Supporting Comment Moderators in Identifying High Quality Online News Comments

Deokgun Park, Simranjit Sachar, Nicholas Diakopoulos, and Niklas Elmqvist
University of Maryland
College Park, USA
{intuinno, ssachar, nad, elm}@umd.edu

ABSTRACT
Online comments submitted by readers of news articles can provide valuable feedback and critique, personal views and perspectives, and opportunities for discussion. The varying quality of these comments necessitates that publishers remove the low quality ones, but there is also a growing awareness that by identifying and highlighting high quality contributions this can promote the general quality of the community. In this paper we take a user-centered design approach towards developing a system, CommentIQ, which supports comment moderators in interactively identifying high quality comments using a combination of comment analytic scores as well as visualizations and flexible UI components. We evaluated this system with professional comment moderators working at local and national news outlets and provide insights into the utility and appropriateness of features for journalistic tasks, as well as how the system may enable or transform journalistic practices around online comments.

Author Keywords
Computational journalism; online comments; asynchronous collaboration;

ACM Classification Keywords
H.3.3. Information interfaces and presentation

INTRODUCTION
In September 2013, Vladimir Putin published an op-ed article in the New York Times (NYT) [26]. It was an essay critical of the U.S.—some might even say prodding the public. As a result, the comments flooded in: 6,367 of them, in fact. Of those, 4,447 were eventually published along with the piece online, including 85 which were selected as 'NYT Picks', high quality comments with exceptional insights that are highlighted in the commenting interface. What makes this remarkable though is that each of these thousands of comments was read by a human moderator, a trained journalist at the NYT *before* it was published. The New York Times uses a pre-moderation strategy and employs 13 community managers to read such comments, filter out inappropriate ones before publication, and select NYT Picks for highlight [13].

Reader-contributed comments are a double-edged sword: while they increase user engagement, contribute to fostering an online community, and may even provide enriching content to both readers and reporters alike, not all comments are created equally. More specifically, reader comments are sometimes low in quality (in terms of spelling, grammar, or composition), may have a tone not commensurate to the news outlet (e.g., aggressive or obscene), and may be intentionally or unintentionally incorrect or misleading. For this reason, while there is a clear value to including reader comments on online articles, there is also a need for comment moderation to ensure that published comments are representative of the news outlet's policies. Filtering out low-quality comments only addresses half the issue; top news publishers are also interested in selecting high-quality comments that contribute particularly well to the associated article and which set the tone for the site. Crowdsourced approaches have their own limitations: selections don't convey an editorial voice, there is no central oversight to ensure balance, and selections may exhibit undesirable popularity biases.

Managing and moderating online news comments is a particularly challenging task due to the overwhelming volume of content, as well as the nuance and context that moderators sometimes need to understand and consider when dealing with sensitive or political issues. Various strategies for mitigating the scale issue have been tried: leaving comments unmoderated devolves quickly, so post-moderation is often employed to allow the community to flag or report low-quality or otherwise inappropriate comments. The pre-moderation strategy at the New York Times produces a high quality of discourse but is quite resource intensive. As a result they must limit the number of articles where comments are even allowed, as well as the time window for commenting on those limited articles.

This work presents the design and evaluation of a visual analytic tool, CommentIQ, intended to augment comment moderators' capabilities to scale the selection of *high quality* commentary on online news sites. Our design process consisted of iterative requirements gathering from

CHI'16, May 07 - 12, 2016, San Jose, CA, USA

ACM 978-1-4503-3362-7/16/05···$15.00
DOI: http://dx.doi.org/10.1145/2858036.2858389

domain experts, prototyping, piloting, and development of analytics and visual representations that enable the selection of comments that are editorially interesting. We evaluated the tool with seven professional community editors and moderators at leading local and national news outlets who used CommentIQ to identify interesting comments on different articles. The evaluation provides insight into how the various analytic dimensions, filters, and visualizations that we built enabled and supported the comment moderation task. Moreover, this evaluation with actual domain experts allows us to reason about current practices and our potential to augment these practices with visual analytics, as well as offer guidelines for the future design and development of computational journalism tools.

