discussion and allow the discussion to evolve over time.

We finally arrived at a dataset of 5.017 conversation paths, 2.067 conversations, and 51.442 tweets. Overall, 9.652 different users were involved in these conversations; with an average of three users per path. On average, the longest reply-chain per conversation has a length of 123 tweets; the average length is 10.3 tweets. Since we are dealing with social media, most tweets are rather short: the average length of a tweet is 188 characters.

To summarize, the discussions we are looking at are written exchanges that can be perceived as a conversation tree with the original tweet being the root node and answers-relations represented as the edges that lead to the answers that are the nodes in the tree. The leaves of the tree are answers that have no replies (retweets/quotes).

Deliberative Analytics

The generated features are directly linked to the three moderation types: tone policing refers to the polarity of the tweets, elaboration support to the degree of justification, and agenda control to a topic analysis. All of these features are extracted using computational analytics.

Sentiment. To measure the polarity of tweets, we make use of a pre-trained multilingual sentiment model (Barbieri et al., 2022). The model is trained on about 188 million tweets in 8 languages and has been finetuned for sentiment analysis. For each tweet, the model indicates the probability of a positive sentiment, a neutral sentiment, and a negative sentiment. The sum of all three probabilities is 1. Hence, we base our analysis only on the probability of a positive sentiment and neglect the other two categories.

Justification. In principle, we apply a rather similar procedure for argument analytics: Based on a finetuned model, the probability of an argument is calculated for each tweet. However, we were not aware of any reliable model to extract arguments from social media. Hence, we finetuned our own RoBERTa model. The application of the finetuned model has two advantages: first, we tune the model to respect the specific argument characteristics of social media data. Most of the available models were trained on a different type of data, e.g. oral communication. Second, we extend the base model's capacities to a broader range of topics. With regard to the first goal, we rely on two Twitter datasets with manually annotated arguments: GerCTT (Schaefer & Stede, 2022) and PPT4AM (Bhatti et al., 2021). The latter goal – extending the range of topics – is achieved by including the XArgMining dataset (Toledo-Ronen et al., 2020). As the base model, we rely on the Twitter-XML-Roberta- model (Barbieri et al., 2022). With an F1 score of 0.805, the model works well in predicting arguments on social media. When the model is applied, for each tweet, the probability of an argument is given. The probability ranges between 0 (no justification) and 1 (justification).

Topic Analysis. To extract the topics of the tweets, we apply BERTopic (Grootendorst, 2022). BERTopic uses transformers and clustering algorithms to assign documents to one or more topics. Due to the bilingual nature of the conversations, we make use of the multilingual sentence transformers model as proposed by Reimers and Gurevych (2019). After the model has been applied, each tweet is assigned to a topic space. For instance, when 4 topics have been revealed, the probability of the tweet belonging to each of the 4 topics is given – with all probabilities summing up to 1.

To succeed, the applied method relies on a large set of documents as input. Our conversations do, however, not always meet this requirement. Hence, for conversations with less than 1000 tweets, we have extended the respective topic space using data from Wikipedia. The procedure includes three steps: First, using the universal dependency parser (Straka, 2017), we have extracted proper nouns. Second, for each of the nouns, we look up Wikipedia and, if available, download the respective articles. Third, the conversation is extended by a random sample of all retrieved Wikipedia sentences. With this procedure, we make sure the added sentences belong to the same topic space. After the topic model has been applied, all the added sentences were deleted. A qualitative exploration of the results shows improved performance over a model with a smaller amount of data.

Units of Analysis

The derived hypotheses require measuring the effect of moderation. If moderation improves deliberative communication on social media, we expect to see a difference between the tweets after the moderation and the tweets before the moderation. Think about an experimental setup with moderation as the treatment. This line of thinking helps to quantify the effects of moderation. If we take the analogy to an experiment seriously, we also need to think about a control group: So far, the effects are only specified for the three types of moderation: tone policing, engaging participation, and agenda control. A control group with unmoderated conversations is missing.

In principle, every conversation without moderation constitutes the control group – with a number of tweets before and after every tweet in this conversation. The intervention is synthetically created, e.g. by randomly sampling one tweet within conversations without moderation. Given the assumption that our concept to identify deliberative moderation on social media is encompassing, we don’t expect to see effects of moderation in the control group. However, as the discussion naturally progresses, some effects unrelated to moderation might take place. For instance, the tone of the discussion might improve anyways. Users might also feel the need to justify their claims as the discussion progresses. Or users might come back to the original topic regardless of an intervention.

To identify such potentially unrelated effects, we create a control group by taking a random sample of all paths without moderation. Then, in a second step, we randomly sample one tweet as a fictitious non-moderation intervention. The same rules apply to this type of non-intervention: It must take place earliest at the third tweet, and the conversation needs to last longer than one tweet after the sampled tweet. In total, we sampled 2.500 cases of non-moderated conversations.

As stated by the hypotheses, we need to measure the effects both globally and locally. Hence, we create two datasets: The original dataset is used as-is for analyzing the effects of moderation on the global level; no restrictions to the number of tweets apply. We take the full length of conversations. To evaluate the effects locally, we have generated a second dataset that restricts the conversations to five tweets: besides the moderation we keep two tweets before and after the intervention.

When it comes to combining the first and second set of hypotheses, we need to create additional datasets. We are not only interested in comparing pre- to post-moderation tweets, but also in comparing the effects for two types of users: users that have tweeted both before *and* after the intervention and users that only participated before *or* after the intervention.

Hence, we end up with six datasets: (*All*) Complete conversations with all users; (*All n=5)* Subset of 5 tweets with all users; (*Same*) Complete conversations with users that tweet both before and after the intervention; (*Same n=5*) Subset of 5 tweets with users that tweet before and after the intervention; (*Different*) Complete conversation with users that either tweet before or after the intervention; (*Different n=5*) Subset of 5 tweets with users that either tweet before or after the intervention. All of our variables are generated for each of the 6 datasets[3].

Variables

As stated in our hypotheses, we focus on two distinct aspects of deliberative quality: First, the underlying response patterns. Second, we are interested in quantifying the effect size and strength of moderation. When it comes to the independent variables, our most important explanatory variable is the type of moderation. Additionally, to quantify the effect size, we also add moderation tweet characteristics and some control variables to the models.

