Facilitative Moderation for Online Participation in eRulemaking

Joonsuk Park
Dept. of Computer Science
Cornell University
Ithaca, NY, USA
jpark@cs.cornell.edu

Claire Cardie
Dept. of Computer Science
Cornell University
Ithaca, NY, USA
cardie@cs.cornell.edu

Cynthia Farina
Law School
Cornell University
Ithaca, NY, USA
crf7@cornell.edu

Sally Klingel
ILR School
Cornell University
Ithaca, NY, USA
slk12@cornell.edu

Mary Newhart
Law School
Cornell University
Ithaca, NY, USA
mjn3@cornell.edu

Joan-Josep Vallbé
Dept. of Political Science
University of Barcelona
Barcelona, Spain
vallbe@ub.edu

ABSTRACT

This paper describes the use of **facilitative moderation** strategies in an online rulemaking public participation system. Rulemaking is one of the U.S. government's most important policymaking methods. Although broad transparency and participation rights are part of its legal structure, significant barriers prevent effective engagement by many groups of interested citizens. Regulation Room, an experimental open-government partnership between academic researchers and government agencies, is a socio-technical participation system that uses multiple methods to lower potential barriers to broader participation. To encourage effective individual comments and productive group discussion in Regulation Room, we adapt strategies for facilitative human moderation originating from social science research in deliberative democracy and alternative dispute resolution [24, 1, 18, 14] for use in the demanding online participation setting of eRulemaking. We develop a moderation protocol, deploy it in "live" Department of Transportation (DOT) rulemakings, and provide an initial analysis of its use through a manual coding of all moderator interventions with respect to the protocol. We then investigate the feasibility of automating the moderation protocol: we employ annotated data from the coding project to train machine learning-based classifiers to identify places in the online discussion where human moderator intervention is required. Though the trained classifiers only marginally outperform the baseline, the improvement is statistically significant in spite of limited data and a very basic feature set, which is a promising result.

dg.o '12, June 04 - 07 2012, College Park, MD, USA

Categories and Subject Descriptors

H.5.3 [**Information Interfaces and Presentation**]: Group and Organization Interfaces—*Asynchronous Interaction*; I.2.6 [**Artificial Intelligence**]: Learning; I.2.7 [**Artificial Intelligence**]: Natural Language Processing

General Terms

Algorithms, Human Factors

1. INTRODUCTION AND BACKGROUND

Rulemaking, the process by which agencies of the federal government issue new regulations, has become one of the most important public policymaking methods in the U.S. Between fiscal years 2001 and 2010, federal agencies finalized more than 38,000 rules [21]. Substantial transparency and public participation are built into rulemaking's formal legal structure. These include requirements that the agency notify the public of what it is proposing and why and give the public a period (average: 60 days) to comment—hence the name "notice-and-comment rulemaking." However, "public" comment continues to be dominated by submissions from large corporations, professional and trade associations, and national level interest groups [5]. This lack of broader participation is problematic because, although many rulemakings concern only limited populations, a significant subset of proposed new regulations will directly and substantially affect individuals, small businesses, local governments and not-for-profits[16].

For these reasons, for nearly 20 years rulemaking has been a target of U.S. e-government efforts to increase participation. In 2002, the E-Government Act directed that agencies provide essential rulemaking documents online and allow for electronic submission of comments [12]. This "first generation" eRulemaking essentially put the conventional process online, through a government-wide portal, Regulations.gov. Although Regulations.gov has created easier access to rulemaking materials and made comment submission simpler, it has not significantly broadened public awareness of, or effective engagement in, rulemaking [4].

Second generation eRulemaking has been launched by Obama Administration mandates requiring that agencies use Web

2.0 technologies to increase transparency and participation in federal policymaking. Regulation Room is an experimental "Rulemaking 2.0" participation platform, in which multidisciplinary expertise is being used to develop socio-technical strategies for lowering the barriers to broader effective rulemaking participation. The site is designed and operated by the Cornell eRulemaking Initiative (CeRI) at Cornell University, which comprises univerity researchers from communication, computer science, information science, law, and conflict resolution.[1] Launched in Fall 2009, the project is an unusual collaboration between academia and federal rulemaking agencies that allows for development and field testing of hypotheses on how to alert and engage the kinds of stakeholder groups (as well as interested members of the general public) who are typically silent or ineffective in the conventional process.

We perceive **four principal barriers to rulemaking participation** that an online public participation system must address: (1) *Unawareness*: Despite the formal "notice" agencies give of rulemakings, most of the public do not realize when a new regulation that would directly affect them is being proposed; (2) *Information Overload*: The legal requirements for rulemaking combined with the economic and scientific nature of many of regulatory problems cause agencies to produce long, linguistically and technically complex materials about their proposals. These impose high cognitive and attentional demands on new commenters; (3) *Participation Illiteracy* : Newcomers to rulemaking tend not to understand that it is a rational analytical process driven by data and reasoned argumentation, rather than a majoritarian process of preference aggregation. Lack of understanding leads to low-cost but ineffective forms of participation such as mere expression of sentiment, or mass e-comment "voting;" (4) *Motivation*: Particularly in light of the information intensity of rulemaking proposals and the demands of meaningful participation, individuals often lack motivation to engage effectively even if the outcome directly affects them.

Regulation Room uses multiple methods to address these four barriers, including social and conventional media outreach, site design and functionality, online educational materials, and information translation and layering [10, 11]. This paper focuses on one critically important method: **facilitative human moderation**. Based on research in deliberative democracy and alternative dispute resolution [24, 1, 18, 14], we are evolving online facilitative moderation strategies to encourage both effective individual commenting and knowledge-creating group interchange.