Our contributions include (1) the characterization of the comment moderation domain, which includes use cases, user tasks, and other design needs, (2) a web-based moderation tool called CommentIQ that was designed and developed to meet these needs, and (3) the evaluation of the designed tool with domain experts in a way that allows us to reason about current practices and the potential for technological augmentation of those practices. Our results underscore the applicability of visual analytics to enable journalistic activity, and uncover opportunities for future technology development that can enable their practices.

BACKGROUND

This work contributes to a growing body of research in the area of computational journalism, which includes tools that are tailor-designed to suit journalistic tasks and workflows, and which take into account the professional norms and use-cases of journalists. Like previous work in this vein, such as the Overview [4], CityBeat [28], or SRSR prototypes [11], we pursue a design study in this domain. Yet rather than focus on the needs of reporters, we target the CommentIQ system at an underexplored but increasingly important task and user sub-population within this domain: comment moderators. In designing the system we were informed by related work in three areas that we describe next, including *community moderation*, *analytics of comment quality*, and *discourse visualization*.

Community Moderation

Discourse quality and incivility in online comment forums is an issue that has not gone unnoticed in the research literature [6,12,30]. Research has shown that if low-quality commentary goes unchecked it can lead to detrimental and polarized risk perceptions of scientific information [1]. One approach for dealing with discourse quality is that of post-moderation: a user community can flag or report comments that they deem inappropriate and these flags can then be reviewed by professionals [12] to determine whether they should be removed. Users can also rate, tag, and vote on comments which feed into end-user interfaces for sorting and filtering comments [19]. A substantial downside to this approach is that it can take a long time for good comments to be identified, and a reliable ranking depends on having enough votes in the system [20].

Studies have shown that users are more interested in engaging with discussion that is moderated [35]. Evidence is mounting which shows that by signaling norms and expectations for behavior, the overall tenor of discourse can be improved. For instance, lower levels of incivility and a greater use of evidence in comments was found when a reporter engaged directly in a news outlet's comment threads on Facebook [32]. In another study, thoughtfulness cues in comments led to participants posting longer comments, spending more time, and writing more relevant comments [33]. The practice of selecting high quality comments by outlets such as the New York Times fits this strategy of social signaling to set community standards. The CommentIQ system was specifically designed with this approach towards comment moderation in mind, thus we focus not on the removal of low-quality comments (which is still a valid problem in its own right), but on the identification of high-quality contributions that could act as cues for a positive feedback loop with a community.

Analytics of Comment Quality

Various efforts have been undertaken to measure and rank the quality of written texts and comments, including both low quality [31] and high quality written outputs. Natural language processing of text content as well as data analysis of community information (e.g. user history and interactions) have been applied. For instance, Louis and Nenkova [22] predicted the article quality of science journalism based on lower level linguistic features, such as sentence structure. Other work in this domain has considered the measurement of dimensions of text readability [25]. Efforts have tried to automatically predict the quality of online comments, although the reported accuracy of such models makes them difficult to apply practically [3,34]. In addition to textual features such as informativeness and cohesion of text, user features can also be leveraged to rank comments, such as activity level, history of ratings, and degree to which other people respond. In contrast to Hsu et al. [17], however, we do not use community ratings as ground-truth for quality as this can reflect popularity bias. For our ground truth of "quality" we instead utilize a source of professionally curated and selected comments: the NYT "Picks" comments.

Studies in the literature describe journalistic efforts to identify high quality contributions from the public, including how letters to the editor are selected [18], how online comments are selected for print publication [23], and how on-air radio comments are chosen at NPR [27]. Specifically in the domain of online news comments, recent work by Diakopoulos [8,9] has synthesized these journalistic criteria into a set of twelve human-centric criteria including argument quality, criticality, emotionality, entertainment, readability, personal experience, internal coherence, thoughtfulness, brevity, relevance, fairness, and

novelty. In this work, we utilize the validated analytic operationalization of several of these criteria to score comments, including readability, personal experience, length, and relevance (i.e., to the article and conversation). We also derive user-based scores of quality by averaging these criteria over user history. We incorporate understanding from other literature in the domain of online reviews which suggests that a measure of user activity level will be usefully correlated to quality [2]. Finally, we train a model on a set of collected comments to arrive at default weights of these various analytic criteria that can orient moderators towards the top comment candidates.