Dependent Variables

Our dependent variables indicate the difference in either user behavior (first set of hypotheses) or deliberative quality (second set of hypotheses) before and after the

[3] The data is available at *[url hidden]*. Due to Twitter API terms, we only provide the datasets including the differences of the dependent and independent variables.

moderation. User behavior is operationalized by counting the number of users that take part in the conversation before or after the moderation, or both. For tone policing to be successful, we need to see a difference in sentiment values before and after the moderation. The same is true for the degree of justification: If a user asks for elaboration, other users might more likely be inclined to provide arguments for their claims. Regarding topics, users should get back to the original topics that were addressed before the moderation.

While sentiment and justification can be compared more easily by subtracting the mean of pre-moderation values from the mean of post-moderation values – comparing topics is more difficult. The main challenge is to compare a set of values over a range of topics before and after the moderation. We addressed this challenge by calculating the cosine similarity of the mean of pre- and post-moderation values. Higher values indicate similarity of topics, i.e. the tweets after the moderation address the same topics; lower values indicate that the topic distribution changes. Please note the different scale of the dependent variables: The difference in sentiment and argumentation varies between -1 and 1, the cosine similarity ranges between 0 and 1.

Figure 2 visualizes our dependent variables for deliberative quality. The values of the difference between post- and pre-moderation behavior are shown for each of the six datasets. Values greater than 0 indicate that the post-moderation values are greater than the pre-moderation values; values lower than 0 indicate the reverse scenario. For example, when users engage in tone policing, the change in sentiment values is positive when all participating users are considered (All). The same is true for users that tweeted both before and after the intervention (Same). However, users who only tweet before or after the intervention, show to decrease polarity – the conversation gets more unfriendly. This effect is amplified by a more narrow window of tweets around the intervention (n = 5); it even changes its effect direction for all users. While we see a rather similar picture for engaging participation, both agenda control and no moderation improve the conversation.

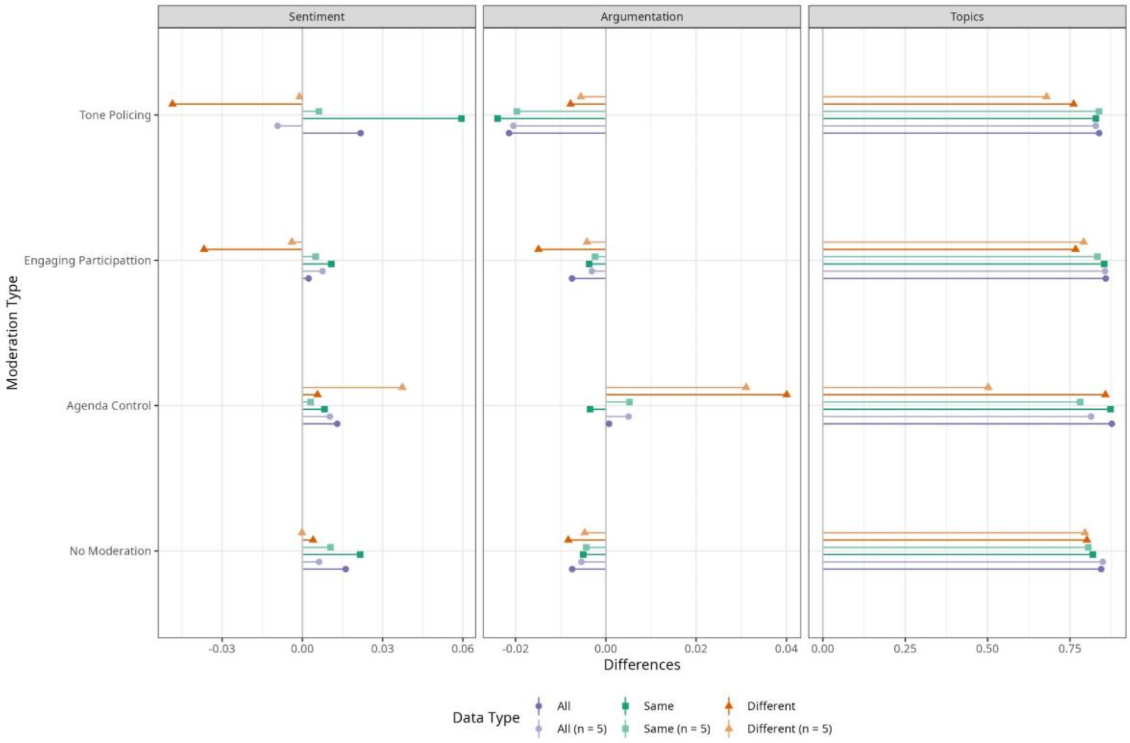

Figure 2: Values of the dependent variables for all six datasets

The level of justification decreases for both tone policing, engaging participation, and no moderation. If agenda control is employed, a different picture emerges: only when the differences for the full conversation with the same users are calculated, the results are in line with the other types of moderation. For the other types of data, the tweets after the intervention see more justification.

The descriptive statistics for topics have to be interpreted based on the different scale: Most values are higher than 0.75 – indicating that post-intervention tweets address similar topics as pre-moderation tweets. There is one trend to be seen from

Figure 2: users that only participate before or after the intervention show less motivation to take up previous topics.

Independent Variables

Our main independent variable is the **type of intervention**. In total, the dataset includes 2.610 moderation interventions. Due to missing data when generating the six different datasets, the numbers differ and are shown in Table 1. Besides no moderation (we sampled 2500 conversations), engaging participation is seen most often.

Table 1: Number of moderation types

	All			**n = 5**		
Moderation Type	**Prepost**	**Same**	**Diff**	**Prepost**	**Same**	**Diff**
no moderation	1999	1707	1739	1480	1274	842
agenda control	329	314	190	273	266	57
engaging participation	1393	1253	1094	1051	972	625
tone policing	221	191	167	161	140	105

We also include some important tweet characteristics into our models. In principle, we assume post-moderation behavior to be influenced by the type of moderation, but we also believe more friendly, more justified, and more topic-linked moderation interventions to make a difference. Hence, moderation behavior is included by **sentiment prediction**, **argument prediction**, and **cosine value**. The operationalization is the same as for the dependent variables.