Here we describe the Moderator Protocol that has been developed over four "live" Department of Transportation rulemakings in 2010-2011. As currently conceptualized, it identifies seven distinct moderator roles for facilitating informed commenting and productive group discussion, operationalized through nineteen types of possible facilitative interventions moderators can make (Section 2). We also describe the data (Section 3) and the process of developing detailed manual coding of moderator interventions for two of the Regulation Room rulemakings (Section 4).

Then, to lower costs and increase the scalability of the system, we investigate the feasibility of employing the annotated data from the coding project to train machine learning-based classifiers to identify places in the online discussion where human moderator intervention is required (Section 5). Though the trained classifiers only marginally outperform the baseline, the improvement is statistically significant in spite of limited data and a very basic feature set, which is a promising result.

[1]See regulationroom.org.

2. THE ROLE OF MODERATION IN ONLINE RULEMAKING

2.1 Goals and challenges of online public participation

As noted earlier, rulemaking is a deliberative, technocratically rational process in which the agency considers all the relevant facts and weighs reasoned positions and arguments to discover the best outcome for the public good, within the boundaries of the legal authority it has been given by Congress [10]. Because of the nature of this process, an online participation system can increase effective public engagement by highlighting relevant facts and issues and presenting important rulemaking materials in ways that enable previously silent stakeholders to make the kinds of comments that have weight in the agency's final decision. Beyond helping newcomers make individually better comments, the system might foster dialogue among those with an interest in the proposed rule. In the conventional commenting process, the most significant and substantive comments tend to be strategically filed on or near the closing date of the comment period [11]. This provides little opportunity for affected constituencies to engage in discussion that might generate new knowledge and ideas and reveal the basis for mutually beneficial solutions to difficult regulatory problems. An online participation system might be able to follow a period of stakeholder discussion with a consensus-building process that develops common ground and presents the agency with regulatory approaches having cross-group support.

As the next subsection explains, in the context of face-to-face discussion, the value of active, facilitative moderation to achieving informed participation, deliberative discussion, and consensus-building is well-recognized. So far, however, prominent federal government online participation efforts have lacked this element, and the participatory outcomes have often been disappointing. Users have tended to engage each other only with conclusory or discursive remarks, or by simple operations such as "thumbs-up/down" ratings. Moreover, users strongly concerned with tangential or irrelevant issues have been able to hijack the discussion. For example, in the White House Open Government 2009 online brainstorming session on national open government priorities, users voted up to the top of the priority list legalizing marijuana and resolving questions about President Obama's birth certificate; at the same time, germane and thoughtful suggestions went undiscussed and were "demoted" from view for lack of votes [19]. There was also some evidence of orchestrated campaigns to vote down certain ideas and comments in order to intimidate and suppress opposing views. In the collaborative drafting phase of this event, even with technology that enabled easy incorporation and attribution of text from multiple participants, users tended to substitute their own preferred text wholesale – and then elicit friends and members to vote for their version – rather than genuinely engaging language offered by others.

2.2 Contribution of face-to-face participation design

Both the goals and challenges of online participation are familiar to those who study and practice deliberative democracy and alternative dispute resolution (ADR). Deliberative democracy practitioners design processes to mitigate problems of unequal access to information and other participation resources that can privilege the interests of the most powerful groups in public policymaking [24]. ADR practitioners recognize that, even when access to information has been equalized and support for participation provided, reasoned deliberation does not naturally and necessarily follow [1]. ADR consensus-building processes assume conflict among participants and recognize that groups require help to manage conflict in order to reach satisfactory deliberative outcomes [18].

The following conditions are generally recognized as necessary for effective group dialogue and consensus-building [14]:

– A meaningful task with a defined impact
– Full inclusion of relevant stakeholders
– Accessible information, equally available to all participants
– Dialogue processes that foster respect and listening, and equal ability to participate
– A process that permits questioning of assumptions and current agreements
– A focus on surfacing stakeholder interests and using them as the basis for mutually satisfactory agreements

Facilitators of face-to-face group deliberative processes are responsible for promoting these conditions using techniques drawn from communication theory and cognitive psychology, studies of group behavior, theories of conflict, and participatory action research and action inquiry [24, 13]. They understand themselves to be advocates of the process, rather than of any particular position or outcome. Modeling and maintaining an atmosphere of respect for all participants, their task is to create and maintain the conditions in which effective individual engagement and group deliberation can occur.

2.3 Facilitative moderation in Regulation Room

Using the techniques demonstrated to be effective in face-to-face settings, the (human) moderators of Regulation Room work to motivate meaningful participation from previously unengaged stake holders, bridge information gaps, and promote knowledge-creating and collaborative problem-solving discussion among participants. Their charge is to maintain a process that (i) helps each participant craft his/her own comment in the way most useful to the agency—specific and clear, with articulated justifications; and (ii) encourages interchange among participants.

Of course, numerous challenges exist in translating "in the room" deliberation and consensus-building techniques to the environment of an online rulemaking participation system. Discussion is asynchronous, both among users and between commenters and moderators (who do not monitor the discussion 24-hours a day). The frequency and fluency of participation varies greatly across individual commenters. Participants often have quite different expectations of the norms and purposes of participants in this online setting, as well as different levels of computer skills and familiarity. For this reason, the Moderator Protocol for Regulation Room has evolved over the four live rulemakings that have been offered on the site.

