Moderation in the Wild: Investigating User-Driven Moderation in Online Discussions

Neele Falk[*], **Eva Maria Vecchi**[*], **Iman Jundi**[*], and **Gabriella Lapesa**[†]
[*]Institute for Natural Language Processing, University of Stuttgart, Germany
[†]GESIS - Leibniz Institute for Social Sciences and Heinrich-Heine University of Düsseldorf
[*]first[-middle].last@ims.uni-stuttgart.de, [†]gabriella.lapesa@gesis.org

Abstract

Effective content moderation is imperative for fostering healthy and productive discussions in online domains. Despite the substantial efforts of moderators, the overwhelming nature of discussion flow can limit their effectiveness. However, it is not only trained moderators who intervene in online discussions to improve their quality. "Ordinary" users also act as moderators, actively intervening to correct information of other users' posts, enhance arguments, and steer discussions back on course.

This paper introduces the phenomenon of user moderation, documenting and releasing UMOD, the first dataset of comments in which users act as moderators. UMOD contains 1000 comment-reply pairs from the subreddit r/changemyview with crowdsourced annotations from a large annotator pool and with a fine-grained annotation schema targeting the functions of moderation, stylistic properties (aggressiveness, subjectivity, sentiment), constructiveness, as well as the individual perspectives of the annotators on the task. The release of UMOD is complemented by two analyses which focus on the constitutive features of constructiveness in user moderation and on the sources of annotator disagreements, given the high subjectivity of the task.

1 Introduction

Moderation is often employed to enhance the productivity and civility of online discussions (Park et al., 2012, 2021). In more deliberative contexts, such as civic participation forums, moderators go beyond merely censoring problematic comments; they actively assist participants in improving and guiding their commenting behaviour. The overarching goal is to articulate diverse viewpoints optimally, ensure their visibility, and foster an environment where everyone feels comfortable contributing their opinions (Kuhar et al., 2019; Lampe et al., 2014). In these scenarios, the moderators are trained experts who facilitate the discussions while maintaining a neutral and respectful tone.

In online discussions, however, it is surprisingly common to encounter "regular" users who take up moderator roles. Consider, for example, these two comments from the argumentative subreddit r/changemyview: "Can you give a summary of your understanding of what sociology is and what people who study it are attempting to do? I think in order to rebut your view, we need to know what your concept of the field is." and "Do you have anything to back up this statement, or is it just hyperbole?". In the first case, the "user moderator" is asking for a clarification that will enable a better discussion; in the second, the "user moderator" is requesting (in a slightly aggressive way) more evidence to support an argument.

User moderation (**UM**) is as common as it is unexplored in NLP: our work fills this gap. We document and release the **UMOD** dataset (User Moderation in Online Discussions)[1] which comprises 1000 comment-reply pairs sourced from the argumentative subreddit r/changemyview (Tan et al., 2016) and is annotated with a fine-grained annotation schema which allows us to build a comprehensive picture of the different facets of this phenomenon. To what extent do tone and sentiment play a role? Are UM comments inevitably constructive, or in which cases are they not? Which moderation functions of expert moderators (e.g., keeping discussion on topic, improve comment quality) are taken over by users more frequently? Clarifying these questions can help to identify which functions moderators should prioritize and which type of moderation can be successfully taken over by users.

Each comment-reply exchange in UMOD is annotated to determine whether the reply includes a form of moderation with respect to its parent comment, the specific moderation function performed,

[1]UMOD and the annotation guidelines are publicly available at [https://github.com/Blubberli/userMod]

Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 992–1013

the writing style of the reply (subjectivity and aggressiveness) of the reply, the sentiment expressed, and whether it is generally constructive. As the very notion of moderation, constructiveness and the perception of writing style are inherently subjective, we conducted a large-scale crowd-sourcing study involving between 7 and 10 different annotators for each instance, from a pool of 84 different annotators. Additionally, we solicited annotators' personal definitions of UM as free text. We could therefore capture and characterize a wide range of perspectives which we provide in a non-aggregated version of our dataset (Cabitza et al., 2023).

The release of the dataset is complemented by two studies employing statistical analyses to gain insights into *the empirical properties of constructive behavior* and *the sources of disagreements in the identification of UM*. We observe that users tend to engage more in content-oriented moderation functions, such as assisting others in improving their arguments or clarifying misconceptions and misunderstandings. A deeper analysis of what constitutes constructiveness in our dataset reveals that it is characterized by sufficient length, a more positive and appreciative tone, and appropriate complexity. For disagreements, we find that as the writing style deviates more from that of expert moderators, the perception of whether something qualifies as moderation becomes more subjective.

This work addresses a critical challenge attributed to high costs associated with human moderation, which introduces a significant bottleneck for large-scale discussions. A promising avenue for addressing this challenge is to identify the specific types of moderation functions that users can effectively take over, potentially reducing the burden on expert moderators. Alternatively, we can utilize this dataset to train models for that can predict whether a comment requires moderation, thus enabling semi-automatic moderation.

The contributions and potential impact of this work are therefore manifold. At the level of the *core phenomenon and research questions*, we are the first to shift the focus from expert to UM and to propose a taxonomy of UM properties which is encoded in our annotation schema. Accordingly, at the level of the *contributed resource*, UMOD fills an obvious gap and it does so combining a fine-grained annotation schema with a large pool of annotators. Last, at the level of the *potential applications* UMOD can be used to support effective semi-automatic moderation by overcoming the low-resource bottleneck (UM can be used to supplement the scarce training data from expert moderation) but also by informing content moderators about the types of moderation actions that are more popular among forum users (Vecchi et al., 2021).

2 Related Work

Online moderation in general focuses primarily on maintaining a healthy environment online. While on many newspaper discussion platforms experts are employed to remove inappropriate content, on platforms like Reddit, dedicated and engaged users take over this role by being officially granted moderation rights by the community. In this context, Park et al. (2021) have assembled a dataset about moderating community norms on Reddit. Note that we do not consider Reddit moderators as user moderators under our definition because they are appointed and acknowledged as such by the community. A more restricted definition of this type of moderation refers to the elimination of hate speech and abusive language, often called automatic content moderation (McMillan-Major et al., 2022). A broader definition of moderation involves the quality of argumentation and identifying what is appropriate (Ziegenbein et al., 2023).

In deliberative contexts, moderators aim to foster a *productive discussion*. They assist participants in articulating their arguments more effectively (making them clearer or providing evidence) and in staying on topic; they also structure the discussion by summarizing or bringing similar opinions together.

Automatic models targeting moderation in deliberative contexts (Falk et al., 2021; Falk and Lapesa, 2023) have been developed based on the dataset constructed by Park et al. (2012), which contains at online discussions from the deliberation platform RegulationRoom. Expert moderators on this platform have been trained on guidelines describing different actions to be taken over (eRulemaking Initiative et al., 2017) and the dataset contains a small set of comments annotated with these actions. Other research efforts in this field focus on investigating the effect of human moderation on participation processes (either qualitative (Skousen et al., 2020) or empirically (Esau et al., 2017)), or different ways of integrating automated support on deliberation platforms, e.g. forms of intelligent nudging if participants did not contribute over a certain amount of time (Gelauff et al., 2023).

As for UM, Malinen (2022) explore the behavior of voluntary user moderators on Facebook trough qualitative interviews. They find that the user moderator's primary goal is to improve the quality of discussions by offering personal feedback during the conversation. However, as the scale of these discussions continues to grow, their ability to provide such feedback becomes increasingly challenging. Consequently, they often find themselves compelled to employ stronger forms of moderation, including the removal of inappropriate content. More recent work on how people would moderate or evaluate appropriateness in discussions was conducted by Hettiachchi and Goncalves (2019) who collect crowd-sourced perspectives on appropriateness. They examine annotator-specific preferences and reflect on a resulting moderator bias. Related to that is the work by by Shen and Rose (2019) who conducted a meta-analysis of what Reddit users think about content moderation.

Research gaps Reviewing the background and related work relevant for investigating (user) moderation reveals two main research gaps which we aim to tackle with this work. First of all, there is limited data available for empirically investigating expert moderation in deliberative contexts and for training robust and effective models. Datasets systematically covering the broader spectrum of moderation tasks are scarce and completely absent when it comes to UM. Second, while the issue of perspectivism and subjectivity in defining appropriate behavior within discussions has recently gained prominence (Sachdeva et al., 2022; Cabitza et al., 2023) datasets that delve deeper into subjectivity and model the behavior of different annotators are very scarce, or, when it comes to moderation, completely absent. We fill both these gaps by introducing the first UM dataset, UMOD, and designing our annotation study in a way that different perspectives can be captured and the inherent subjectivity of the phenomenon accounted for.

3 Annotation

3.1 Data Collection and Pre-processing

The data for the study is sampled from the ChangeMyView corpus of Tan et al. (2016), a dataset that consists of discussion threads from the `/r/ChangeMyView` subreddit. Each thread is initiated by an original post (OP) that explains a view with several justifications. Other participants then discuss the opinion at issue and try to convince the original poster to, effectively, change their view. If they are successful they will be rewarded a 'delta'. They can also respond to each other (and the original poster can intervene as well), allowing the discussion tree to develop in-depth.

This platform is particularly suitable for the investigation of UM for the following reasons: (a) It consistently maintains a high level of discussion quality as discussions are monitored by CMV-designated moderators (particularly dedicated members of the community). Therefore, it is likely that a broader range of different forms of moderation can be found. An excessive amount of hate speech and spamming would prevent a productive unfolding of a discussion. (b) The participants themselves have a strong interest in productive discourse. It is likely that they actively contribute to controlling the quality of the discussion through various forms of moderation (Srinivasan et al., 2019; Chang and Danescu-Niculescu-Mizil, 2019; Chandrasekharan et al., 2022).