Moreover, some control variables are used: First, we use the **number of characters** of the intervention as a measure for the length of the intervention. The underlying assumption is that longer interventions have a stronger effect. The **number of posted tweets** quantifies the difference in participation frequency of users between pre- and post-moderation tweets. Higher values indicate different user behaviors of pre-

and post-moderation, e.g. when many new users join the conversation right after the intervention. The occurrence of the moderation within the conversation is measured by the variable **position**. As stated in our previous sections, the lowest value is 3. We also control for the number of tweets within the complete conversation (**conversation length**) and the number of tweets of the conversation path (**path length**). The length of the path is at least two tweets longer than the position of the moderation. Finally, to control for incomplete conversations, we include variables indicating the **number of deleted tweets** and whether the path was **fixed to the root post**. The descriptive statistics are shown in Table A *1* (full datasets) and Table A *2* (subset of n=5) in the appendix.

Method

The first set of hypotheses is operationalized by count variables: the length of moderated and unmoderated conversations and the number of participating users – both before and after the moderation intervention. To compare these frequencies, we both apply t-tests (to compare the average length between unmoderated and moderated conversations) as well as chi square tests (to compare user patterns). Our three main dependent variables to proxy deliberative quality – sentiment, argumentation, topics – are normally distributed and range between -1 and 1, for topics between 0 and 1. Consequently, we apply linear regression models to estimate the effects of the independent variables. We chose to take “no moderation” as the reference group for our main independent variable. To demonstrate the effects over the range of possible moderation types, we compute and visualize predicted effects. The predicted values are based on holding all other independent variables either at their mean values (for continuous variables) or at their mode (for categorical variables). In the following section, we only show the results for our most important set of independent variables, namely the type of moderation and the moderation behavior. The full models including the controls are available in the appendix.

Results

Our findings are summarized for each of our sets of hypotheses. We first present results for the analysis of the patterns, then we describe results for sentiment, justification, and topics.

Patterns of Autor Compositions and Conversation Lengths

The data shows that unmoderated conversations combine on average 9.53 (std. dev. 6.82) tweets; moderated conversations prove to be longer with 14.1 (std. dev. 21.8) tweets. With a mean of 29.5 tweets (std. dev. 39.5), conversations with agenda control are the longest, followed by conversation of engaging participation (mean 11.4, std. dev. 15.1), and conversation of tone policing (mean 8.53, std. dev. 6.03). With the exception of tone policing, the average lengths of conversations differ statistically significantly. Hence, hypothesis 1a is mostly confirmed: User moderated conversations are longer than unmoderated conversations.

Similarly, we confirm both hypothesis 1b and 1c. In total, our sample of data includes 2.676 moderated conversations in which users both participated before and after the intervention. In 1.237 conversations, users only participated after the intervention. The same ratio applies to unmoderated conversations: the frequencies are 2.411 and 1426, respectively. When the chi squared test is applied, the differences are significant: moderation both motivates users to participate again, as well as attracts new users. The ratio also holds if moderation is included as a categorical variable. We also see that the majority of the conversations within the 5-tweet window are dialogues, i.e. conversations with only two participants.

Sentiment

In general, as seen by Table *2,* throughout the models, the estimated effects of the differences are rather low. This is to be expected and confirms our expectation. This pattern also applies to the other two dependent variables – degree of argumentation and cosine similarity (see Table *3* and Table *4*).

Table 2: Linear regression of sentiments

	All		**Same**		**Diff**	
	Full	**n=5**	**Full**	**n=5**	**Full**	**n=5**
Intercept	-0.0304	-0.0199	0.0049	-0.0347	-0.0252	-0.0263
Moderation Type (Ref: No Moderation)						
Tone Policing	-0.0055	-0.0108	0.0330*	0.0058	-0.0522^{+}	0.0149
Engaging Participation	-0.0144^{+}	0.0115	-0.0073	0.0179	-0.0343*	0.0068
Agenda Control	-0.0092	0.0093	-0.0094	0.0053	0.0276	0.0262
Moderation Behaviour						
Sentiment(+)	0.1139***	0.0447*	0.0646***	0.0201	0.2073***	0.1478***
Argument(1)	0.0730*	0.0607	0.0594	0.0484	-0.0114	0.0188
Cosine Value	0.0134	-0.0049	-0.0108	-0.0289*	0.0175	-0.0083
Controls	*(see appendix)*					
Num.Obs.	3703	2826	3244	2524	1565	1302
R2 Adj.	0.017	0.005	0.007	0.004	0.039	0.010
AIC	-2682.1	-1758.2	-1970.6	-1021.4	-50.0	109.6

Dependent variable: Sentiment value. *** $p < 0.001$, ** $p < .01$, * $p < 0.05$, + $p < 0.1$

When the effects of the three moderation types are compared to the reference type of unmoderated conversations, almost no significant differences can be seen. In comparison to unmoderated conversations, tone policing increases the overall tone of the conversations; but only for users who have already taken part in the discussion before the intervention. On the contrary, the overall tone decreases for different users –

which is in line with our hypothesis. These contradictory results point towards the fact that users with a history in the discussion react differently towards the intervention than users who are new to the conversation. The first type of users is directly confronted by the intervention and might disagree with the intervention; the latter type of users only jumps into the discussion at a later stage. Interestingly, the effects are only significant for the complete conversations. The narrow window of 5 tweets does not make a difference.

Engaging participation shows significant effects for both all users as well as different users. In both scenarios, the effect is negative, i.e. the overall tone of the conversation decreases. Again, this is in line with our hypothesis. No significant effects are seen for engaging participation. In sum, with regard to the type of intervention, there is mixed evidence. Under some circumstances, moderation makes a statistically significant difference.

The results in Table *2* also show significant effects for moderation behavior: being friendly while at the same time confronting other users to not being friendly increases the differences in polarity. The conversations become more friendly. The same effect is seen for the level of argumentation – but only for the general pre- vs. post-moderation data. Providing justifications when users intervene increases polarity.

With regard to the control variables (see Table A *3*), systematic effects are only seen for different users. First, the position of the interventions makes a difference – with later interventions to increase the sentiment for new users. Second, the longer the conversation, the more unfriendly the reaction of new users.

Justification

The results for justification are given in Table *3*. When argumentation is taken as the dependent variable, the differences between unmoderated and moderated conversations becomes apparent: throughout the models, the intercept is negative and, in 3 of the 6 models, statistically significant. However, the effects differ per type of intervention. While tone policing decreases the degree of justification in the subsequent tweets even more, agenda control increases justification. The effect of agenda control is only significant for new users joining the discussion, for both the full as well as the 5-tweet window of conversations. This, again, points towards the fact that the success of an intervention depends upon the type of user: users that have been participating before the intervention do not feel the need to provide arguments. On the contrary, new users show a different pattern. In sum, hypothesis 2b is only partly confirmed.