Table 1 is a high-level depiction of the current version of the Protocol. It is shows seven distinct Moderator roles – each of which is operationalized through one or more facilitative interventions. These roles create the conditions for effective deliberation and consensus-building by increasing task clarity and focus, helping commenters articulate their interests and contributions, fostering shared group process norms, and ensuring that individuals have the substantive and site use information required to participate effectively.

The Moderator Roles address participation barriers by helping commenters find and understand relevant information about the agency's proposal, mentoring them in creating the kinds of comments that will have value in the final decisionmaking, and motivating participation through creation of an environment of robust but respectful interchange of views. In many moderator-commenter interactions, the moderator is playing more than one role and is undertaking more than one intervention. For example, a moderator might welcome a first time commenter (Social Functions) and ask him/her to provide more details (Improving Comment Quality). Or a moderator might express appreciation for a thoughtful comment criticizing the agency proposal (Social Functions) and ask the commenter or the community at large to suggest alternative approaches (Improving Comment Quality; Broadening Discussion).

We expect to continue to refine the Moderator Protocol over time. The objectives, however, remain constant: encourage participation and safeguard the legitimacy of each individual commentator, while at the same time taking advantage of the communal platform of online commenting to encourage participants to learn from each other and engage in dialogue that clarifies separate and shared issues.

Full evaluation of facilitative moderation techniques on the outcome of the rulemakings has been difficult, as DOT has only announced the final decision on one of the four "live" rulemakings moderated on Regulation Room. In that final decision (on the APR rule, see Section 3), comments from Regulation Room users are mentioned in DOT's discussion of almost 20 different sections of the new rule. For example, DOT states (in a section on Full Fare Advertising) how comments on Regulation Room influenced its decision to require clearer information in fare advertising, even though comments from airline companies had opposed this change.[2] Note that many of the Regulation Room comments mentioned in the final rule were proposed by, or discussed among, multiple commenters, and several of these are the consequences of moderations stimulating the collective efforts by referring users to others with similar ideas. In addition, users often provided more details in response to moderator interventions, making their comments more sound and influential to DOT.

3. MODERATION DATA FOR "LIVE" RULEMAKINGS

We have developed and used the facilitative moderation

[2]The full regulatory provisions are available from http://regulationroom.org/airline-passenger-rights/agency-documents/final-rule/.

Table 1: Moderator Roles and Interventions.

Moderator Roles	Interventions
Social Functions	Welcoming Encouragement; appreciation of comment Thanking users for participating
Resolving Site Use Issues	Resolving technical difficulties Providing information about the goals/rules of moderation Providing information about role of CeRI
Organizing Discussion	Directing user to another issue post more relevant to his/her expressed interest
Policing	Redact and quarantine for inappropriate language or content Maintaining/encouraging civil deliberative discourse
Keeping Discussion on Target	Explaining why comment is beyond agency authority or competence or outside scope of current rule Indicating irrelevant, off point comments
Improving Comment Quality	Providing substantive information about the proposed rule Correcting misstatements or clarifying what the agency is looking for Pointing to relevant information in primary documents or other data Pointing out characteristics of effective commenting Asking users to provide more information, factual details, or data to support their statements Asking users to make or consider possible solutions/alternative approaches
Broadening Discussion	Encouraging users to consider and engage comments of other users Posing a question to the community at large that encourages other users to respond

Table 2: Regulation Room: Basic data for four rulemakings, 2009-2011.

Rule	Days open	Unique Visitors	Visitors registered as users	Total Comments	Users who submitted comments	Moderator Responses
TEXTING	34	3665	54	32	18	16
APR	110	19320	1189	931	348	197
EOBR	106	5328	121	235	68	104
ATA	112	12631	53	103	31	60

strategies in four "live" DOT rulemakings on Regulation Room:

1. TEXTING: a proposed ban on texting while driving by commercial motor vehicle operators (the "texting rule" of 2010) [7];

2. APR: a proposal to increase airline passenger rights in areas such as bumping, tarmac delay, and fee advertising (the "APR rule" of 2010) [6];

3. EOBR: a proposal to require commercial motor vehicle operators to purchase and install electronic on-board recorders to verify compliance with maximum driving time rules (the "EOBR rule" of 2011) [9]; and

4. ATA: a proposal to require that air travel websites and automated airport check-in kiosks be made accessible to people with disabilities (the "Air Travel Accessibility" rule) [8].

Moderators were students from the Cornell University Law School, trained according to the Moderator Protocol (Section 2.3). Rather than all students moderating the full online discussion for the rulemaking, moderators were assigned to cover issue-specific threads of the discussion. Basic participation and moderation data for these four rulemakings are summarized in Table 2.

This paper analyzes moderation data from the APR and EOBR rules only. The Moderation Protocol was insufficiently developed in the very brief first rule (TEXTING) and data from the fourth rule (APR) were only entering the coding process when this paper was written.

4. CODING PROJECT

The Regulation Room software records all user comments and moderator interventions. Moderators use a specially designed interface that facilitates their interventions, but this interface did not, at the time of the EOBR and APR rules, allow moderators to record the reason for the intervention — i.e., neither the Moderator Role associated with the intervention nor an indication of which of the 19 possible intervention types was intended was recorded. A first step for the analysis, therefore, was to manually code each intervention, keeping in mind that more than one type could be associated with each Moderator intervention.