Discourse Visualization

Previous work on comment or discourse visualization has often approached the issue from the end-user's perspective. For instance, the ForumReader tool [7] was designed to help orient and guide readers to areas of interest within large scale online forums, such as Slashdot. Another more recent effort in this area is the ConVisIt system [16], which utilizes flexible user-driven topic modeling to provide an interface that allows for exploration of asynchronous online discussions and an ability to find useful and insightful comments. The Arkose system [24] was designed to help visually distill large online discussions into more succinct summaries. A somewhat related effort is the Opinion Space system, which visualizes comments by projecting sets of elicited scalar opinions relating to controversial statements and led to increased user engagement with comments [14].

Many discourse visualization systems in the literature, including those cited above are not oriented towards sensemaking of comments that can directly enable moderation. In contrast, we designed the CommentIQ system specifically for comment moderators and we present a persona and characterization of the task and use-cases involved which inform our design. More specifically we use analytics to score comments along various dimensions of interest to journalist moderators as discussed above, and we provide interactive visualizations of these scores including map-based and temporal views that align with user needs and requirements in the domain and help orient moderators towards comments that may be high quality.

THE DESIGN PROCESS OF COMMENTIQ

We conducted a multi-phase design study with the purpose of applying both text analytics algorithms and interactive visual interfaces to this domain. The design goal was to develop an interface that would help comment moderators identify high quality comments from online news discourse, and which would allow us to study and evaluate how new tools can augment comment moderation practices.

Our design study process was inspired by the core phase of the nine-stage framework proposed by Sedlmair et al. [29]. More specifically, our work was organized into four distinct stages roughly matching the discover, design, implement, and deploy stages in that framework: (1) Domain Characterization: (discover) characterizing personas, use-cases, and tasks; (2) Design and Analytics: (design) developing design rationale as well as concrete visual, interaction, and algorithm design for analytics; (3) Prototyping and Implementation: (implement) interface prototypes as well as client-side and server-side components; and (4) In-Field Evaluation: (deploy) validation through domain expert feedback. The following sections discuss each of phases 1,2, and 4 in more detail while we elide details on phase 3 as we rely on standard web technologies for implementation.

THE DOMAIN: ONLINE NEWS COMMENTS

In designing the CommentIQ system we adhered to a user-centered design methodology, including undertaking several early semi-structured interviews with domain experts. In particular, we were interested in developing knowledge from our informants that would allow us to (1) build a persona of a comment moderator that would guide our design thinking, and (2) understanding the use-cases and tasks for comment moderation on news sites. Our interviews were informal and targeted individuals who were embedded in newsrooms and had experience moderating online comments in the context of news. We spoke to eight people from news organizations including the Washington Post, National Public Radio (NPR), the New York Times, and the Wall Street Journal. Six of the interviews were conducted face-to-face, two were conducted via phone, and all were uncompensated and lasted roughly one hour. Copious notes were taken during the interviews and these were typed and analyzed afterwards to facilitate our design process. We approached the interviews with several questions in mind, aiming to understand more about comment moderator workflow and goals, editorial criteria for identifying high quality selections, indications of how moderation decisions were made, and challenges, frustrations, and pain points with current tooling.

Persona Development

Personas can be useful design tools that work by capturing and communicating the objectives, motivations, behaviors, and expectations of a group of target users [5]. They are often useful as grounding artifacts to assess how different design options may impact users. In our case, the personas allowed us as designers to crystallize the important facets of moderator work and effectively communicate among the design team and with other stakeholders. Based on our initial interviews, we developed a persona of The Comment Moderator (TCM), an archetype reflecting our understanding of the comment moderators we interviewed.