Table 3: Linear regression of argumentation

	All		**Same**		**Diff**	
	Full	**n=5**	**Full**	**n=5**	**Full**	**n=5**
Intercept	-0.0390***	-0.0367**	-0.0192	-0.0192	-0.0674**	-0.0266
Moderation Type (Ref: No Moderation)						
Tone Policing	-0.0153*	-0.0249**	-0.0180*	-0.0312**	-0.0033	-0.0028
Engaging Participation	-0.0002	0.0004	0.0033	-0.0023	-0.0074	-0.0006
Agenda Control	0.0058	0.0009	-0.0001	-0.0016	0.0553***	0.0334^{+}
Moderation Behaviour						
Sentiment(+)	0.0081	0.0183^{+}	0.0166^{+}	0.0148	-0.0120	0.0282
Argument(1)	0.0578***	0.0503**	0.0115	0.0272	0.1341***	0.0946*
Cosine Value	0.0051	0.0048	0.0091	0.0071	-0.0185*	-0.0328**
Controls			*(see appendix)*			
Num.Obs.	3585	2784	3148	2487	1467	1237
R2 Adj.	0.007	0.010	0.006	0.009	0.025	0.011
AIC	-8649.5	-5952.5	-6634.4	-4576.4	-2726.9	-1925.6

Dependent variable: Argumentation. *** $p < 0.001$, ** $p < .01$, * $p < 0.05$, + $p < 0.1$

The models also reveal the degree of justification of the intervention to make a significant difference: justified interventions attract more justification in the subsequent tweets. This relationship is statistically significant when the generic full pre-post or different users dataset are taken into account. Users who have participated before the intervention do not provide more arguments. When previous topics are addressed, less justified tweets are seen for users joining the conversation. Sentiment is only (marginally) significant for some of the user scenarios.

With regard to the control variables (as shown in Table A *4*), with two exceptions, no clear patterns can be seen. First, for the 5-tweet window, the later the intervention, the less likely users justify their positions. Second, as described earlier, when generating the conversation paths, some missing references were found. In such cases, we took the root as the start of the path, but kept a variable to control for our decision. The control variable is statistically significant in 3 out of the 6 scenarios: the degree of justification increases. Hence, our results need to be interpreted with some caution.

Topics

The results for topic similarity in Table *4* also show a clear difference between unmoderated and moderated conversations. Unmoderated conversations are more coherent in their topics. When agenda control is applied, topics mentioned before the intervention are less likely addressed. The effect of agenda control is statistically significant when only 5 tweets are analyzed; it does not affect the full conversation. Hence, hypothesis 1c is partly confirmed.

Table 4: Linear regression of topics

	All		**Same**		**Diff**	
	Full	**n=5**	**Full**	**n=5**	**Full**	**n=5**
Intercept	0.5612***	0.5298***	0.5136***	0.4228***	0.3414***	0.4266***
Moderation Type (Ref: No Moderation)						
Tone Policing	0.0040	-0.0041	-0.0099	-0.0057	0.0234	0.0170
Engaging Participation	0.0099	0.0086	0.0109	0.0112	0.0103	0.0350^{+}
Agenda Control	-0.0187	-0.0635***	-0.0136	-0.0536**	0.0065	-0.2605***
Moderation Behaviour						
Sentiment(+)	-0.0054	0.0099	0.0166	0.0377	-0.0306	-0.0212
Argument(1)	-0.0469	-0.0638	-0.0809	-0.0705	0.0055	-0.1237
Cosine Value	0.3371***	0.3999***	0.4146***	0.4966***	0.4514***	0.4509***
Controls			*(see appendix)*			
Num.Obs.	3703	2826	3244	2524	1565	1302
R2 Adj.	0.154	0.193	0.204	0.246	0.226	0.224
AIC	-895.4	-591.5	-252.1	131.4	336.0	406.6

Dependent variable: Cosine value. *** p < 0.001, ** p < .01, * p < 0.05, + p < 0.1

Moderation behavior is only significant for cosine similarity: if users refer to previous topics when intervening, the subsequent tweets also more likely address the same topics. This effect is seen throughout the models, no matter which of the six scenarios is analyzed. The other two moderation behaviors – sentiment and justification – do not influence the post-moderation tweets.

When it comes to the control variables (as seen in Table A 5), longer moderation interventions motivate subsequent users to come back to the original topics. However, the effect is rather small. Also, the position of the intervention within the conversation is significant: Later interventions decrease topic similarity for users that have been participating both before and after the intervention. Similarly, the length of the conversation correlates positively with cosine similarity. This is to be expected as longer conversations provide more possibilities to address more topics, and increase the topic space for later tweets. It becomes easier to address at least one of the various topics. The finding that, for new users, the number of deleted tweets decreases topic similarity, is also to be expected. In many instances, users refer to the later deleted tweet, which results in a low topic similarity.

Predicted Effects

In Figure *3*, the predicted effects are shown. In principle, the figure summarizes our previous findings: unmoderated and moderated conversations differ. However, these differences are only statistically significant for certain specific scenarios. In most scenarios, tone policing increases the overall tone of the conversations, decreases the degree of argumentation, and motivates users to address the original topics. Engaging participation shows a similar pattern: the sentiment value increases, justification decreases, and original topics are addressed. Agenda control is the most restricted and only proves to be successful given very specific settings. The reference group unmoderated conversations also reveal effects unrelated to specific interventions: The sentiment value increases, the degree of justification decreases, and original topics are addressed to a large degree.