We expected that coding all moderator interventions with respect to the full set of 19 possible interventions would be difficult. We therefore employed a series of coarser annotation schemes at the level of Moderator Role. We tested both the validity of the coarse coding scheme and coder reliability, across a total of six rounds.

As can be observed in Table 3, four of these rounds, aimed at training the coding team and testing the coding scheme,

Table 3: Basic information for all coding rounds carried out for the EOBR rule: number of comments, number of coders, and number of intervention categories in the coding scheme.

session	# comments	# coders	# categories
1	22	6	5
2	21	6	6
3	23	6	6
4	19	3	7
5	105	3	7
6	198	3	7

were performed with a sample of approximately 20 comments each. The last two rounds were done with a considerably larger number of comments (all comments for the two rules). The first three rounds were performed by a team of six coders; in the last three rounds, the team was reduced to three members due to availability of student coders. (Later rounds had fewer coders because they were done during the summer and we lost coders to the bar exam.)

As Table 3 also shows, in the first round coders were given a Moderator Protocol that had only five categories of intervention (i.e., five Moderator Roles) as the coding scheme. After each coding round, the coders were debriefed by the project team in order to receive feedback regarding the coding scheme and discuss the results and any doubts that might have been raised during the coding task. As a result of this process, one Moderator Role was added at the second round. After the third round debriefing, the addition of another Role resulted in the final set of Roles and interventions shown in Table 1. More detail about this process is given in Section 4.2.

4.1 Measures of inter-coder agreement

The results of each coding session were recorded in the form of a m x n matrix, where m were coders and n were the comments coded. There are numerous ways of assessing the level of agreement among coders that have been widely discussed in the literature [17]. In this project we have used three different types of agreeement measures: percentage of agreement, measures that take agreement by chance into account, and measures of covariation.

4.1.1 Percentage of agreement

The percentage of agreement is the most simple measure and perhaps the least reliable of all measures usually used in coding. Its general form is presented in Equation 1 (adapted from [20]):

$$PA_o = \frac{A}{n} * 100 \tag{1}$$

where, PA_o is the observed percentage of agreement, A is the observed number of agreements between coders, and n is the total number of comments coded by the coders. The literature on inter-coder reliability measures has stressed [20] that the main problem affecting this measure is that it does not control for the agreement that happens by chance (e.g., a probability of 0.5 with two coders and a binary coding scheme). For this reason, the percentage of agreement is usually complemented with other measures that take chance into account.

4.1.2 Alternative measures of agreement

Measures such as Krippendorff's α (alpha), Cohen's κ (kappa), Scott's π (pi), or Fleiss' κ (kappa), which take into account the probability of agreement by chance, are considered more adequate than the percentage of agreement [20].

The adequacy of each measure depends heavily on both the type of data resulting from the coding activities, and the number of coders. Yet, when more than one statistic is available for the same type of data and number of coders, their results are most times interchangeable. In this case, both Krippendorff's α and Fleiss' κ (which is a generalization of Scott's π [20]) are the most adequate statistics for describing intercoder agreement on binary data with more than two coders [17]. Fleiss' κ is displayed in equation 2:

$$\kappa = \frac{Pr(a) - Pr(e)}{1 - Pr(e)} \tag{2}$$

where $Pr(a)$ is the observed level of agreement, and $Pr(e)$ refers to the agreement that should be expected to happen by chance. Note, though, that both observed and expected agreement must be computed taking into account k raters, n comments, and m categories. A similar statistic is Scott's π (pi), which is based on the same idea [20].

Krippendorff's α, on the other hand, takes a slightly different and simpler form:

$$\alpha = 1 - \frac{D_o}{D_e} \tag{3}$$

where D_o is the observed disagreement, and D_e is the expected disagreement. The values of these statistics range generally from 1 to -1, with 1 indicating perfect agreement, and -1, systematic disagreement. Values close to zero are given when there is not agreement at all or agreement is reached only by chance [17]—the observed disagreement is equal to the expected disagreement by chance.

Regarding what constitutes an acceptable value of these agreement measures, we follow the reasonably strict criterium set by [17], despite the low level of consensus existing among social science researchers in this matter: data are reliable when $\alpha \geq .800$, may be considered for drawing preliminary conclusions when $.667 \leq \alpha \leq .800$, and should be discarded when $\alpha \leq .667$.

4.1.3 Measures of covariation

Finally, a different family of statistics is used in order to assess, if not coders' agreement, their covariation. These statistics measure the extent to which coders behave similarly when coding, even when disagreeing. In this framework, covariation is expressed through the intraclass correlation coefficient (ICC), an extension of the popular Pearson's standardized correlation coefficient (r), which was designed to account for the linear correspondence between two sets of data points. The intraclass correlation coefficient is especially adequate for computing correlation when data are organized into groups [23]. It then computes to what extent there is correlation within each group. In our case, each comment counts as a different group, and we are therefore interested in the extent to which coders behave similarly on each comment in order to assess their consistency. The coefficient has values in the range [-1,+1], where +1 means perfect *positive* correlation (two coders always code the same), 0 means no covariation at all (correspondence between the

values assigned to each comment by each coder), and −1 means perfect *negative* correlation (two coders always code differently). It should be kept in mind that correlation coefficients are fair indicators of the general reliability of our data (they help identifying outlying cases and inconsistencies in the coding activities), but they are not measures of agreement *per se*.