Perhaps more than any other goal, TCM is motivated by a desire to produce a quality discussion. They want to not only remove off-topic, impolite, or critical and unconstructive comments, but also to identify and highlight original ideas, cogent points, or contributions that rise above the noise. They are interested in discovering local voices and top contributors to give them a spotlight and to create a feedback loop where good behavior is rewarded

and readers aspire to have their comment selected. Different outlets employ different terminology for this idea: at the New York Times such selected comments are termed "NYT Picks", at the Washington Post they sometimes badge users as "preferred", and at the Wall Street Journal they refer to them as "featured" comments.

The New York Times is perhaps most sophisticated in their thinking about highlighting perspectives as NYT Picks.[1] The NYT Picks are the most popular comment queue and NYT TCMs try to select those with a broad range of viewpoints. In other words, they strive for diversity in the selections. This might include geographic diversity if that is relevant to the story, or it could involve other types of diversity such as along political perspectives. They want the selections to be representative but they do not necessarily need to be balanced between different viewpoints.

The overall goal of TCM is to set the tone and uphold commenting policies to provide a positive atmosphere for discussion. TCM sees comments not as an appendage but as any other piece of journalistic content, and as such they apply an editorial eye: it's not about finding the "most liked" comment. For instance, they strive for fairness in the editorial standards they apply, and are willing to turn comments off for sensitive topics or stories where they believe civil discussion is unlikely. They seek to build trust and increase engagement with their community.

Use-Cases and Tasks

In our conversations with comment moderators, we learned of several use-cases and analytic tasks that may be accomplished with comments. These include: (1) exclusion of low quality comments, (2) selection, highlighting, or picking high quality comments, and (3) taking other journalistic actions based on comment content.

The first task, that of identifying and filtering out low quality comments is one that dominates the analytic workflow for moderators. To a large extent this is about upholding the community standards and providing a venue for discussion that is free from profanity, hate speech, or personal attacks. In many cases, moderators examine flags that are passed in by community moderators, or in some cases by automated systems (e.g., http://keepcon.com/). Sometimes moderators examine the context of a user to see if they are a habitual violator of community norms and as a result may take additional action such as blocking the user. While this is surely an important analytic task which has received some attention in the research literature [31], we instead focus the current work on the underexplored strategy of selecting high-quality comments.

To select high-quality comments for highlighting on a site, moderators consider many different criteria. At the New York Times, they consider five criteria when choosing NYT Picks: overall quality, such as spelling and grammar, argumentation, and literary value; broad representation and diversity of perspective; conversation between two people making opposing points; unexpected short, funny, or unusual commentary; and relevant personal stories and experiences. Interviews with moderators exposed the importance of flexibility and adaptability in applying these criteria. Different quality criteria apply for different stories and communities: there isn't a one-size-fits-all model for when to apply a given editorial criteria, but rather there are many contingencies. Interviews elucidated openness to employing automation to uncover higher quality comments, with an acceptance of some errors and the understanding that a human moderator is making the final decision.

The final task that moderators might engage in involves some other journalistic action, such as correcting a story based on a comment, or passing a comment on to a reporter for follow-up. Several moderators we interviewed believed that comments were valuable leads for news reporting. People often write fascinating stories about how they are personally impacted by an issue at hand, and this can fuel additional reporting by journalists. For niche communities or blogs, insiders may sometimes comment with valuable knowledge and insight that would otherwise be unavailable. This task is conceptually similar to the previous task insofar as it is about identifying comments of a specific ilk, but instead of choosing comments that should be highlighted it is about using the content of those comments for internal purposes. The essential difference in tasks is thus the final step being one of publication, or one of internal use.

DESIGN AND ANALYTICS

We developed several design goals based on our user-centered requirements gathering, persona development, and user task modeling, particularly the second task relating to selecting high quality comments:

- **DG1. Custom ranked list:** The users can customize the ranking based on their own needs and the contingencies of their context, maintaining user agency in the automation process;
- **DG2. Score by multiple criteria:** Comments are scored by several quality criteria that capture different facets of interest in various discussion contexts;
- **DG3. Overview and filter:** To reflect representative but diverse opinions the system should be able to show the distribution of selected comments;
- **DG4. Learning from user feedback:** The system should learn from the user to accommodate specific recurring scenarios in a newsroom.