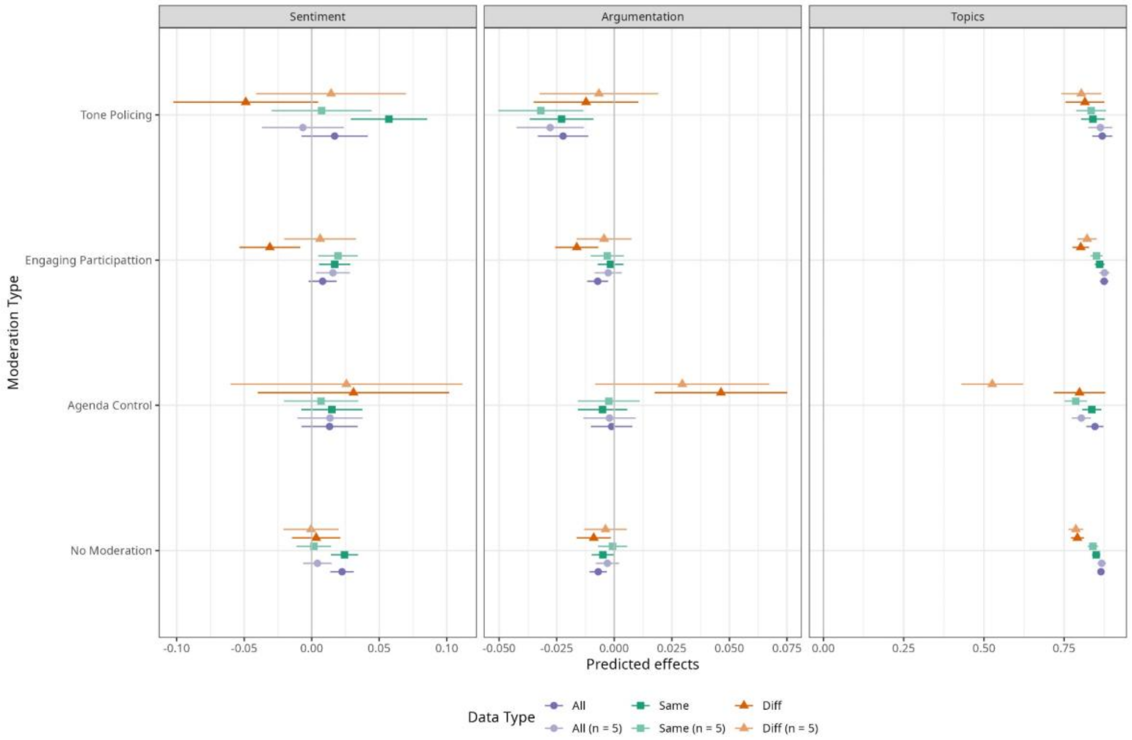

Figure 3: Predicted effects

The most important findings in Figure *3* are revealed when the differences to the unmoderated conversations are taken into account. Tone policing amplifies the positive effect for sentiments when users take part both before and after the intervention. The same is true for the degree of argumentation: tone policing decreases the willingness to provide arguments. No clear difference for topic similarity is seen. For new users, engaging participation shows an opposite effect: the sentiment value decreases. Also, engaging participation further decreases the degree of justification. Again, no clear difference can be seen for topic similarity. Finally, the effect of agenda control is rather different than for unmoderated conversations: New users provide more arguments. When, however, topic similarity is analyzed, only within the window of 5 tweets, a clear difference to unmoderated conversations is seen.

In a nutshell: Figure *3* partly confirms our derived hypotheses. Given certain circumstances – moderation type, type of users, full or narrow window of conversations – user moderation has an effect on the subsequent tweets. In some instances, moderation improves the conversation (which is in line with our hypotheses); in some other instances, moderation shows the opposite or no effects.

Discussion

Firstly, we observed that user moderation, in contrast to more targeted, collective civic moderation, tends to produce smaller effect sizes on deliberative outcomes like the civility of the tone, the amount of justification for claims, or the coherence of the topic at hand. Our study focused on general discussions rather than specifically on hate speech, which allowed us to observe the broader implications of moderation across various conversation types. It can be argued that perfectly deliberative discussions with the highest standard of civility need moderation, too and here the rationalistic idea of moderation might apply the best. Instead of using single posts as the unit of analysis, the conversations were modeled as linear pathways in a reply tree. Most research concerned with deliberation online focuses on single posts or author groups as it is easier to sample discussions based on these criteria. Having a logical access point is important for querying the vast data available. Another explanation might be that it requires more computational methods to model and investigate the conversation trees as a graph.

Using this novel way of modeling conversations online, we found that longer threads of discussion are relatively rare, which can be attributed to the drop-offs in participation following moderation interventions. Our data indicates that dialogues, where two users engage directly with each other, are the most typical form of conversation. In these dialogues, the person addressed usually responds immediately, creating a dynamic interaction. This has certain implications for deliberation research online. As deliberation is thought of as a group process, dialogues need different moderation strategies. It became apparent that strategies tend to be more effective when they are appropriately tailored to the specific context of the conversation. However, it is important to note that our research did not include an in-depth qualitative exploration. Such an analysis was not the focus of this paper, as we concentrated more on quantitative assessments of moderation effects. For instance, it would be interesting to see if moderation techniques can influence the character of the response patterns which could be used as a proxy for openness. This would cater to the idea of supporting more participation to improve deliberative quality.

Finally, we also explored windows of five consecutive posts in contrast to the complete reply paths. Our findings suggest that there are diminishing effects of the moderating intervention the longer the conversation goes on. This is in line with current research on the attention span of users and the limitations of online platforms that don't always show the user the full picture of what happened during the previous communication. This means that the area of effect is small and repeated interventions might be an interesting object for future research.

Limitations

In this study, we focus on the impact of moderation in online discussions, with our analysis being exclusively applied to the Twitter platform. Here, our study does not address sequential effects. This means that we did not analyze how moderation strategies might influence subsequent interactions in a conversation thread over time. Understanding these dynamics could provide deeper insights into the long-term effectiveness of moderation.

Another limitation is that our analysis was confined to just seven moderation phrases. This restriction potentially overlooks a broader range of moderation tactics that could vary in effectiveness and applicability. Exploring a wider array of moderation strategies could offer a more comprehensive understanding of how different approaches influence online discussions.

We also considered the variation of topics within our study. A key question that arises is whether there are specific topics that attract more moderation and subsequently exhibit better effects from these interventions. This aspect highlights the potential variability in moderation effectiveness depending on the subject matter of the discussion.

It is crucial to note that this was not a user-centric study. Our approach focused more on the moderation phrases and their immediate impacts, rather than delving into the perspectives and behaviors of individual users. For this reason, the interpretation of the results is sometimes challenging.

Furthermore, we acknowledge the potential issue of endogeneity in our selection of observations. The reason why users choose to intervene in conversations (or refrain from doing so) remains unclear, which could pose a challenge in fully understanding the dynamics at play. This uncertainty is compounded by the limited validity of our control variables, which may not fully account for all relevant external factors influencing user behavior.

In summary, while our study provides valuable insights into the effects of moderation on Twitter, these findings are shaped by certain limitations regarding scope, methodology, and the depth of user behavior. Future research could benefit from addressing these gaps to gain a more holistic understanding of online moderation practices.