4.2 Results

The results of the tests on our reliability data are shown in Figure 1. In the first coding round, the team of coders was given a coding scheme with the following five Roles:

(a) Social functions
(b) Site use issues
(c) Policing
(d) Stimulate discussion
(e) Improving comment quality

The overall results of this first round were modest. While the consistency of the coding activity was relatively high (ICC), the percentage of agreement was only 80.6%, and the quality threshold of 0.8 in the α and κ coefficients was not reached. When the coding team was debriefed by the project coordinators, the project team decided to add an additional category to the coding scheme: "Directing user to another post."

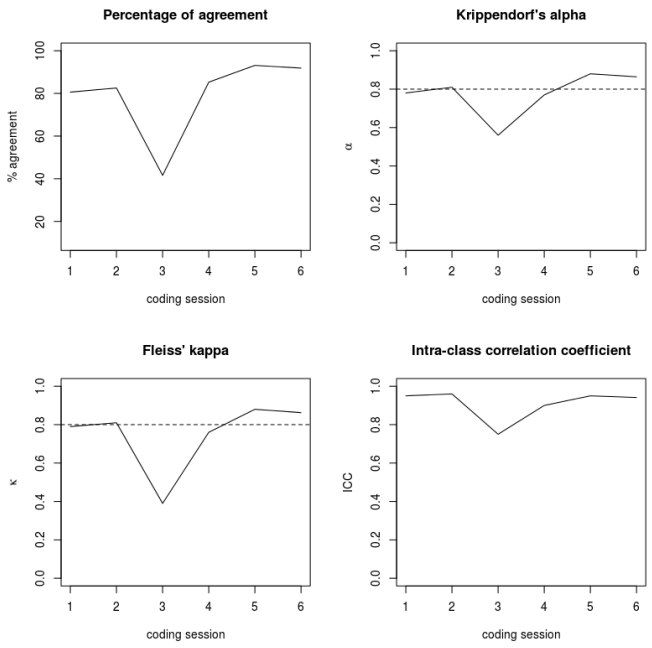

Figure 1: Summary of results of the reliability tests in all coding rounds in terms of averages of percentage of agreement, Krippendorf's alpha, Fleiss' kappa, and intraclass correlation coefficient.

The second round, then, was made with a coding scheme containing six different coding categories. With these changes and the higher familiarity of coders with the coding scheme, results improved a little. While both the percentage of agreement and the intraclass correlation coefficient did experience only slight increases (2% and .01 respectively), the two reliablity coefficients crossed the quality threshold by a small margin (.81 in both cases). Despite the better results, the coding sessions did not lack difficulties for the coders. Debriefing resulted in adjustment of the interventions within the Moderator Roles, but did not create any additional Roles.

Despite all these changes, the results of the third round were far below expectations. Suddenly, the percentage of agreement was just above 40%, ICC was still high, but the coefficients were way below the .667 threshold of minimum acceptance. One of the problems here was a high level of non-responses, possibly due to the unequal levels of internalization of changes among coders. After this round, the debriefing session was used again to thoroughly review conflicts between coders, and changes were made that resulted in the ultimate, seven Role scheme.

Two weeks after the third round, a fourth round was carried out and the changes seemed to be quite effective. Overall results improved dramatically and brought the relability of the data back to the levels of our first coding round (just below the .8 threshold in the α and κ coefficients). After this round, the debriefing was mainly devoted to making small adjustments to the coding scheme explanatory documentation.

After the four training sessions, the team of coders was provided with the entire set of moderator comments from the EOBR and APR rules for coding. The high-level organization of the final coding scheme used for both rules had seven categories, corresponding to the organization of Moderator interventions in Table 1.

As observed in Figure 1, the performance of the coding team with the whole set of comments for both the EOBR and APR rules was far better than in previous results. First, the percentage of agreement for both rules was above 90% (93.1 and 91.9, respectively), and the consistency of the coding activity was quite high (ICC = .94). Moreover, in both cases the acceptance threshold for both reliability coefficients was surpassed (.88 for EOBR and .86 for APR).

5. AUTOMATED MODERATOR INTERVENTION

As discussed in the previous sections, moderator interventions are a key element in lowering the significant barriers to broader, effective public participation in rulemaking. However, a fully human-operated moderation system poses a significant challenge in scalability. In this section, we propose to automate the identification of the moderator intervention type required for an individual comment. We use a supervised learning algorithm called Support Vector Machines (SVM) [25] with the text of the comment encoded as features using a bag of words (BoW[3]) representation [15]. We evaluate the plausibility of the automation scheme through experiments that compare system performance on in-domain data (data from the same rule) and cross-domain data (from different rules).

5.1 Learning Task

From a natural language processing (NLP) perspective, various components of the moderator intervention process

[3]The name comes from the fact that the words themselves are used as features without considering their relative locations, as if the words are thrown into a bag.

Table 4: The number of comments moderated with interventions serving each of the Moderator Roles

Moderator Role	APR	EOBR
Social functions	178	49
Site-use Issues	9	9
Organizing Discussion	2	17
Policing	0	1
Out of Bounds for Agency	2	17
Improving Comment Quality	152	86
Broadening Discussion	52	32
Total Comments	197	104

can be automated. Here, we focus on automating the identification of which types of moderator response a given comment requires. We automate this step, as opposed to others, because of its relative simplicity and usefulness: once a fully functional system is implemented, each comment can be tagged with the desired types of moderator responses. The tags can then be used to guide the moderators as they interact with users or to evaluate moderator interventions during, and after, moderator training sessions.

Thus, we formulate each of the seven moderator response type prediction tasks as an independent binary classification problem in which a comment is to be classified as to whether it requires an intervention of the specified type (*true*) or not (*false*). How this is done is described in the next section.