Candidate Extraction To extract potential candidates for UM, we trained two text classification models to identify whether a comment was written by a moderator or not (model details in Appendix A.1). The *expert moderation model* was trained on data from deliberative discussion forums that were moderated by trained experts.[2] By training a model to distinguish between moderation comments and user comments, we can identify new comments that closely resemble "expert moderation comments" which as a consequence serve as good candidates for our annotation study. As this type of data is extremely scarce we combined two available datasets. ∼3k comments stem from the RegulationRoom dataset (Park et al., 2012), ∼4.3k comments were extracted from the online platform `lasst-uns-streiten`,[3] . The merged dataset consists of 7.3k comments, of which 1k were written by expert moderators and 6.3k by users.

The *ChatGPT moderator model* was trained on data generated from ChatGPT and was developed to ensure a wide coverage of all potential mod-

[2]Moderators received additional qualifications (e.g. a training dedicated to moderation of deliberative discussions) and are paid for their moderation duties.

[3]`https://www.lasst-uns-streiten.de`, an e-participation project organized by the German federal state Saxony (data provided by the company Zebralog). The German data was automatically translated into English with DeepL (`https://www.deepl.com/translator`).

erator actions.[4] This decision resulted from the observation that certain actions, such as policing and maintaining topic relevance, were underrepresented in our expert moderation dataset. To address this, we provided ChatGPT with explicit instructions to generate UM comments that express specific moderation functions as described in our guidelines. As negative examples, we instructed ChatGPT to produce general user comments resembling typical Reddit contributions. It is essential to note that this dataset is relatively small, consisting of only 408 comments (half moderator comments, half user comments) and each moderator function is approximately equally represented in the dataset.

We run inference on the ChangeMyView dataset which resulted in 390k candidates from the expert moderator model and 105k candidates from the ChatGPT moderator model. We noticed a bias towards shorter comments from the ChatGPT model.

Sampling Criteria Since we are particularly interested in subjective perceptions of UM we hypothesize that most disagreement would occur with comments that strongly deviate from the style of expert moderators (neutral, calm, respectful tone). We took the predictions of a toxicity classifier[5] into account to collect annotations for comments with high and low toxicity scores. Our final annotation sample consists of 1000 comment–reply pairs. 40% are sampled from the ChatGPT moderator model, 40% from the expert moderator model and 20% negatives (not predicted as moderation by any of the two models). We restricted the comments to a length between 5 and 200 tokens. For each candidate pool (expert moderator, ChatGPT and negatives) we sampled equally from the lower and upper quartile of toxicity scores. Finally, 70% comment-reply pairs are deeper down a discussion tree, while 30% consist of the OP (the first post to open a discussion) and a direct reply to that.

3.2 Procedure

We conduct our study on the platform Prolific[6] and add a pre-screening which enforces every annotator to be fluent in English and be based in an English-speaking country and to have a high school diploma. We request a 'balanced' sample (regarding sex), an option provided by Prolific.[7] We release all annotators socio-demographic variables with the dataset and a unique, anonymous identifier (overview in Appendix Tbl. 6). Each batch consists of 100 instances and requires 9 annotators (for a total of 90 different annotators). The total costs of the study were 6,903 USD (hourly rate of 12,45 USD, corresponding to minimum wage in Germany), and the average time spent per annotator was 3.1 hours (cf. Appendix Tbl. 6). We offered a bonus payment for correctly annotating three control instances.

3.3 Annotation layers

Our primary objective was to analyze and capture the pragmatic and stylistic characteristics of UM. To achieve this, we task our annotators with evaluating an exchange between two users. Each exchange consists of a comment (OP or comment to an OP) and a reply comment. To provide additional context for the annotators, we also specify the topic of the discussion (cf. Appendix Fig. 8 for an example of the annotation task). The reply comment is the target of the annotation. The annotation layers are described below and summarized in Tbl. 2.[8]

User moderation The annotators have to specify whether the comment is or not an instance of UM. Additionally, they have to identify the moderator actions (multiple may be present within a single comment) and map them to a list of *moderation functions*. We adopt the taxonomy of moderation functions from Park et al. (2012), derived from their study on expert moderation in the RegulationRoom deliberation platform, which defines eight distinct moderator actions. The authors refined their guidelines and taxonomy through multiple iterations, and the use of the shared taxonomy here allows a direct comparison of function coverage between our dataset and theirs, revealing commonalities such as the prevalence of "improving quality" and "broadening discussion." Tbl. 3 outlines the potential functions along with their descriptions.[9]

[4] Examples of candidates extracted by this model are in Appendix Tbl. 4.

[5] SkolkovoInstitute/roberta_toxicity_classifier

[6] https://www.prolific.com

[7] A balanced pool on various demographic attributes would be ideal, but Prolific allows this only by limiting the annotator pool to the US or the UK.

[8] The guidelines were refined iteratively through annotation and development rounds, including input from the paper authors and a student annotator (20-item pilot). One last round was carried with the pre-final version of the guidelines was conducted on Prolific (20 items, 6 annotators). The final guidelines (cf. Sect. A.2) are released together with the dataset.

[9] To ensure the quality and consistency of annotations, we included three 'control instances' (cf. Appendix Tbl. 5, appendix). They serve as examples that clearly and explicitly manifest a specific moderation function. They are integrated

Evaluative features In examining the stylistic aspects and the tone of the comments, we focus on:
Sentiment (positive vs. neutral vs. negative): captures the affective dimension of the text.
Subjectivity (1-5): how strongly the comment reflects the author's personal opinions or own interpretations.
Aggressivity (1-5): evaluates the degree of aggressiveness in the text. It includes elements like sarcasm or direct attacks on the dialogue partner.
Constructiveness (1-5): this is highly relevant in argumentative discourse and deliberation and has been investigated for example in the context of discussions under newspaper articles (Kolhatkar and Taboada, 2017a,b). It covers a respectful and polite tone that ensures a healthy discussion and specific sub-dimensions related to argument and discourse quality, e.g. does the person justify their opinion? Is the comment relevant to the topic?
Annotator perspective We anticipate that the annotation tasks sketched above will be highly subjective. For this reason, we collect additional annotations to better characterize the annotator perspective. For each item, we ask the annotators *whether they agree* with the opinion stated in the reply comment in order to investigate how a bias towards a certain attitude effects the annotation (e.g. if the annotator agrees with the comment it is more likely that they will rate it with a higher constructiveness). Additionally, we ask annotators to describe, in a free-text field, *how they would describe UM*, and to do so before and after having completed the annotation batch. Finally, after the annotation, we ask them whether they carried out the task from the perspective a potential moderator or from that of a user who would receive the candidate UM comment.

3.4 Aggregation

As our study is primarily focused on the subjective perception of UM, we do not establish a classical 'gold standard'. Nevertheless, we provide our dataset in an aggregated form, in addition to our non-aggregated version. As a first step, we filter out low-quality annotations using a heuristic based on overall competence, the time taken for the study, and performance on three control instances. We calculate overall competence using MACE (Multi-Annotator Competence Estimation, Hovy et al. (2013), a probabilistic model that learns competence scores for each annotator. We calculate the overall competence by taking the average of the competence score for UM and all other annotation layers. We then remove all annotators that (a) filled in the study in a very short time ($<$ lower quartile of all minutes), (b) have a low average competence score ($<$ lower quartile of all competence scores by all annotators) and (c) have only assigned the correct function to one of the three control questions. With this heuristic 6 annotators were removed. To reflect subjectivity, we add two soft labels for UM. We calculate the probability for UM with standard normalization (based on the raw annotations) and with MACE. For the layers that were rated on a 5-point ordinal scale we aggregate by averaging. For sentiment and each moderation function we use the majority vote. For the layer 'agree with opinion' we report the number of annotators per label. To analyse disagreements on UM we calculate the normalized entropy for each item based on the raw annotations.

4 Dataset Overview: descriptive statistics

Tbl. 1 shows two example items with a high probability for UM. The upper one, however, has a perfect agreement and therefore a probability of 1.0 for UM, while the lower example has a high entropy with the probability being a bit lower (0.7). The upper example is phrased in a very polite tone with a neutral sentiment and some hedging, the speaker indicates that they are not certain about how the other person defines 'dangerous', asking for a clarification (but not necessarily saying that the person is wrong). In the lower example the user questions the meaningfulness of the parent statement and corrects misinformation about the complexity of the music genre. This is done in a quite aggressive tone ('your argument is basically absurd') but the comment is still rated as very constructive.

Probability of UM across candidate sets A total of 63% of the items of the non-aggregated version of the data has been identified as a form of moderation according to the respective annotator, indicating that the phenomenon is indeed common and that candidate selection was successful. Fig. 1 compares the distributions of the soft label (probability of UM) created with MACE between the different candidate subsets. We can see that all candidate subsets cover a wide range of probabilities, but that the median of the ChatGPT and the expert moderation model is significantly higher than the

into each batch of annotations to serve as a benchmark for annotators' assessments and to maintain the overall annotation quality.

Reply	Properties
How are you defining 'dangerous'? We can see scientifically it's not more physically dangerous than the other drugs, so what precisely do you mean? It's clearly not just that more acceptability == more danger, because there are drugs that are even more accepted (caffeine, aspirin) that are not more dangerous, so I think you need to clarify the 'danger' you're talking about.	usermoderation: 1.0 subjectivity: 2.5 aggressiveness: 1.89 sentiment: neutral constructiveness: 3.11 Functions: improve quality, broadening the discussion entropy: 0.0
I think country is a worthless genre of music, all you have to do is strum a few strings and sing a few words and you have a new "song". Your argument is basically as absurd as it sounds. I know because it is on the computer it seems that it would be easier to make, but there actually is a lot more skill to placing the sounds and arranging them in cool and pleasing ways. It is basically like any other music genre, except with techno you can work around with a whole bunch of different sounds and it really becomes complex. Maybe you're listening to bad techno. Nevertheless, the genre is worth its praise.	usermoderation: 0.7 subjectivity: 4.0 aggressiveness: 2.14 sentiment: negative constructiveness: 4.0 Functions: improve quality, content correction entropy: 0.86
Its ok to not go, the world will keep spinning. It seems that the main reason you don't want to go is the social aspect. Do you think you might have some levels of social anxiety? I myself have dealt with some social anxiety. Do you think that if it were not for the social aspects you would like going? Is there a part of you that wishes you could go? only to be overruled by the parts of you that is uncomfortable? If so. it may be beneficial for you to try to work on being more comfortable in social settings, especially if you have noticed a trend of not wanting to do these kinds of things because of social discomfort.	usermoderation: 0.9 subjectivity: 3.4 aggressiveness: 1 sentiment: positive constructiveness: 4.3 Functions: improve quality, broadening the discussion entropy: 0.86

Table 1: Examples from the dataset of UM.