Conclusion

Even civil conversations benefit from moderation and we evangelize changing the unit of analysis to the conversation threads instead of the posts of uncivil behavior. This way, online deliberation can be researched more analogously to offline deliberation making the results more comparable and generalizable.

Although there is a small but overall significant effect of the user moderation, the intended effect of the moderation strategies is reached only in some circumstances dependent on the stability of the participant composition. In contrast, leading by example works more reliably. For instance, the findings indicate that being polite engenders politeness to some degree. Similarly, staying on topic makes it more likely that future posts will also stay on topic. This should motivate users that want to improve the public sphere online to be role models. It is harder to recommend specific strategies. These depend on the situation. In order to provide advice for platforms, civic user groups or public policy, further analysis of the long term effects of moderation is needed. But it should be a positive sign that ordinary users engage in and have positive effects with moderating metatalk without having an agenda other than their personal interests. This gives a different spin to the bleak image of the lost public deliberative online sphere.

Acknowledgments

The authors would like to thank [*hidden*] for their support in annotating the data.

Funding

This work was supported by the [*hidden*] under Grant [*hidden*].

References

Allen, J. J., Anderson, C. A., & Bushman, B. J. (2018). The General Aggression Model. *Current Opinion in Psychology*, *19*, 75–80. https://doi.org/10.1016/j.copsyc.2017.03.034

Bächtiger, A., & Parkinson, J. (2019). Unpacking Deliberation. *Mapping and Measuring Deliberation*, 19–44. https://doi.org/10.1093/oso/9780199672196.003.0002

Barbieri, F., Anke, L. E., & Camacho-Collados, J. (2022). *XLM-T: Multilingual Language Models in Twitter for Sentiment Analysis and Beyond*.

Bhatti, M. M. A., Ahmad, A. S., & Park, J. (2021). Argument Mining on Twitter: A Case Study on the Planned Parenthood Debate. In *Proceedings of the 8th Workshop on Argument Mining* (pp. 1–11). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.argmining-1.1

Cai, J., & Wohn, D. Y. (2019). What Are Effective Strategies of Handling Harassment on Twitch? Users' Perspectives. In *CSCW '19, Conference Companion Publication of the 2019 on Computer Supported Cooperative Work and Social Computing* (pp. 166–170). Association for Computing Machinery. https://doi.org/10.1145/3311957.3359478

Cialdini, R. B., & Goldstein, N. J. (2004). Social Influence: Compliance and Conformity. *Annual Review of Psychology, 55*(1), 591–621. https://doi.org/10.1146/annurev.psych.55.090902.142015

Cogan, P., Andrews, M., Bradonjic, M., Kennedy, W. S., Sala, A., & Tucci, G. (2012). Reconstruction and analysis of Twitter conversation graphs. In *HotSocial '12, Proceedings of the First ACM International Workshop on Hot Topics on Interdisciplinary Social Networks Research* (pp. 25–31). Association for Computing Machinery. https://doi.org/10.1145/2392622.2392626

Edwards, A. R. (2002). The moderator as an emerging democratic intermediary: The role of the moderator in Internet discussions about public issues. *Information Polity, 7*(1), 3–20.

Friess, D., Ziegele, M., & Heinbach, D. (2021). Collective Civic Moderation for Deliberation? Exploring the Links between Citizens' Organized Engagement in Comment Sections and the Deliberative Quality of Online Discussions. *Political Communication*, *38*(5), 624–646. https://doi.org/10.1080/10584609.2020.1830322

Graham, T., & Witschge, T. (2003). In search of online deliberation: Towards a new method for examining the quality of online discussions.

Grimmelmann, J. (2015). The Virtues of Moderation. *Cornell Law Faculty Publications*. https://scholarship.law.cornell.edu/facpub/1486

Grootendorst, M. (2022). *BERTopic: Neural topic modeling with a class-based TF-IDF procedure*.

Habermas, J. (1981). *Theorie des kommunikativen Handelns*. Suhrkamp.

Habermas, J. (2009). *Diskursethik* (1st ed.). *Philosophische Texte, Bd. 3*. Suhrkamp.

Han, S.-H., Brazeal, L. M., & Pennington, N. (2018). Is Civility Contagious? Examining the Impact of Modeling in Online Political Discussions. *Social Media + Society*, *4*(3), 2056305118793404. https://doi.org/10.1177/2056305118793404

Hwang, H., Kim, Y [Youngju], & Kim, Y [Yeojin] (2018). Influence of Discussion Incivility on Deliberation: An Examination of the Mediating Role of Moral Indignation. *Communication Research*, *45*(2), 213–240. https://doi.org/10.1177/0093650215616861

Joglekar, S., Velupillai, S., Dutta, R., & Sastry, N. (2020). *Analysing Meso and Macro conversation structures in an online suicide support forum*.

Kim, N. (2016). Beyond Rationality: The Role of Anger and Information in Deliberation. *Communication Research*, *43*(1), 3–24. https://doi.org/10.1177/0093650213510943

Kluck, J. P., & Krämer, N. C. (2022). Appraising Uncivil Comments in Online Political Discussions: How Do Preceding Incivility and Senders' Stance Affect the Processing of an Uncivil Comment? *Communication Research.* Advance online publication. https://doi.org/10.1177/00936502221113812

Kraut, R. E., Resnick, P., Kiesler, S., Burke, M., & Chen, Y. (2012). *The MIT Press*. *Building Successful Online Communities: Evidence-Based Social Design* (R. E. Kraut, & P. Resnick, Eds.). MIT Press.

Krippendorff, K. (2004). Reliability in content analysis. *Human Communication Research*, *30*(3), 411–433.

Magnani, M., Montesi, D., & Rossi, L. (2012). Conversation retrieval for microblogging sites. *Information Retrieval*, *15*(3), 354–372. https://doi.org/10.1007/s10791-012-9189-9

Molina, R. G., & Jennings, F. J. (2018). The Role of Civility and Metacommunication in Facebook Discussions. *Communication Studies, 69*(1), 42–66. https://doi.org/10.1080/10510974.2017.1397038

Monnoyer-Smith, L., & Wojcik, S. (2012). Technology and the quality of public deliberation: a comparison between on and offline participation. *International Journal of Electronic Governance, 5*(1), 24–49.