Unfortunately, many of the moderator reponse types (i.e., moderator roles) are not suitable for automation using supervised learning methods, including:

- Social Functions. As described in more detail below, we propose to build classifiers that use the words of each comment as features (i.e., clues for categorization), but identifying whether or not a social function is required is better learned using other types of features, such as user-level activity logs.
- Site-use Issues, Organizing Discussion, Policing, Out of Bounds for Agency. The number of comments in the dataset receiving these types of moderator intervention is quite small (see Table 4) — too small to allow adequate learning.

As a result, we examine only Improving Comment Quality and Broadening Discussion in the experiments below.

For these tasks, we want to train classifiers that are optimized for precision rather than recall.[4] The reason is that, from the perspective of building an automated tagging system that guides human moderators in their selection of comments for intervention, we can aim to train classifiers that either (a) tag few comments for moderation with high accuracy, i.e., "You should really respond to these comments;" or (b) tag many comments for moderation with lower accuracy, i.e., "Just look through these comments; some may not actually require action, but you will not miss comments that need intervention." Though each type of classifier is advantageous in its own right (and we may want to take the middle ground eventually), the available dataset allows for a better training of the former — high-precision classifiers.[5]

[4]See section 5.2.4 for the technical definitions.

[5](b) describes high-recall classifiers.

This is because the negative responses (i.e., no intervention of a particular type is needed) in the dataset are inherently noisy, whereas the positive ones are relatively clean. That is, it is more probable that the moderators did not perform a certain type of intervention that could have been done than that they gave responses of wrong types when they did intervene. Moreover, not responding to a comment can be deliberate: since it is undesirable to overwhelm users with too much feedback, moderators may omit responses of other applicable types once a certain type (hopefully a more crucial type) of intervention has been given already.

5.2 Experiment Methodology

The classifiers are trained using LIBSVM [3], a popular off-the-shelf SVM package, with a linear kernel. SVMs are suitable for this task because they can handle high dimensional data effectively, and the BoW representation typically results in a high dimensional feature set.

5.2.1 Preprocessing

The purpose of preprocessing is to minimize insignificant variations in the text that can hinder learning, while retaining characterizing contents of the text. In this experiment, all letters are lowercased, and numerics are converted to a *NUM* token under the assumption that the actual digits do not serve as distinguishing features. Also, stop words[6] are removed. (The Porter Stemmer from the ntlk package [2] was also initially employed, but it either hurt, by 20% in some cases, or did not improve the performance by a significant degree.)

5.2.2 Features

Each comment is presented to the classifiers in a form that is more or less like a list of its characteristics, called *features*. Determining the information to be contained in the feature set is an important design decision in learning problems. With plain text, the BoW representation is typically used, for it is simple, yet powerful. Simply put, each comment is represented as a binary bit string denoting which of the words known to the classifier[7] it contains.

As an example, consider a short comment from the APR dataset: *"Few if any airlines offer peanuts anymore. But peanuts are not the only allergen- what about pets? More people are allergic to pets than peanuts."* Its feature representation would contain a 1 in positions that correspond to "airlines," "offer," "peanuts," and the remaining content words in the comment; it would contain 0's in positions of the bit string that correspond to content words that appear in other comments but not in the current one.

The job of the SVM is to learn how to weight the features so that their weighted sum (i.e., the cross product of the weights and the binary feature values) is high for the desired moderation types and low for others. In our example, one such desirable moderation is Improving Comment Quality: the moderator should ask the commenter for data to support the claim that "More people are allergic to pets than peanuts." And, after training, we indeed find that the Improving Comment Quality classifier recognizes that this comment contains words associated with highly positive weights, such as "people," and classifies the

[6]Stop words are terms that appear too frequently in documents to be distinguishing features, such as the word "the."

[7]These are words that appear at least twice in the data.

Table 5: The distribution of positive and negative comments in each dataset. "Combined" refers to the union of APR and EOBR.

Task	Dataset	+	-	Ratio
Improving Comment Quality	APR	152	86	3.38:1
	EOBR	86	18	4.78:1
	Combined	238	63	3.78:1
Broadening Discussion	APR	52	145	1:2.78
	EOBR	32	72	1:2.25
	Combined	84	217	1:2.58

comment as *true*, i.e., the comment will benefit from an Improving Comment Quality intervention. (An intuitive explanation is that sentences containing the word "people" are often generalizations that need further substantiation.[8]) Here, the word "people" has a highly positive weight, because it frequently appeared in comments in the training set that are moderated via an Improving Comment Quality intervention.

5.2.3 Training and Testing

The system consists of a 2-level 5-fold cross validation. In the outer level, a dataset is randomly split into 5 equally sized portions that do not cross comment boundaries. Then SVM classifiers are trained and tested 5 times: in each iteration, one of the partitions is used as the test set, and the other 4, as the training set. In addition, each time a classifier is tested on the test set, it is also tested on the other rule-making dataset in its entirety to measure the cross-domain performance. Results are averaged across the 5 iterations. The same procedure is repeated for a baseline classifier — a trivial classifier that predicts every comment to be positive w.r.t. the response type of the classifier.[9] By reiterating the train-test process 5 times and averaging the results, we make sure that the evaluation of the classifiers is not spoiled by a single train-test split that happened to result in an abnormally low, or high, performance.