Annotation layer	**Labels**
User Moderation	[y | n]
Moderation Function	8 possible functions, [y | n] for each
Constructiveness	[1-5 scale]
Sentiment / Tone	[positive | neutral | negative]
Subjectivity	[1-5 scale]
Aggressivity	[1-5 scale]
Agreement with comment	[yes | no | opinion not clear]
Describe the task	free text
Annotator perspective	[user | moderator]

Table 2: UMOD annotation layers: overview. All layers but the last two are at the item (comment) level. The last two layers are at the annotator level.

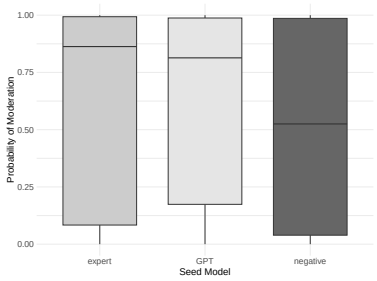


Figure 1: Distribution of soft label (MACE) for UM for seed model (median is marked).

one of the negatives. This suggests that the models used for candidate selection provide a good proxy for identifying potential candidates for UM, but also that the phenomenon is common enough to occur frequently, even in the negative sample.

What are frequent moderation functions of user moderators? We find that in general, user moderators engage more in moderator actions that target the content of comments, e.g. improve the quality, correct information or broaden the discussion (cf. Appendix Fig. 12 for a summary of each moderation function across the candidate sets). They operate less frequently on the meta-level (policing, helping with site issues). If we compare the amount of each function across candidate sets we can see that instructing ChatGPT with examples and explanations for each function helped to identify comments with less frequent actions, such as policing and helping with site issues. While for the other, more content-oriented functions and the social aspect of moderation the expert model candidates were selected more frequently, the ChatGPT-based candidates annotated contain a higher proportion of content correction.

What characterizes style and tone of user moderator comments? Comparing the distribution of ratings for constructiveness, subjectivity, and aggressiveness across different probability ranges for UM (cf. Fig. 2, aggregated dataset), the following pattern can be observed: UM comments are noticeably more constructive and exhibit reduced aggressiveness. For subjectivity, the trend is less pronounced, yet the tendency remains that user moderators' comments are less subjective.[10] UM comments of medium and high probability have a significantly lower proportion of negative sentiment. Conversely, the proportion of positive and neutral sentiment is higher in these ranges. This shows that the comments of the user moderators

[10]See Appendix Fig. 11 for a detailed look at sentiment, further supporting this trend.

follow similar characteristics expected by expert moderators, although our guidelines stated that UM does not need to conform to the neutrality and politeness typical of expert moderation.

However, in a qualitative inspection of the annotators definitions of UM, we found a prevalence towards neutrality and politeness as essential ingredients for UM. We can thus conclude that moderation is not solely understood as carrying out a specific function (e.g., asking for clarification), but annotators also consider tone and style.[11]

5 Analysis

We calculate the Krippendorff alpha for each annotation layer (average for the moderation functions).[12] Unsurprisingly given the phenomenon we are investigating, we observe a low agreement: all annotation layers are very subjective (especially constructiveness and subjectivity). Subjectivity (and the disagreement that it causes) is, however, a defining feature and not a bug of UM. Getting a better understanding the key properties of UM, and of the source of annotator disagreements is therefore the straightforward next step in our investigation.

In what follows, we employ regression analyses to address the following research questions: (a) How do annotators define constructiveness? and (b) What properties causes annotators to disagree about UM? We operate on the aggregated version of the dataset and for each item (i.e., an annotated reply comment) we predict constructiveness (1-5) as a dependent variable (DV) for (a), and the entropy of the UM annotation (disagreement) for (b). We use all other annotated properties (e.g. sentiment, aggressiveness) as independent variables (IV).

5.1 What defines constructiveness?

Defining what makes a comment constructive in argumentation or deliberation is not easy, despite growing interest in defining this notion (Napoles et al., 2017; Kolhatkar and Taboada, 2017a; Del Valle et al., 2020; Reveilhac, 2023). Here we aim to carve a clear notion of constructiveness based on the linguistic, stylistic, and pragmatic features that predict it.

Regression analysis We predict constructiveness as a dependent variable (DV), and use the other annotation layers in UMOD as predictors (IV). Additionally, we gather annotations with freely available tools, e.g. linguistic or textual complexity, emotion, and syntax. Specifically, we examine sentiment and emotions by identifying the amount of specific words from databases like the Geneva Affect Label Coder (GALC) or General Inquirer (GI). We assess linguistic or textual complexity through various metrics like type-token ratio variations or word frequencies. Additionally, we consider syntactic features such as the frequency of verbs and the usage of 1st and 2nd person pronouns.[13]

In total, our regression model contains 209 features, 4 from the UMOD annotated layers, and 205 linguistic/pragmatic/stylistic as described above. The next step is to perform model selection, i.e., to identify the most explanatory regression model (subset of the candidate features). To do so, we start with a simple model and perform a step-wise increase in complexity, selecting IV terms that improve the fit significantly.[14] We measure model fit in terms of explained variance (R^2). The explained variance of each predictor (e.g., sentiment), in turn, quantifies the strength of its impact in predicting the modulation of the dependent variable (e.g., in this case, constructiveness). The final model selected consists of 79 features.[15]

Results The most influential factor, explaining 20.8% of the variance in the model, is the number of words – lengthier posts tend to demonstrate higher levels of constructiveness (cf. Fig. 3). This underscores the idea that comments offering greater explanation and information are generally more constructive compared to shorter, less detailed comments. The strong predictive influence of more informative comments, those with numbers (e.g. statistics, dates, etc), .com links, and mentions of affiliations, further supports this finding. Additionally, while sufficient length is important, the choice of language should be familiar (high HDD42 AW in Fig. 3), i.e. use frequent words (high KF FREQ AQ LOG), which in turn correspond to a higher proportion of frequent trigrams.

Sentiment, accounting for 17% of the explained variance, plays a noteworthy role in indicating constructiveness. Comments that come across as more positive tend to exhibit higher levels of constructiveness. Additionally, we observe that comments

[11] A more detailed analysis and discussion of the annotators definitions on UM can be found in Appendix D.

[12] See Appendix Tbl. 7 for values.

[13] A high-level overview of the feature categories is in Appendix Tbl. 8, a detailed description is in the repository.

[14] Implemented with the standard stepAIC package in *R*.

[15] A breakdown of explained variance for features of the selected model is reported in Appendix Tbl. 9.

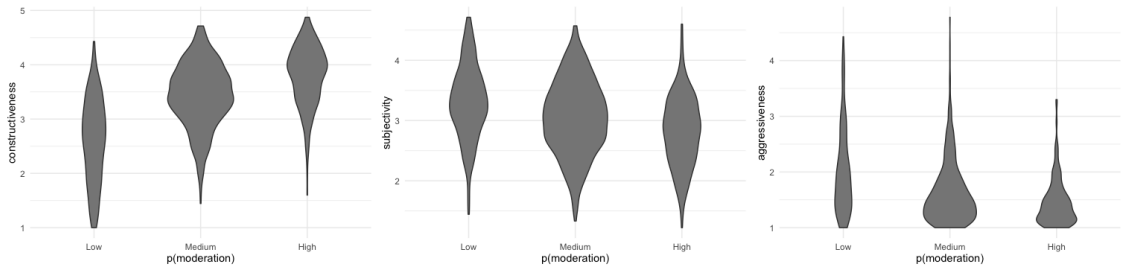


Figure 2: Violin Plots of the ordinal scores for aggressiveness, constructiveness and subjectivity compared across different ranges of probabilities for UM.

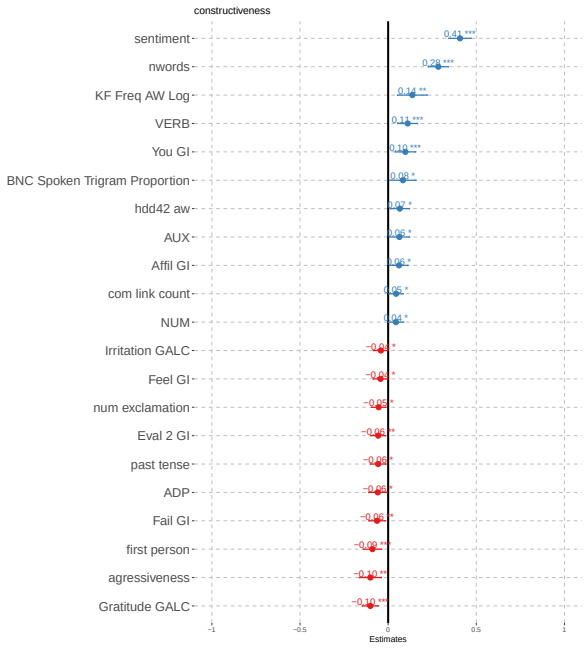


Figure 3: **Constructiveness.** Standardized beta values of selected terms after model selection. Showing how strongly each feature affects the average constructiveness ($R^2 = 61\%$): forest plot.

displaying elevated levels of aggressiveness, irritation, use of exclamation marks, or variables linked to negative contexts or judgments (FAIL GI, EVAL 2 GI) are all notable predictors for lower constructiveness scores. The use of first person pronouns, frequent use of words describing feelings (FEEL GI), or words associated to gratitude all have a negative impact on constructiveness. This is likely caused by the self-referential and emotional impact such features produce, leading to less neutral and unbiased comments. While comments with second person pronouns (YOU GI) likely encourage an interactive, and thus more constructive, quality (see, for example, the third example in Tbl. 1).