Mutz, D. C. (2008). Is Deliberative Democracy a Falsifiable Theory? *Annual Review of Political Science, 11*(1), 521–538. https://doi.org/10.1146/annurev.polisci.11.081306.070308

Nishi, R., Takaguchi, T., Oka, K., Maehara, T., Toyoda, M., Kawarabayashi, K., & Masuda, N. (2016). Reply trees in Twitter: data analysis and branching process models. *Social Network Analysis and Mining*, *6*(1), 26. https://doi.org/10.1007/s13278-016-0334-0

Papacharissi, Z. (2004). Democracy online: Civility, politeness, and the democratic potential of online political discussion groups. *New Media & Society*, *6*(2), 259–283. https://doi.org/10.1177/1461444804041444

Porten-Cheé, P., Kunst, M., & Emmer, M. (2020). Online Civic Intervention: A New Form of Political Participation Under Conditions of a Disruptive Online Discourse. *International Journal of Communication*, *14*(0). https://ijoc.org/index.php/ijoc/article/view/10639

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing.* Association for Computational Linguistics.

Schaefer, R., & Stede, M. (2022). GerCCT: An Annotated Corpus for Mining Arguments in German Tweets on Climate Change. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference* (pp. 6121–6130). European Language Resources Association. https://aclanthology.org/2022.lrec-1.658

Seering, J. (2020). Reconsidering Self-Moderation: The Role of Research in Supporting Community-Based Models for Online Content Moderation. *Proc. ACM Hum.-Comput. Interact.*, *4*(CSCW2). https://doi.org/10.1145/3415178

Steenbergen, M. R., Bächtiger, A., Spörndli, M., & Steiner, J. (2003). Measuring Political Deliberation: A Discourse Quality Index. *Comparative European Politics*, *1*(1), 21–48. https://doi.org/10.1057/palgrave.cep.6110002

Straka, M. (2017). *UDPipe*. http://ufal.mff.cuni.cz/udpipe

Stromer-Galley, J. (2007). Measuring Deliberation's Content: A Coding Scheme. *Journal of Public Deliberation*, *3*(1, Article 12).

Toledo-Ronen, O., Orbach, M., Bilu, Y., Spector, A., & Slonim, N. (2020). Multilingual Argument Mining: Datasets and Analysis. In T. Cohn, Y. He, & Y. Liu (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2020* (pp. 303–317). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.findings-emnlp.29

Wang, Y.-C., Mahesh Joshi, M. J., & Cohen, W. (2021). Recovering Implicit Thread Structure in Newsgroup Style Conversations. *ICWSM*, *2*(1), 152–160. https://doi.org/10.1609/icwsm.v2i1.18629

Wohn, D. Y. (2019). Volunteer Moderators in Twitch Micro Communities: How They Get Involved, the Roles They Play, and the Emotional Labor They Experience. In *CHI '19, Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1–13). Association for Computing Machinery. https://doi.org/10.1145/3290605.3300390

Appendix

Table A 1: Descriptive statistics of independent variables, full datasets

	Prepost					Same					Diff				
	N	**Mean**	**SD**	**Median**	**Range**	**N**	**Mean**	**SD**	**Median**	**Range**	**N**	**Mean**	**SD**	**Median**	**Range**
Moderation Behavior															
Sentiment(+)	3942	0.14	0.18	0.07	0.93	3465	0.14	0.18	0.07	0.92	3190	0.12	0.16	0.06	0.93
Argument(1)	3703	0.54	0.08	0.54	0.73	3244	0.54	0.08	0.54	0.73	2966	0.53	0.08	0.54	0.70
Cosine Value	3942	0.84	0.27	0.99	1.00	3465	0.83	0.29	0.99	1.00	3190	0.80	0.33	0.99	1.00
Controls															
Tweet Length	3942	27.35	15.61	26.00	63.00	3465	27.72	15.65	26.00	63.00	3190	27.53	15.74	26.00	62.00
Posted Tweets	3942	2.50	5.44	1.58	71.54	3465	3.23	5.96	2.00	69.51	1634	0.43	1.13	0.33	15.36
Mod. Position	3942	6.86	9.00	4.00	84.00	3465	7.21	9.51	5.00	84.00	3190	6.11	4.89	4.00	55.00
Conv. Length	3942	251.37	182.54	224.00	493.00	3465	250.27	183.69	219.00	493.00	3190	277.38	180.35	282.00	493.00
Path Length	3942	11.41	12.46	8.00	119.00	3465	12.13	13.10	9.00	119.00	3190	10.04	7.16	7.00	64.00
Deleted tweets	3942	0.00	0.06	0.00	2.00	3465	0.00	0.06	0.00	2.00	3190	0.00	0.06	0.00	2.00
Fixed Path to Root	3942	0.39	0.49	0.00	1.00	3465	0.40	0.49	0.00	1.00	3190	0.44	0.50	0.00	1.00

Table A 2: Descriptive statistics of independent variables, n=5 datasets

	Prepost					Same					Diff				
	N	**Mean**	**SD**	**Median**	**Range**	**N**	**Mean**	**SD**	**Median**	**Range**	**N**	**Mean**	**SD**	**Median**	**Range**
Moderation Behavior															
Sentiment(+)	2965	0.14	0.18	0.07	0.92	2652	0.14	0.18	0.07	0.92	1629	0.12	0.16	0.06	0.90
Argument(1)	2826	0.54	0.08	0.54	0.73	2524	0.54	0.08	0.54	0.73	1547	0.53	0.08	0.54	0.63
Cosine Value	2965	0.84	0.28	0.99	1.00	2652	0.84	0.30	0.99	1.00	1629	0.78	0.33	0.99	1.00
Controls															
Tweet Length	2965	28.16	15.67	26.00	63.00	2652	28.28	15.67	26.00	63.00	1629	26.63	15.35	24.00	62.00
Posted Tweets	2965	1.12	1.16	1.00	24.50	2652	1.52	0.57	1.50	3.17	1369	0.19	1.55	0.00	26.00
Mod. Position	2965	7.31	10.15	4.00	84.00	2652	7.62	10.57	5.00	84.00	1629	5.26	4.35	4.00	55.00
Conv. Length	2965	252.10	180.35	224.00	493.00	2652	247.89	181.46	210.50	493.00	1629	255.26	175.94	227.00	493.00
Path Length	2965	13.03	13.86	9.00	118.00	2652	13.47	14.39	10.00	118.00	1629	10.17	6.77	8.00	57.00
Deleted tweets	2965	0.00	0.07	0.00	2.00	2652	0.00	0.07	0.00	2.00	1629	0.00	0.07	0.00	2.00
Fixed Path to Root	2965	0.39	0.49	0.00	1.00	2652	0.40	0.49	0.00	1.00	1629	0.40	0.49	0.00	1.00