To optimize for precision, the SVM hyperparameter, C[10], is tuned during training, using 5-fold cross validation on just the four training set partitions. In particular, because the datasets are unbalanced for both the Improving Comment Quality and Broadening Discussion tasks (see Table 5), misclassifications of comments from the minority class are more heavily penalized. For instance, Improving Comment Quality for the APR dataset has a positive to negative ratio of 3.38:1; thus, a false positive is penalized by a factor of 3.38.[11]

5.2.4 Evaluation Measures

There are numerous evaluation measures for quantifying the performance of classifiers. However, not all are suitable for our application. We will first review terminology frequently used in defining different evaluation measures and then define the evaluation measures used in our study.

- **TP** (True Positive):
  positive instance classified as *positive*
- **FN** (False Negative):
  positive instance classified as *negative*
- **TN** (True Negative):
  negative instance classified as *negative*
- **FP** (False Positive):
  negative instance classified as *positive*

When these terms appear in the equations below, they denote the number of instances of the given type.

Though accuracy, the percentage of predictions that are correct, is typically used to evaluate the performance of classifiers, it bears little or no significance when the dataset is unbalanced [22]. For instance, imagine having a dataset with positive to negative ratio of 75:25. A classifier that simply predicts every example to be positive achieves an accuracy of .75 for that dataset, while such a classifier is undesirable. Therefore, we use the balanced accuracy measure instead.

1. Balanced Accuracy (BAC):

$$BAC = \frac{\text{Sensitivity} + \text{Specificity}}{2} = \frac{\left(\frac{\text{TP}}{\text{TP+FN}}\right) + \left(\frac{\text{TN}}{\text{TN+FP}}\right)}{2}$$

BAC measures the performance with respect to both the positive and negative class and takes the average, instead of considering only the positive class. To be more specific, it is the average of the recall rate of the positive class and that of the negative class. Thus, it can be considered a reliable measure even when the dataset is heavily skewed toward one of the classes.

2. Precision / Recall / F_1-Measure:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$F_1\text{-measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Precision measures the percentage of correct predictions with respect to the positive class, whereas recall measures the percentage of positive examples in the test set that were correctly predicted to be positive. Notice that there are trivial classifiers that can easily give high precision or high recall: A classifier always predicting positive, as our baseline classifier does, is guaranteed to achieve a recall of 1.0, and a classifier that predicts positive on the most confident example and negative on the rest is highly likely to yield perfect precision. Thus, both measures have to be considered together to gain insight into the true performance. F_1-measure seeks to capture such information in a single number, the weighted harmonic mean of precision and recall.

5.3 Results and Analysis

The experiment results for the Improving Comment Quality and Broadening Discussion prediction tasks are shown

[8]Note that often it not so intuitive why certain words are assigned large weights by the classifier.

[9]Even for Broadening Discussion, whose majority class is negative, the baseline is to predict every comment to be positive. This is because we are optimizing for precision, not accuracy.

[10]The soft margin constant C determines the sensitivity to comments that are close to, or on the other side of, the decision boundary.

[11]Otherwise, the trained classifier may simply predict every instance to be in the majority class.

Table 6: Improving Comment Quality prediction results (with the baseline performance in the parentheses)

Train	Test	BAC	Prec.	Rec.	F_1
APR	APR	.56(.50)	.80(.77)*	.82(1.0)	.80(.87)
	EOBR	.56(.50)	.86(.83)*	.57(1.0)	.68(.90)
EOBR	EOBR	.49(.50)	.83(.82)	.96(1.0)	.88(.90)
	APR	.50(.50)	.77(.77)	.98(1.0)	.86(.87)
Combined	Combined	.52(.50)	.80(.79)*	.83(1.0)	.81(.88)

* The difference is statistically significant according to the χ^2 test.

Table 7: Broadening Discussion prediction results (with the baseline performance in the parentheses)

Train	Test	BAC	Prec.	Rec.	F_1
APR	APR	.55(.50)	.35(.26)*	.33(1.0)	.32(.42)
	EOBR	.55(.50)	.36(.31)*	.49(1.0)	.41(.47)
EOBR	EOBR	.47(.50)	.33(.30))*	.16(1.0)	.16(.45)
	APR	.49(.50)	.17(.26)*	.09(1.0)	.12(.42)
Combined	Combined	.53(.50)	.33(.28)*	.32(1.0)	.32(.43)

* The difference is statistically significant according to the χ^2 test.

in Table 6 and Table 7, respectively. For each task, the performance of the classifiers on each dataset, both in-domain (rows 1 & 3) and cross-domain (rows 2 & 4), and the combined dataset (row 5) are provided, along with indications of statistical significance w.r.t. the baseline (results in parentheses).

As discussed in section 5.1, the classifiers in this experiment are trained with the aim of building a system that can tag comments with the types of desirable interventions with high confidence (by employing classifiers for each task independently), even if it succeeds in identifying only a portion of such comments. And while the overall performance of the classifiers across all evaluation measures is not significantly better — even worse in some cases — than the baseline, it is encouraging to see the classifier outperforming the baseline in terms of precision, with only one exception. (Note that the baseline is highly biased toward recall, which also makes the F_1-measure fairly biased. Thus, it is acceptable, and expected, to achieve lower scores according to these metrics.)

In addition, we observe the following from the results. First, the Improving Comment Quality task seems to be domain independent: both the in-domain and cross-domain results improve over the baseline by roughly the same amount, or match the baseline, with respect to BAC and precision.