5.2 What causes annotators to disagree?

As discussed before, the assessment of whether a comment represents an instance of UM is a subjective task, leading to annotator disagreements. Here, we conduct a statistical analysis to "mine" such disagreements. We hypothesize that if the style deviates from expert moderation, annotators will disagree more. Beyond linguistic features which broadly speaking represent the style of a comment, we consider two additional features which lend themselves to a better description of the patterns of disagreement: the number of different moderation functions assigned to an item (if there are many different possibilities the item is harder to interpret or its intention is less clear which could cause higher disagreement) and the discrepancy between the individual conceptualizations of the task that annotators built while reading the guidelines and performing the task. We model the latter as the average semantic similarity between the definitions of UM given by the annotators of a specific item. If annotators have a (semantically) similar definition we expect them to agree more.

Regression analysis Our regression model takes the entropy of the prediction of UM as a dependent variable. A high entropy indicates high disagreement. As independent variables, we consider subjectivity, constructiveness, sentiment, definition similarity, the number of different functions and all their pair-wise interactions. The model selection procedure is the same as in the analysis in Sect. 5.1.

Results The final model explains 34% of the variance. Its most explanatory IVs are subjectivity (8%) and three interaction terms: subjectivity and constructiveness (6%), constructiveness and sentiment (5%) and constructiveness and the number of different functions (4%). As expected, we can observe a positive effect of subjectivity on disagreement and a negative effect of constructiveness (see

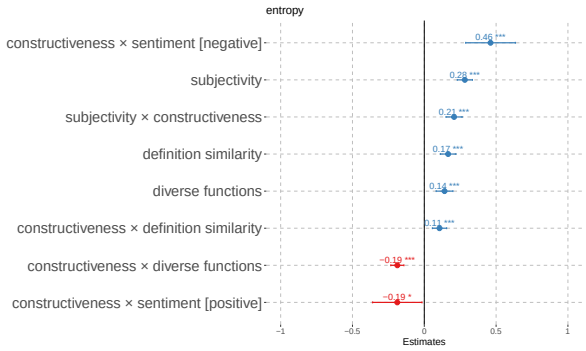
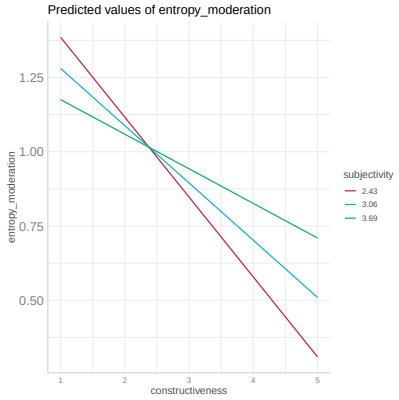
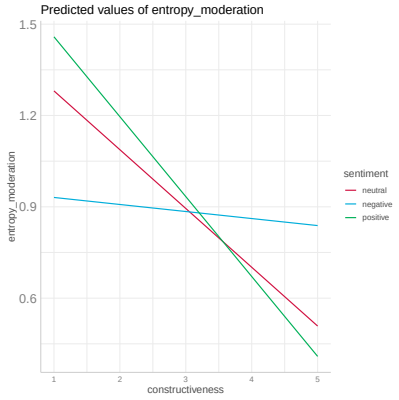


Figure 4: **Disagreement.** Standardized beta values of selected terms. Showing how strongly each feature affects the disagreement ($R^2 = 34\%$): forest plot



(a) Interaction: constructiveness & subjectivity



(b) Interaction: constructiveness & sentiment

Figure 5: Disagreement prediction: marginalized effect of interaction terms

positive and negative effect of these variables in the forest plot in Fig. 4). Highly subjective items therefore cause annotators to disagree on whether it can be perceived as UM. When a reply is more constructive, there tends to be less disagreement. However, the effect of reduced disagreement for more constructive comments is weakened when the comments are highly subjective. This can be seen in the visualization of the interaction terms (cf. Fig. 5(a)): the slopes of the lines representing the effect of constructiveness on disagreement vary. The line illustrating the highest level of subjectivity shows a noticeably weaker decline.

A similar pattern can be observed with sentiment: annotators are more likely to agree on constructive comments with a neutral or positive sentiment (red and green line show a steep slope, Fig. 5(b)) but a negative sentiment mitigates the effect of constructiveness (weak slope of the blue line). Surprisingly we can observe a positive effect of the semantic similarity between definitions – a higher similarity leads to higher disagreement (Fig. 4). When we look at how constructiveness, subjectivity, and semantic similarity interact, we find that these factors have a stronger impact on annotators who share a similar definition of UM. One possible explanation could be that this group of annotators has a more nuanced interpretation of UM, while others have a more general perspective. As a result, even small differences in language use lead to greater variations in their annotations.

6 Conclusion

This work is the first to introduce and study the concept of UM. We released UMOD, a dataset of 1000 online comments annotated by a large pool of crowdsourcers for different aspects of UM (e.g., tone, style, but also annotator perspectives on the task). Our analysis shows that the further the language deviates from this professional standard, the more controversial the perception of its validity as moderation becomes and that constructiveness is characterized by politeness, positive sentiment, and appropriate language complexity and length; it also involves personally addressing others without being excessively self-focused. We believe that UMOD will significantly contribute to research on semi-automatic content moderation. Additionally, UMOD bridges the gap between theories of effective moderation as implemented in moderation guidelines and the needs of forum users.

Future work could apply this annotation schema to other platforms, enabling a comparison of UM characteristics across diverse domains, languages and cultures. This allows for novel research directions, such as exploring research questions related to the factors that promote various types of UM behaviors across those various platforms.

7 Limitations

Due to budgetary constraints and limitations inherent in the annotation platform used, achieving a broad diversity in annotator perspectives is challenging. Platforms like Prolific do not provide options for a fine-grained the pre-screening of annotator pools. Currently, the demographic statistics of annotators, especially for countries outside the U.S., are not available in Prolific. This makes it difficult to ensure a balanced and representative sample.

The dataset is exclusively in English. While this focus allows for a depth of analysis within the English-speaking and -writing population, it restricts the dataset's utility for studies aiming at linguistic diversity and cross-language analyses.

All data has been sourced from a single domain - Reddit. Other platforms e.g. Facebook and X often have issues with both data availability and content moderation dynamics that are not transparent. It is also worth noting that Reddit is one of the most prominent discussion platforms, it is highly argumentative and deliberative in nature, and covers a large variety of discussion topics. This diversity allows us to analyze how certain patterns persist across various discussion topics. We selected the subreddit of changemyview because of its availability and the fact that participants actively argue and deliberate about a variety of different topics. Although Reddit is known for its diverse array of topics and discussions, this limitation may affect the generalizability of the dataset to other online platforms or domains. While the dataset is limited to one domain, the resource itself consists of a large range of discussion topics and language variability.

Reddit is still predominantly used by a specific demographic group, mainly younger, white males. However, a more precise analysis is possible since the comments in our dataset are derived from an existing collection which to date is a standard dataset in computational argumentation research. Numerous works have contributed annotation layers or developed tools tailored to it. Each comment is tagged with a unique identifier, allowing it to be easily matched with its corresponding metadata for a detailed examination. This process enables a deeper understanding of the dataset's composition and diversity, despite the known demographic tendencies of the Reddit user base.

The current work does not conduct a direct comparison with expert moderation in terms of annotated or linguistic features. This was beyond our study's scope, but can reveal further important insights about the difference in expert and UM and their relationship to constructive and productive discourse.

8 Ethics

Revisiting the ethical considerations associated with this work, it's crucial to note that our dataset comprises interactions sourced from Reddit. As previously stated, this platform is characterized by a distinct user demographic, rendering it unrepresentative of the broader society. Consequently, models trained on this dataset might inherit and amplify the existing biases.

However, it is important to emphasize that the focus of our dataset is on proactive moderation behavior, not exclusively on identifying and censoring problematic content. This nuanced focus mitigates the potential reinforcement of bias to some degree compared to models that focus strictly on enforcing civility.

Furthermore, we recognize that the language used on Reddit, and some of the topics within the "changemyview" discussions, may be triggering for some individuals. In response to this, we've implemented an option for all annotators to skip particular instances they find uncomfortable and to label them as disturbing. We will incorporate this information into the final dataset. This will enable immediate filtering, allowing future users and researchers to focus on instances that haven't been flagged as problematic.

Acknowledgements

We acknowledge funding by the Bundesministerium für Bildung und Forschung (BMBF) through the project E-DELIB. We would also like to thank ZebraLog for their contribution of data.

References

Federico Cabitza, Andrea Campagner, and Valerio Basile. 2023. Toward a perspectivist turn in ground truthing for predictive computing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(6):6860–6868.

Eshwar Chandrasekharan, Shagun Jhaver, Amy Bruckman, and Eric Gilbert. 2022. Quarantined! examining the effects of a community-wide moderation intervention on reddit. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 29(4):1–26.

Jonathan P Chang and Cristian Danescu-Niculescu-Mizil. 2019. Trouble on the horizon: Forecasting the derailment of online conversations as they develop. 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 4743–4754.