Table A 3: Linear regression of sentiments

	All		**Same**		**Diff**	
	Full	**n=5**	**Full**	**n=5**	**Full**	**n=5**
Intercept	-0.0304	-0.0199	0.0049	-0.0347	-0.0252	-0.0263
Moderation Type (Ref: No Moderation)						
Tone Policing	-0.0055	-0.0108	0.0330*	0.0058	-0.0522^{+}	0.0149
Engaging Participation	-0.0144^{+}	0.0115	-0.0073	0.0179	-0.0343*	0.0068
Agenda Control	-0.0092	0.0093	-0.0094	0.0053	0.0276	0.0262
Moderation Behaviour						
Sentiment(+)	0.1139***	0.0447*	0.0646***	0.0201	0.2073***	0.1478***
Argument(1)	0.0730*	0.0607	0.0594	0.0484	-0.0114	0.0188
Cosine Value	0.0134	-0.0049	-0.0108	-0.0289*	0.0175	-0.0083
Controls						
Tweet Length	-0.0001	-0.0001	-0.0003	-0.0003	0.0004	0.0008
Posted Tweets	-0.0014	-0.0075*	0.0010	0.0174*	0.0015	-0.0018
Mod. Position	0.0012	0.0019*	0.0004	0.0008	0.0119**	0.0053^{+}
Conversation Length	0.0000	0.0000^{+}	0.0000	0.0001***	0.0000	0.0000
Path Length	-0.0007	-0.0016*	-0.0010	-0.0010	-0.0061***	-0.0032*
Deleted Tweets	-0.0018	0.0091	0.0044	0.0041	-0.0108	0.0296
Fixed Path to Root	-0.0075	-0.0126	-0.0107	-0.0140	-0.0120	-0.0309^{+}
Num.Obs.	3703	2826	3244	2524	1565	1302
R2 Adj.	0.017	0.005	0.007	0.004	0.039	0.010
AIC	-2682.1	-1758.2	-1970.6	-1021.4	-50.0	109.6

Dependent variable: Sentiment value. *** p < 0.001, ** p < .01, * p < 0.05, + p < 0.1

Table A 4: Linear regression of arguments

	All		**Same**		**Diff**	
	Full	**n=5**	**Full**	**n=5**	**Full**	**n=5**
Intercept	-0.0390***	-0.0367**	-0.0192	-0.0192	-0.0674**	-0.0266
Moderation Type (Ref: No Moderation)						
Tone Policing	-0.0153*	-0.0249**	-0.0180*	-0.0312**	-0.0033	-0.0028
Engaging Participation	-0.0002	0.0004	0.0033	-0.0023	-0.0074	-0.0006
Agenda Control	0.0058	0.0009	-0.0001	-0.0016	0.0553***	0.0334^{+}
Moderation Behaviour						
Sentiment(+)	0.0081	0.0183^{+}	0.0166^{+}	0.0148	-0.0120	0.0282
Argument(1)	0.0578***	0.0503**	0.0115	0.0272	0.1341***	0.0946*
Cosine Value	0.0051	0.0048	0.0091	0.0071	-0.0185*	-0.0328**
Controls						
Tweet Length	-0.0001	-0.0001	0.0000	0.0000	-0.0003	-0.0002
Posted Tweets	0.0007	-0.0002	0.0007	-0.0045	0.0076*	-0.0002
Mod. Position	0.0005	0.0002	0.0008^{+}	0.0000	0.0045**	0.0014
Conversation Length	0.0000	0.0000	0.0000	0.0000**	0.0000	0.0000
Path Length	-0.0006*	0.0002	-0.0009*	0.0005	-0.0014*	-0.0005
Deleted Tweets	-0.0027	0.0013	-0.0076	-0.0009	0.0102	0.0087
Fixed Path to Root	0.0091**	0.0059	0.0139***	0.0140**	0.0070	-0.0015
Num.Obs.	3585	2784	3148	2487	1467	1237
R2 Adj.	0.007	0.010	0.006	0.009	0.025	0.011
AIC	-8649.5	-5952.5	-6634.4	-4576.4	-2726.9	-1925.6

Dependent variable: Argumentation. *** p < 0.001, ** p < .01, * p < 0.05, + p < 0.1

Table A 5: Linear regression of topics

	All		**Same**		**Diff**	
	Full	**n=5**	**Full**	**n=5**	**Full**	**n=5**
Intercept	0.5612***	0.5298***	0.5136***	0.4228***	0.3414***	0.4266***
Moderation Type (Ref: No Moderation)						
Tone Policing	0.0040	-0.0041	-0.0099	-0.0057	0.0234	0.0170
Engaging Participation	0.0099	0.0086	0.0109	0.0112	0.0103	0.0350^{+}
Agenda Control	-0.0187	-0.0635***	-0.0136	-0.0536**	0.0065	-0.2605***
Moderation Behaviour						
Sentiment(+)	-0.0054	0.0099	0.0166	0.0377	-0.0306	-0.0212
Argument(1)	-0.0469	-0.0638	-0.0809	-0.0705	0.0055	-0.1237
Cosine Value	0.3371***	0.3999***	0.4146***	0.4966***	0.4514***	0.4509***
Controls						
Tweet Length	0.0009***	0.0008**	0.0007*	0.0010**	0.0012*	0.0010^{+}
Posted Tweets	-0.0013	0.0088*	-0.0020	0.0063	-0.0064	0.0043
Mod. Position	-0.0037***	-0.0005	-0.0034**	0.0000	0.0015	-0.0049
Conversation Length	0.0000	0.0000	-0.0001**	-0.0001*	0.0001	0.0001
Path Length	0.0045***	0.0009	0.0046***	0.0008	0.0031^{+}	0.0051**
Deleted Tweets	-0.0915	-0.0790	-0.0488	-0.0131	-0.2518**	-0.1901*
Fixed Path to Root	0.0020	-0.0089	0.0048	0.0024	-0.0070	-0.0310^{+}
Num.Obs.	3703	2826	3244	2524	1565	1302
R2 Adj.	0.154	0.193	0.204	0.246	0.226	0.224
AIC	-895.4	-591.5	-252.1	131.4	336.0	406.6

Dependent variable: Cosine value. *** p < 0.001, ** p < .01, * p < 0.05, + p < 0.1