Second, the results on both tasks confirm that a sufficient amount of training data is needed to adequately train a classifier, and that a dataset consisting of about 100 comments is not enough for either task. The EOBR dataset contains only 104 comments, compared to 197 comments in the APR dataset (Table 4), and the classifiers trained on EOBR make predictions that are statistically insignificantly different from those of the baseline classifier. (Predicting every example to be positive on a positive majority dataset means useful learning has not taken place.) However, the classifiers trained on APR or the combined dataset manage to outperform the baseline, especially in terms of BAC. Once enough training data is secured, however, having additional data does not necessarily improve the performance: we see that classifiers trained on APR and the combined dataset exhibit similar performance according to the majority of the measures.

The above observations are further substantiated by the fact that the classifiers trained on APR outperform those trained on EOBR when tested on the EOBR dataset. This would not occur if there were an inadequate amount of training data or if the tasks were domain dependent.

6. CONCLUSION

We apply facilitative moderation strategies in Regulation Room, an experimental online rulemaking public participation setting. Facilitative moderation helps maximize individual contribution and promote quality discussions among the users, with the aim of generating knowledge useful for aiding the formation of new federal regulations. Site moderators follow a predefined protocol based on studies from deliberative democracy and alternative dispute resolution.

The data gathered from Regulation Room are then used to evaluate the plausibility of automating moderator intervention. Classifiers that outperform, or match, a majority-class baseline in terms of BAC and precision are successfully built. Though the resulting classifiers only marginally outperform a majority-classifier baseline, precision-based classifiers show much promise, outperforming the baseline at statistically significant levels in spite of limited data and a very basic feature set.

Further development is planned in many areas of this research. The Moderator protocol will benefit from further tuning, which is possible with the experience and data gained from this work. Also, automated moderation can be enhanced by adopting existing NLP techniques or developing new ones optimized for this task.

7. ACKNOWLEDGEMENTS

This material is based upon work supported by the National Science Foundation under Grants No. IIS-1111176 and IIS-0968450. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

[1] H. N. Aragaki. Deliberative democracy as dispute resolution? conflict, interests, and reasons. *Ohio State Journal on Dispute Resolution*, 24(3):406–478, 2009.

[2] S. Bird. Nltk: the natural language toolkit. In *Proceedings of the COLING/ACL on Interactive presentation sessions*, COLING-ACL '06, pages 69–72, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics.

[3] C.-J. L. Chang, Chih-Chung. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1 – 27:27, 2011.

[4] C. Coglianese. Citizen participation in rulemaking: past, present, and future. *Duke Law Journal 55*, pages 943–968, 2006.

[5] CSFFR. Achieving the potential:the future of federal e-rulemaking. Technical report, Committee on the Status & Future of Federal e-Rulemaking/American Bar Association, Washington, DC, 2009.

[6] DOT. Enhancing airline passenger protections. *75 Fed. Reg. 32318*, 2010.

[7] DOT. Limiting the use of wireless communication devices. *75 Fed. Reg. 16391*, 2010.

[8] DOT. Nondiscrimination on the basis of disability in air travel: Accessibility of web sites and automated kiosks at u.s. airports. *76 Fed. Reg. 59307*, 2011.

[9] DOT. On-board recorders and hours of service supporting documents. *76 Fed. Reg. 5537*, 2011.

[10] C. R. Farina, M. J. Newhart, C. Cardie, D. Cosley, and CeRI. Rulemaking 2.0. *Miami Law Review*, 65(1), 2011.

[11] C. R. Farina, M. J. Newhart, P. Miller, C. Cardie, and D. Cosley. Rulemaking in 140 characters or less: social networking and public participation in rulemaking. *Pace Law Review*, 31(1):382–463, 2010.

[12] GPO. E-government act of 2002, 44 u.s.c. § 3601 et. seq., 2002.

[13] D. Greenwood and M. Levin. *Introduction to Action Research: Social Research for Social Change*. Sage Pub, Thousand Oaks, CA, 2007.

[14] J. E. Innes. Consensus building: Clarifications for the critics. *Planning Theory*, 3(1):5–20, 2004.

[15] T. Joachims. Text categorization with support vector machines: Learning with many relevant features, 1998.

[16] S. R. F. Kerwin, C. M. *Rulemaking: How government agencies write laws and make policy*. CQ Press, Washington D. C., 2011.

[17] K. Krippendorff. *Content Analysis: An Introduction to its Methodology*. Sage, Thousand Oaks, CA., 2nd edition, 2004.

[18] C. W. Moore. *The Mediation Process: Practical Strategies for Resolving Conflict*. Jossey, San Francisco, 3rd edition, 2003.

[19] NAPA. Open government dialogue. http://opengov.ideascale.com/a/ideafactory.do?id=4049&mode=top, 2012. National Academy of Public Administration.

[20] K. A. Neuendorf. *The Content Analysis Guidebook*. Sage, Thousand Oaks, CA, 2002.

[21] OMB. 2011 report to congress on the benefits and costs of federal regulations and unfunded mandates on state, local, and tribal entities. Report to congress, Office of Management and Budget, 2011.

[22] F. Provost. Learning with imbalanced data sets 101. *In AAAI 2000 workshop on imbalanced data sets*, 2000.

[23] P. E. Shrout and J. L. Fleiss. Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86(2):420–428, 1979.

[24] L. Susskind. Deliberative democracy and dispute resolution. *Ohio State Journal on Dispute Resolution*, 24(3):1–12, 2009.

[25] V. N. Vapnik. *The nature of statistical learning theory*. Springer-Verlag New York, Inc., New York, NY, USA, 1995.