Marc Esteve Del Valle, Rimmert Sijtsma, Hanne Stegeman, and Rosa Borge. 2020. Online deliberation and the public sphere: Developing a coding manual to assess deliberation in twitter political networks. *Javnost-The Public*, 27(3):211–229.

Cornell eRulemaking Initiative et al. 2017. Ceri (cornell e-rulemaking) moderator protocol. *Cornell e-Rulemaking Initiative Publications*, 21.

Katharina Esau, Dennis Friess, and Christiane Eilders. 2017. Design matters! an empirical analysis of online deliberation on different news platforms. *Policy & Internet*, 9(3):321–342.

Neele Falk, Iman Jundi, Eva Maria Vecchi, and Gabriella Lapesa. 2021. Predicting moderation of deliberative arguments: Is argument quality the key? In *Proceedings of the 8th Workshop on Argument Mining*, pages 133–141, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Neele Falk and Gabriella Lapesa. 2023. Bridging argument quality and deliberative quality annotations with adapters. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2469–2488, Dubrovnik, Croatia. Association for Computational Linguistics.

Lodewijk Gelauff, Liubov Nikolenko, Sukolsak Sakshuwong, James Fishkin, Ashish Goel, Kamesh Munagala, and Alice Siu. 2023. Achieving parity with human moderators: A self-moderating platform for online deliberation 1. In *The Routledge Handbook of Collective Intelligence for Democracy and Governance*, pages 202–221. Routledge.

Danula Hettiachchi and Jorge Goncalves. 2019. Towards effective crowd-powered online content moderation. In *Proceedings of the 31st Australian Conference on Human-Computer-Interaction*, pages 342–346.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 2013. Learning whom to trust with MACE. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1120–1130, Atlanta, Georgia. Association for Computational Linguistics.

Varada Kolhatkar and Maite Taboada. 2017a. Constructive language in news comments. In *Proceedings of the First Workshop on Abusive Language Online*, pages 11–17, Vancouver, BC, Canada. Association for Computational Linguistics.

Varada Kolhatkar and Maite Taboada. 2017b. Using New York Times picks to identify constructive comments. In *Proceedings of the 2017 EMNLP Workshop: Natural Language Processing meets Journalism*, pages 100–105, Copenhagen, Denmark. Association for Computational Linguistics.

Metka Kuhar, Matej Krmelj, and Gregor Petrič. 2019. The impact of facilitation on the quality of deliberation and attitude change. *Small Group Research*, 50(5):623–653.

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.

Sanna Malinen. 2022. Moderation of deliberation: How volunteer moderators shape political discussion in facebook groups? In *International Conference on Human-Computer Interaction*, pages 602–616. Springer.

Angelina McMillan-Major, Amandalynne Paullada, and Yacine Jernite. 2022. An interactive exploratory tool for the task of hate speech detection. In *Proceedings of the Second Workshop on Bridging Human–Computer Interaction and Natural Language Processing*, pages 11–20, Seattle, Washington. Association for Computational Linguistics.

Courtney Napoles, Joel Tetreault, Aasish Pappu, Enrica Rosato, and Brian Provenzale. 2017. Finding good conversations online: The yahoo news annotated comments corpus. In *Proceedings of the 11th linguistic annotation workshop*, pages 13–23.

Chan Young Park, Julia Mendelsohn, Karthik Radhakrishnan, Kinjal Jain, Tushar Kanakagiri, David Jurgens, and Yulia Tsvetkov. 2021. Detecting community sensitive norm violations in online conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3386–3397, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Joonsuk Park, Sally Klingel, Claire Cardie, Mary Newhart, Cynthia Farina, and Joan-Josep Vallbé. 2012. Facilitative moderation for online participation in eRulemaking. In *Proceedings of the 13th Annual International Conference on Digital Government Research*. ACM.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Maud Reveilhac. 2023. Comparing and mapping difference indices of debate quality on twitter. *Methodological Innovations*, page 20597991231180531.

Pratik Sachdeva, Renata Barreto, Geoff Bacon, Alexander Sahn, Claudia von Vacano, and Chris Kennedy. 2022. The measuring hate speech corpus: Leveraging rasch measurement theory for data perspectivism. In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022*, pages 83–94, Marseille, France. European Language Resources Association.

Qinlan Shen and Carolyn Rose. 2019. The discourse of online content moderation: Investigating polarized user responses to changes in Reddit's quarantine policy. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 58–69, Florence, Italy. Association for Computational Linguistics.

Tanner Skousen, Hani Safadi, Colleen Young, Elena Karahanna, Sami Safadi, and Fouad Chebib. 2020. Successful moderation in online patient communities: inductive case study. *Journal of medical Internet research*, 22(3):e15983.

Kumar Bhargav Srinivasan, Cristian Danescu-Niculescu-Mizil, Lillian Lee, and Chenhao Tan. 2019. Content removal as a moderation strategy: Compliance and other outcomes in the changemyview community. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–21.

Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2016. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of the 25th international conference on world wide web*, pages 613–624.

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

Timon Ziegenbein, Shahbaz Syed, Felix Lange, Martin Potthast, and Henning Wachsmuth. 2023. Modeling appropriate language in argumentation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4344–4363, Toronto, Canada. Association for Computational Linguistics.

A Details Annotation Study

A.1 Models for selecting candidates

We use a `roberta-large` model for training on each source pool: *expert moderation dataset* and *ChatGPT dataset*. The ChatGPT model was trained for 5 epochs with a learning rate of 2e-5 and a batch size of 16 (we used 3 GPUs (NVIDIA RTX A6000, each GPU has 49GB, CUDA Version 11.7). The same parameters applied to the expert model with a higher number of epochs (10). The ChatGPT model achieved a perfect F1-score on the development set, the expert model an F1-score of 0.94.

The ChatGPT model identified 105,807 comments as moderator comments (total of 9.5 percent) of the whole ChangeMyView dataset, and the expert model 392,226 comments (35 percent).

A.2 Guidelines

Fig. 6 provides the full guidelines for the annotation study, as presented to the annotators.

We collected the ratings with Google Forms via Prolific. Before annotators started with the study, they were informed about the main content of the study, risks and benefits, approximate estimated time and were asked for their consent. We also added a trigger warning and guidelines how to skip an instance and flag it as problematic. The consent form is depicted in Fig. 7. Tbl. 2 in the main text summarizes all 7 annotation layers and the respective labels. Figures 8, 9 and 10 depict an example as it was shown to the annotators in Google Forms and the interface which they used for annotation.

Annotators could to justify and comment on their annotation in a free-text field.

Moderation Function	**Description**
Broadening Discussion	The reply encourages users to consider and engage comments of other users; or it promotes a more expansive or broader discussion on the topic by the author of the OP or the community
Improving Comment Quality	The reply asks for more information, factual details, or data to be provided to support the statements made; or asks the author of the OP to make or consider possible solutions or alternative approaches.
Content Correction	The reply provides substantive information about the parent comment; corrects misstatements or clarifies details about the OP/parent comment; or points to relevant information such as websites or specific documents with the goal of correcting the content of the parent comment.
Keeping Discussion on Topic	The reply explains why the parent comment is beyond the authority or competence of the platform, or outside the scope of the discussion; or it indicates irrelevant, off-point statements.
Organizing Discussion	The reply directs the author of the parent comment to another post or comment that is more relevant to their expressed interest.
Policing	The reply aims to maintain/encourage civil deliberative discourse; or it points to inappropriate language or content in the parent comment.
Resolving Site Use Issues	The reply is to resolve technical difficulties; or it provides information about the goals/rules of the platform.
Social Functions	The reply takes on the function of welcoming/greeting, encouragement or appreciation of the parent comment, or thanking for participation.

Table 3: Different Moderation functions and their description.

Candidate	**Moderation Function**
It's important to recognize the interconnectedness of social issues. How might this issue intersect with other social issues, such as race, gender, or sexuality?	Broadening Discussion
Let's try to stay focused on the topic at hand and avoid getting sidetracked by personal attacks or unrelated issues.	Policing
I'm not sure how your comment is related to the original post. Can you please clarify how your perspective is relevant to the current discussion?	Keeping Discussion on Topic
This post from last week might be of interest to you: https://www.reddit.com/r/changemyview/comments/example_post	Organizing Discussion
Your argument rests on a number of implicit assumptions that I'm not sure are accurate. For example, you seem to be assuming that all people have equal access to resources and opportunities, which is not necessarily the case. Can you speak to these assumptions and provide evidence to back them up?	Improving Comment Quality

Table 4: Examples of candidate instances extracted using the *ChatGPT moderation model*.

A.3 Control Instances

Tbl. 5 shows the three control instances added to each batch of the annotation study. We expected annotators to mark these as a form of moderation with a high probability and we captured three different functions.

Guidelines for User Moderation Annotation

User Moderation Annotation: Goal

Moderation in most platforms relies generally on *expert moderators*, who are trained specifically for the role and whose contribution in the platform most often is specifically that of moderation. However, the role of moderation in deliberation and argumentation platforms can often be seen in general *user comments*; and the impact or contribution of that comment to the discussion is in line with that of a moderator. The goal of this study is to annotate *user comments* that align with the characteristics of [expert] moderation within a discussion or argument.

What is Moderation?[1]

The goal of *moderation* in deliberation and argumentation platforms is to create an environment of informed and thoughtful participation, as well as mentor effective commenting behavior.

A moderator moves participants past "voting and venting" behaviors to effectively contributing the information they possess. They also make participants feel that their voices have been heard and that they are part of a forum for [civil] engagement.

Moderators have the role of advocating for the commenting process; as they encourage a "knowledge building community" that supports commenters' access to, participation in, and learning about the process and topic under discussion. Whether the goal of the process is policymaking, converging perspectives, or arguing one's view, moderation helps commenters to contribute as individuals as well as collaborate with each other.

Expectation of Moderators

1. Neutrality: Expert moderators are strongly encouraged to remain *neutral*, avoiding taking a position on the substance of the discussion, or forming biases or making assumptions about participants' comments. However, users are not restricted to this requirement and comments that do indeed have the role of moderation from a user may (e.g. in the case of clarification comments) or may not (e.g. signaling erred information to another user) have this characteristic.
2. Maintaining the norms: Expert moderators are responsible for maintaining the norms of the platform community and its regulations. Users might mirror this role in subtle ways, such as reminding others of the goal of the discussion or pointing out inappropriate contributions.
3. Choice of wording: Expert moderators are asked to use plain language, calm tones, avoid condescending responses, and limit the number of questions. For example:
    a. *That clarification is available in several forms on the website http:[…]*
    b. *DOT has estimated that the benefits of this discussion will outweigh the costs.*
    c. *This is an interesting suggestion, thanks. Could you provide a little more information on this, and perhaps a link.*

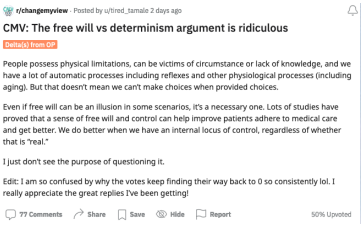
    Again, users are not expected to uphold these standards in their comments, however they may still perform similar contributions to the discussion, with or without a careful choice of wording.

[1] Moderation overview is adapted from the Moderator Protocol of RegulationRoom.com

The Data: Change My View

The data you will be annotating is extracted from the online subreddit entitled Change My View.[2] The platform is dedicated to civil discourse, aimed at promoting productive conversation to resolve differences by understanding others' perspectives.

The format of CMV is as follows. First, a user (original poster, or OP) posts a *view*, defined as a particular way of considering or regarding something, an attitude or opinion, on a specified *topic* issue, and asks the community to "change my view". For example:



Users are then able to interact with the OP as comments to argue their perspective in order to change the OP author's view. The interaction between users and OP author may be a simple back-and-forth comment, or may be an extended discussion. At the end of the interaction, if the user's argument has successfully changed the OP's view, the user is awarded a *Delta* (Δ) by the OP author.

Annotation Task

The annotator will be shown two texts: the ***preceding comment*** (for example, the OP or a post in the comment thread) and the ***reply comment***. The preceding comment as well as the topic of the OP are provided to the annotator to offer context. The *reply comment* is the comment to which the annotation questions refer. For each reply comment, the annotators are asked a set of questions, described in detail below:

1. **User Moderation** [y/n]
    Do you consider this user comment to behave as a form of moderation in the discussion?

[2] https://www.reddit.com/r/changemyview/

2. **Moderation Function**[3]
    In the case that the user comment behaves as a form of moderation, please provide information on the type of *moderation function* the comment performs. Please select the most appropriate function(s), understanding that the language use of users may lead to more flexibility and interpretation of the definitions of these moderation functions. After selecting the relevant functions, the annotators may provide additional comments or justification for their selection as a short answer.
    a. **Broadening Discussion**. [y/n] The comment encourages users to consider and engage comment of other users; or it promotes a more expansive or broader discussion on the topic by the author of the preceding comment or the community.
    b. **Improving Comment Quality**. [y/n] The comment asks for more information, factual details, or data to be provided to support the statements made; or asks the author of the preceding comment to make or consider possible solutions or alternative approaches.
    c. **Content Correction**. [y/n] The user comment provides substantive information about the preceding comment; corrects misstatements or clarifies details about the preceding comment; or points to relevant information such as websites or specific documents with the goal of correcting the content of the preceding comment.
    d. **Keeping Discussion on Topic**. [y/n] The user comment explains why the preceding comment is beyond the authority or competence of the platform, or outside the scope of the discussion; or it indicates irrelevant, off-point statements.
    e. **Organizing Discussion**. [y/n] The comment directs the author of the preceding comment to another post or comment that is more relevant to their expressed interest.
    f. **Policing**. [y/n] The comment aims to maintain/encourage civil deliberative discourse; or it points to inappropriate language or content in the preceding comment.
    g. **Resolving Site Use Issues**. [y/n] The comment is to resolve technical difficulties; or it provides information about the goals/rules of the platform.
    h. **Social Functions.** [y/n] The user comment takes on the function of welcoming/greeting, encouragement or appreciation of the preceding comment, or thanking for participation.

3. **Justification** (Optional)
    You can provide a short justification or any details you would like to offer for your answers to questions (1) and (2). Please note, there is a limit of 225 characters for this answer.

4. **Constructiveness** [1-5 scale]
    Considering the user comment in general – whether or not it behaves as a form of moderation – do you consider this comment to be constructive to the discussion?
    Constructive comments can be defined as *high-quality comments that make a contribution to the conversation*. Such comments are considered to offer an opinion or perspective, and provide support, reasoning, or background for that view. They are characterized as comments that intend to create a civil dialogue through remarks that are relevant to the discussion/topic and not intended to merely provoke an emotional response.

5. **Sentiment / Tone** [positive | neutral | negative]
    How would you evaluate the overall tone of the user comment? Would you consider the underlying feeling, attitude, evaluation, or emotion associated to the comment as positive, negative, or neutral?

6. **Subjectivity** [1-5 scale]
    Does the user comment refer to the user's personal opinions or feelings regarding a particular subject matter, based on their unique interpretation of an idea or their own thoughts, feelings, and background; or is the comment rather neutral in this respect?

7. **Aggressivity** [1-5 scale]
    Do you consider the user comment to be aggressive, actively or passively? Examples could include (but are not limited to) sarcasm, blaming, intimidation, threats, or attacks.

8. **Agreement with comment opinion** [yes | no | opinion not clear]
    Do you agree with the opinion expressed in the reply comment?

[3] Taken from Moderator Roles and Interventions (Park et al., 2012)

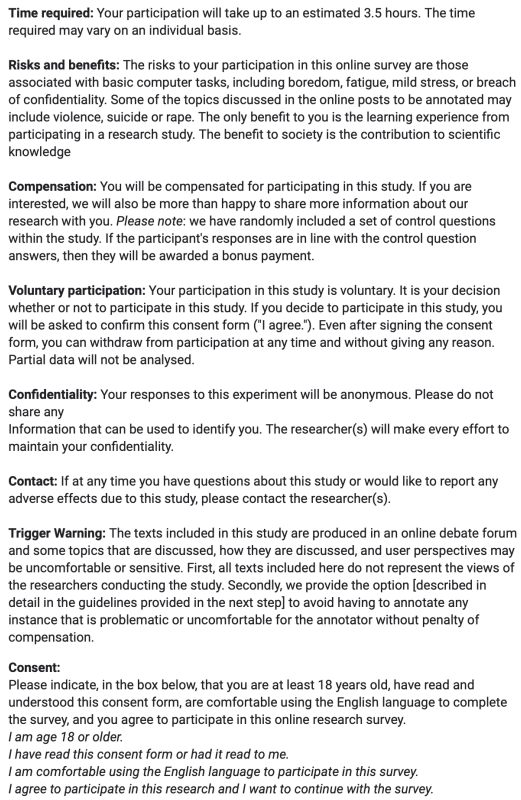
Trigger Warning!

As mentioned in the consent form you agreed to, the texts included in this study are produced in an online debate forum and some topics that are discussed, how they are discussed, and user perspectives may be uncomfortable or sensitive. First, all texts included do not represent the views of the researchers conducting the study. Secondly, we provide the option to avoid having to annotate any instance that is problematic or uncomfortable for the annotator without penalty of compensation.

To do so, please answer the annotation questions as outlined below. Note, although you will have provided answers, if you include the following text in the Justification, your answers to this instance will be automatically discarded and not considered in the study.

1. User Moderation: *No*
2. Moderation Function: *None of these*
3. Justification: (please copy and paste) *I am uncomfortable annotating this text and voluntarily skip this instance.*
4. Constructiveness: *No*
5. Sentiment: *Neutral*
6. Subjectivity: *Neutral*
7. Aggressiveness: *Neutral*
8. Do you agree with opinion: *Opinion not clear*

Figure 6: Annotation Guidelines.



Time required: Your participation will take up to an estimated 3.5 hours. The time required may vary on an individual basis.

Risks and benefits: The risks to your participation in this online survey are those associated with basic computer tasks, including boredom, fatigue, mild stress, or breach of confidentiality. Some of the topics discussed in the online posts to be annotated may include violence, suicide or rape. The only benefit to you is the learning experience from participating in a research study. The benefit to society is the contribution to scientific knowledge

Compensation: You will be compensated for participating in this study. If you are interested, we will also be more than happy to share more information about our research with you. *Please note*: we have randomly included a set of control questions within the study. If the participant's responses are in line with the control question answers, then they will be awarded a bonus payment.

Voluntary participation: Your participation in this study is voluntary. It is your decision whether or not to participate in this study. If you decide to participate in this study, you will be asked to confirm this consent form ("I agree."). Even after signing the consent form, you can withdraw from participation at any time and without giving any reason. Partial data will not be analysed.

Confidentiality: Your responses to this experiment will be anonymous. Please do not share any
Information that can be used to identify you. The researcher(s) will make every effort to maintain your confidentiality.

Contact: If at any time you have questions about this study or would like to report any adverse effects due to this study, please contact the researcher(s).

Trigger Warning: The texts included in this study are produced in an online debate forum and some topics that are discussed, how they are discussed, and user perspectives may be uncomfortable or sensitive. First, all texts included here do not represent the views of the researchers conducting the study. Secondly, we provide the option [described in detail in the guidelines provided in the next step] to avoid having to annotate any instance that is problematic or uncomfortable for the annotator without penalty of compensation.

Consent:
Please indicate, in the box below, that you are at least 18 years old, have read and understood this consent form, are comfortable using the English language to complete the survey, and you agree to participate in this online research survey.
I am age 18 or older.
I have read this consent form or had it read to me.
I am comfortable using the English language to participate in this survey.
I agree to participate in this research and I want to continue with the survey.

Figure 7: Consent Form to be filled in by the annotators.

A.4 Socio-demographic variables

Tbl. 6 shows the socio-demographic variables of the pool of the final annotators (after filtering). Competence refers to the average competence score computed with MACE. Annotation time is reported in minutes. We asked annotators for feedback on which roles they identified with when analyzing the data. A clear majority assigned themselves to the perspective of a moderator. In contrast, a smaller proportion identified with the user's point of view and an even smaller proportion could not clearly identify with one of the given roles.

A.5 Quantitative analysis: plots

Fig. 12 displays the relative frequency of the different moderation functions in the aggregated dataset.

Fig. 11 shows the relative amount of positive, negative and neutral sentiment across three probability bins for moderation. The trend shows that highly probably moderation comments are significantly more positive or neutral and non-moderation comments more negative.

A.6 Agreement

Tbl. 7 reports the Krippendorff alpha for each annotation layer.

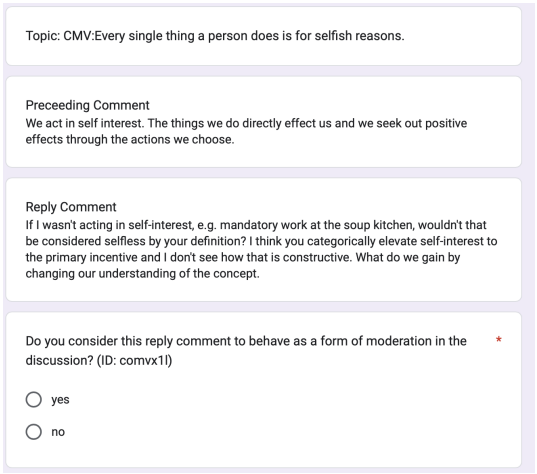


Figure 8: Example of comment-reply pair in the annotation form.

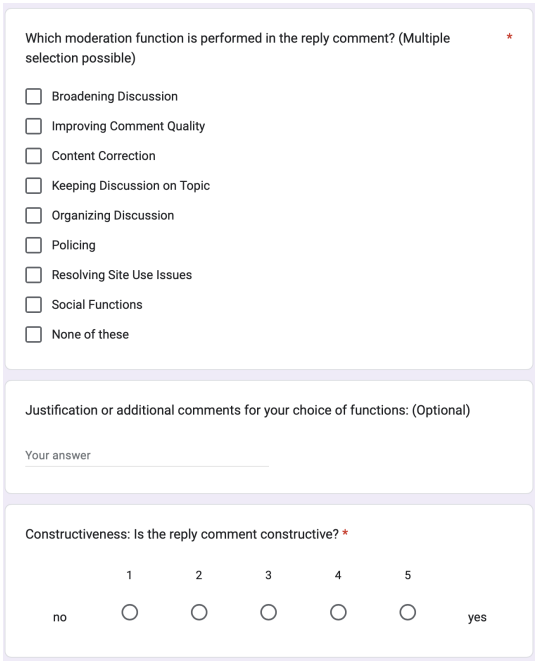


Figure 9: Annotation of moderator functions and constructiveness.

Title / Topic	**parent comment**	**reply**	**function**
Religious equality is irrational.	Religious freedom is just an excuse for discrimination and if you believe in it, then you're ignorant and intolerant. It's just a way for people to justify their bigotry and hate towards marginalized groups. Religion has caused more harm than good in history, and it's time we stop giving it a free pass. Wake up and join the 21st century already.	Please stick to healthy discussions while remaining respectful and avoiding personal attacks.	policing
The bank bailouts took advantage of taxpayer money.	The bank bailouts were nothing more than a blatant misuse of taxpayer money. Instead of holding financial institutions accountable for their reckless behavior, the government rewarded them with bailouts, allowing them to escape the consequences of their actions.	That's an interesting perspective, could you provide some evidence or support for that claim?	improve quality
Private healthcare just works better.	Privatized healthcare systems are inherently more efficient and cost-effective compared to publicly funded healthcare systems. They promote competition, innovation, and personalized care, ultimately benefiting the patients.	That information was proven wrong in a recent study looking into exactly this [https://www.ncbi.nlm.nih.gov/pmc/articles/PMC]. Double check your claims before posting.	content correction

Table 5: Three examples for different functions of UM. These examples were used as control questions during the study.

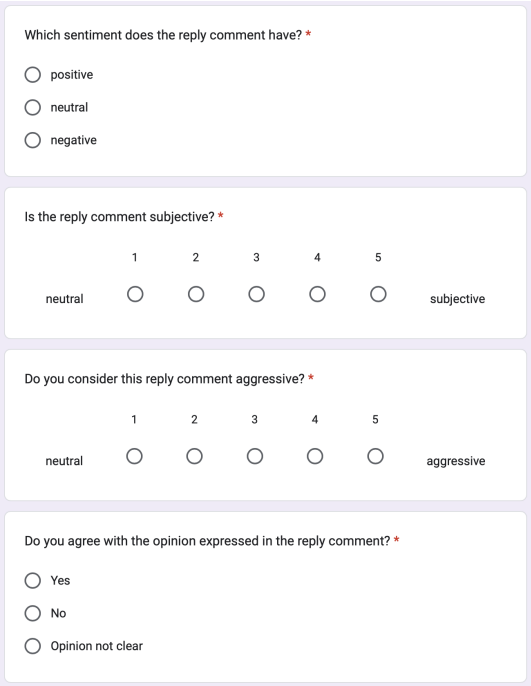


Figure 10: Annotation of evaluative features and annotator opinion

annotator feature	mean value with std or distribution
competence	0.34 ± 0.15
age	32 ± 11
annotation time	185 ± 81
sex	female: 47, male: 34
race	white: 59, asian: 8, mixed: 7, black: 5, other: 2
role	moderator: 59, user: 15, none of the two: 10

Table 6: Summary of socio-demographic variables of the annotators of our study.

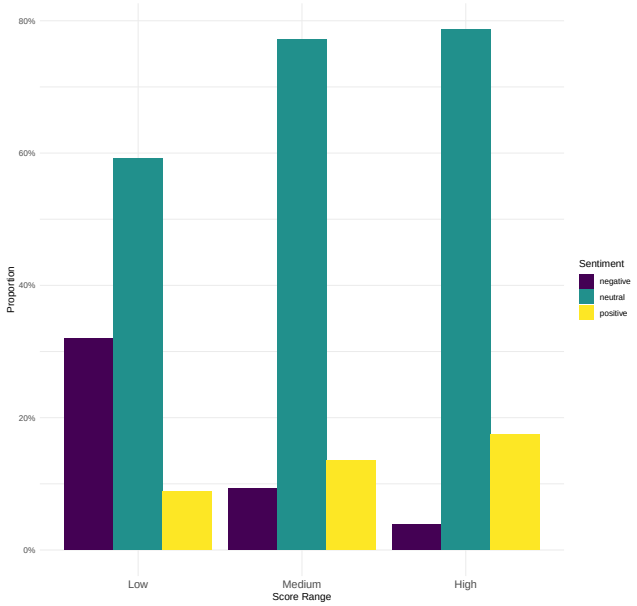


Figure 11: Sentiment across low, medium and high probability for UM.

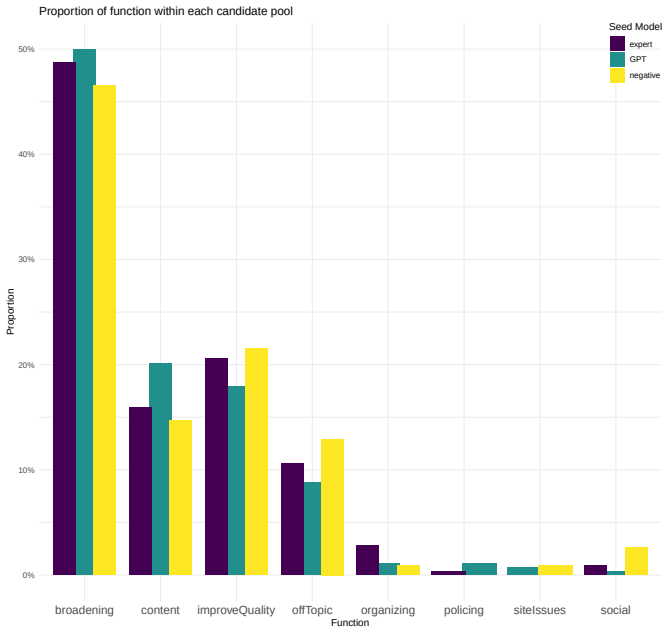


Figure 12: Moderation function (aggregated dataset): rel. frequency for candidate sets from different seed models

Annot. Layer	Krippendorff alpha
user moderation	0.12
moderation function	0.06
sentiment	0.14
constructiveness	0.07
agressiveness	0.10
subjectivity	0.04

Table 7: Krippendorff alpha for all annotation layers.

B Regression analysis on constructiveness: additional materials

feature name	explanation	type
BNC Spoken Trigram Normed Freq	proportion of frequent trigrams normed	complexity
BNC Spoken Trigram Proportion	proportion of frequent trigram	complexity
KF Freq AW Log	mean word frequency / lexical complexity	complexity
hdd42 aw	score for Vocabulary frequency / familiarity: for each word type, compute the probability of encountering one of it's tokens in a random sample of 42 tokens, same range as type token ratio	complexity
OG N H FW	linguistic complexity: Number of phonographic neighbors; i.e., words differing in one letter and one phoneme (e.g., stove and stone); different from orthographic neighbors, which are formed by substituting one letter w/ another (e.g., stove and shove); includes homophones	complexity
Ortho N	Number of orthographic neighbors	complexity
Freq N P FW	linguistic complexity: Ave freq (Freq_HAL) of phonological neighborhood; excludes homophones	complexity
BNC Spoken Bigram Normed Freq Log		complexity
lexical density types	ratio of different types (basically some form of TTR), lexical density / complexity	complexity
BNC Spoken Bigram Proportion	proportion of frequent bigrams	complexity
Decreas GI	Decrease: 82 words, Quality and quantity, e.g. abate, alleviate, amputate, atrophy, cheapen	emotion
Fail GI	Fail: 137 words indicating that goals have not been achieved, negative emotion words e.g. abandon, abandonment, absence, absent, absent-minded	emotion
Longing GALC	arousal, e.g. crav*, daydream*, desir*, fanta*, hanker*	emotion
Eval 2 GI	Evaluation: 205 words which imply judgment and evaluation, whether positive or negative, including means-ends judgments	emotion
Feel GI	Feel: 49 words describing particular feelings, including gratitude, apathy, and optimism, not those of pain or pleasure	emotion
Irritation GALC	negative emotion words realted to irritation: annoy*, exasperat*, grump*, indign*, irrita*	emotion
Disappointment GALC	amount of words expressing disappointment: comedown,disappoint*,discontent*,disenchant*,disgruntl*, disillusion*,frustrat*, jilt*, letdown,resign*, sour*, thwart*	emotion
Natobj GI	Natural Objects: 61 words for natural objects including plants, minerals and other objects occurring in nature other than people or animals, e.g. ash, atom, atomic, bed, boulder	lexical
You GI	You: 9 pronouns indicating another person is being addressed directly, e.g. thee, thou, thy, you, you	lexical
Rcethic Lasswell	Ethics: 151 words of values concerning the social order., e.g. adhere, adherence, appall, appall, betray	lexical
Affil GI	Affiliation: 557 words indicating affiliation or supportiveness e.g. abide, absorption, accede, acceptance, accompany	lexical
Coll GI	Human Collectives: 191 words referring to all human collectivities (not animal) (e.g. administration, agency, air, alliance, army)	lexical
Submit GI	Submit: 284 words connote submission to authority or power, dependence on others, vulnerability to others, or withdrawal (e.g. abdicate, abject, abscond, accept, adjust) Topic: Dominance, respect, money, and power	lexical
Rspoth Lasswell	Respect Other: 182 words regarding respect that are neither gain nor loss, e.g. admirable, admiral, admiral, admiration, age	lexical
Bldgpt GI	Building parts: 46 words for buildings, rooms in buildings, and other building parts	lexical
Vehicle GI	Vehicle: 39 words	lexical
Gratitude GALC	Gratitude, like „great", „thank you"	lexical
nwords	number of words in comment	surface
com_link_count	amount of links in the comment	surface
ADP	prepositions and postprepositions	syntax
first_person	relative amount of first person pronouns	syntax
VERB	relative amount of verbs	syntax
AUX	relative amount of auxiliary verbs	syntax
past_tense	relative amount of past tense verbs	syntax

Table 8: Overview of linguistic features (emotions, lexical, surface, syntax, textual complexity) with short description and features type.

C Additional analysis: probability of UM (regression)

In order to identify which annotated properties have a significant impact on UM we conducted additionally conducted another linear regression. With this we aim to investigate the relationships between the annotated properties of the interaction and the probability for the item being a form of UM. More specifically we would like to know (a) Which annotated features are strong signals for UM? and (b) Which moderation functions are most prevalent in UM? We treat the soft label for UM (according to standard normalization) as the dependent variable (DV) and the values of the other annotation layers in the aggregated dataset as independent variables (IV). We start with a model which only has one IV and incrementally increase model complexity by adding an IV if it significantly improves the fit of the model (in terms of explained variance). We compare the significance between the simpler model and the more complex one using anova.

The final model explains 62% of the variance. The forest plot in Fig. 13 summarizes the effects of the significant terms. We can draw the following conclusions:

The analysis supports the finding that replies that are associated with a higher constructiveness are more likely to be perceived as UM. The analysis also reveals a significant negative effect of subjectivity on the

IV	sign.	explvar
Residuals		36.237
nwords	0.000	20.840
sentiment	0.000	17.983
first_person	0.000	2.949
Disappointment_GALC	0.000	1.280
hdd42_aw	0.000	0.812
Eval_2_GI	0.000	0.640
Feel_GI	0.000	0.610
VERB	0.000	0.586
BNC_Spoken_Trigram_Normed_Freq	0.000	0.529
com_link_count	0.000	0.528
Bldgpt_GI	0.000	0.489
BNC_Spoken_Bigram_Proportion	0.001	0.424
Vehicle_GI	0.001	0.420
KF_Freq_AW_Log	0.001	0.400
agressiveness	0.002	0.396
LD_Mean_RT_FW	0.003	0.343
BNC_Spoken_Trigram_Proportion	0.005	0.318
KF_Freq_AW	0.005	0.308
Submit_GI	0.006	0.302
Ortho_N	0.007	0.290
ADP	0.008	0.274
Brown_Freq_AW	0.010	0.265
Irritation_GALC	0.012	0.251
BG_Mean	0.017	0.224
num_exclamation	0.019	0.219
BNC_Spoken_Bigram_Normed_Freq_Log	0.022	0.207
NUM	0.029	0.187
Rspoth_Lasswell	0.032	0.182
Gratitude_GALC	0.040	0.166
Freq_N_P_FW	0.052	0.149
Fail_GI	0.084	0.118
past_tense	0.101	0.106
Decreas_GI	0.110	0.101
OG_N_H_FW	0.113	0.099
You_GI	0.123	0.094
Coll_GI	0.143	0.085
Affil_GI	0.235	0.056
Natobj_GI	0.278	0.046
AUX	0.323	0.039
Rcethic_Lasswell	0.366	0.032
Longing_GALC	0.676	0.007
lexical_density_types	0.976	0.000

Table 9: Significant terms of the most explanatory regression model for predicting constructiveness, with degrees of freedom, statistical significance and explained variance.

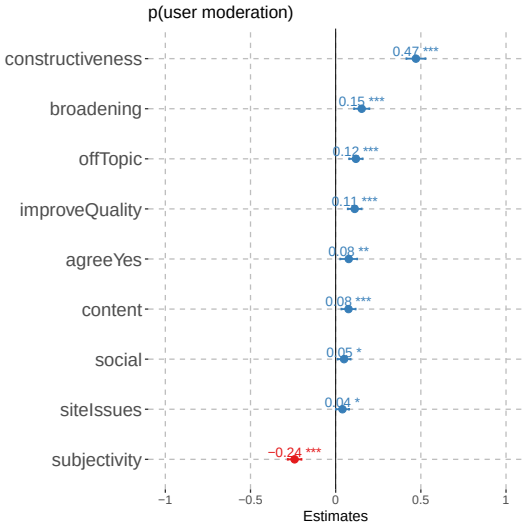


Figure 13: Standardized beta values of selected terms (most explanatory regr. model, $R^2 = 62\%$): forest plot

probability of UM.

In isolation, both aggressiveness and sentiment exhibit a similar effect (higher aggressiveness or negative sentiment reduce the probability of moderation). However, these two variables do not account for additional variance, indicating that the phenomenon of 'constructiveness' encompasses both, and possibly other factors.

In terms of moderation functions that are covered by users, we can see that users focus more on improving the content of the post they are replying to, such as correcting false information, giving feedback to improve the argument, asking the participants to stay on topic or asking questions to broaden the perspective. Functions that operate on the meta level (resolving site issues, social functions) and policing are less relevant. The analysis further supports this finding by showing that the effect of the content-oriented functions is much larger and significant.

D Additional analysis: task definitions by the annotators

The following analysis was conducted in order to gain a better understanding about the annotators personal definitions about UM and how they understood that concept before and after reading the guidelines. To compare the definitions within one annotator (before and after the study) and between different annotators converted the textual definitions into vector representations using SBERT (Reimers and Gurevych, 2020). We use the `all-MiniLM-L6-v2` model and the transformer library from huggingface. We compute semantic similarity as the cosine similarity between two encoded definitions.

How semantically similar are the definitions (before and after the study)? The average semantic similarity between the definitions before and after the study is 0.63, with a standard deviation of 0.22, indicating noteworthy variations among annotators in how extensively they revise their definitions post-study. When we compare the average pair-wise similarity of definitions across all annotators, we observe a slight decrease in their average similarity (0.477 before and 0.462 after the study).

We apply k-means clustering to group the encoded definitions into four clusters, both before and after the study. This analysis reveals that the observed trend cannot be universally applied to all annotators; instead, it points to specific subgroups that either become more similar (exhibiting higher within-cluster similarity) or more diverse (larger decrease in average similarity for the cluster demonstrating the highest within-cluster similarity before the study).

Qualitative inspection of the definitions reveals that before the task annotators tend to rely more on copying and pasting textual fragments from the guidelines and express these definitions more in their own words after the study which can explain the variation in increasing/decreasing semantic similarity between different groups of annotators.

Can we identify patterns of similar definitions? In our qualitative review of the distinct clusters of UM definitions, we uncover notable trends. Some annotators exhibit distinct priorities regarding certain moderator functions, with a focus either on fostering a civil discourse and enforcing rules or on ensuring topic relevance. Additionally, a prevalent pattern emerges concerning annotators' biases toward a particular style. Despite the guidelines explicitly stating that UM need not conform to the neutrality and politeness standards typical of expert moderation, there is a group of annotators that consistently perceive neutrality and politeness as essential ingredients for UM according to their definitions.

This supports the findings from the analysis as users do perceive UM comments as more neutral, constructive and respectful. We can thus conclude that moderation is not solely understood as employing particular pragmatic speech acts. Annotators also consider tone and style of the comments when evaluating moderation, and although it may differ from the style of expert moderators, it remains an essential factor for their interpretation of moderation.