In the following subsections we describe the design decisions and rationale that support these goals in the analytics and user interface that we developed for CommentIQ. The process that was informed from our requirements gathering is depicted in Figure 1.

[1] http://www.nytimes.com/times-insider/2014/04/17/a-comments-path-to-publication/?_r=0

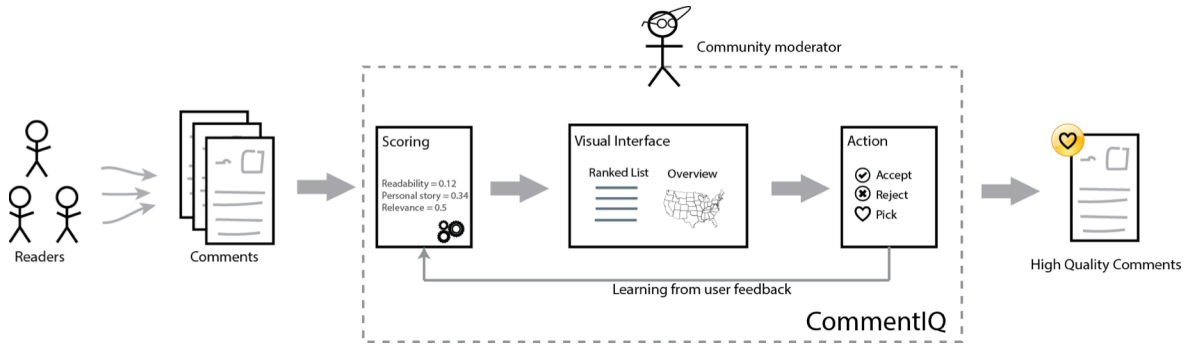

Figure 2. Our design learning suggests the following process for selecting high quality comments: (1) comments submitted by readers are scored using multiple criteria; using these scores, (2) we provide a ranked list of comments as well as the distribution of scores and other meta-data to give an overview and ability to filter; (3) user actions and selection rationale are fed back into the system so that we can learn from users

Analytics

We designed CommentIQ to flexibly support different comment contexts. Thus, instead of using a binary classifier to identify "high quality" comments we employ a customizable weighted ranking of various features (**DG1, DG2**). By allowing the end-user to adjust weights we put the power in their hands to decide when one feature may be more or less salient for their identification of "quality" in that particular context. In addition we provide (1) a smart default weighting for the ranking [10], and (2) several presets to the weighting so that the end-user can quickly switch between contexts.

Selection of Criteria

Instead of computing a large number of textual features and then doing feature selection, we limited our selection of criteria to be legible by humans as motivated by our literature review. Thus we selected criteria that were understandable in this editorial context. This allows us to provide end user customization that is more straightforward (**DG1**). The criteria are based on the content of comments and the history of the user as shown in **Table 1**.

Development of Presets

Tuning of a ranking involves modifying any of 12 weights, one for each criterion. In order to provide a smart starting point for the ranking we trained a classifier to produce a set of default weights. We developed the default weights using a dataset collected via the Times Community API (https://developer.nytimes.com/) which supplies full comment text and metadata such as "NYT Pick" status, and how many recommendation votes it received. The recommendation score was removed during the training of the classifier because a high recommendation score may be a result of a comment being a NYT Pick, rather than the other way around.

Since the final score for the comment is a weighted sum of weights and scores from each criteria, we tested a linear support vector machine (SVM) and a logistic regression. We used 94 'picked' comments and 1574 'not picked' comments. To compensate for the bias in samples, class weight corresponding to the ratio of samples was used. Average precision score using 5-fold cross validation result was 0.13±0.07 with 95 percent confidence interval using linear SVM classifier, and 0.13±0.08 with logistic regression classifier. Average recall for SVM was 0.60±0.39, and 0.60±0.43 for logistic regression. Though these benchmarks are not that impressive, they are adequate to our application since we're interested in providing default settings so that human users can have a useful

Comments Criteria	
Article Relevance	How relevant a comment is with respect to the article, based on word feature vector similarity [8]
Conversational Relevance	How relevant a comment is with respect to preceding comments, based on word feature vector similarity [8]
Length	How long a comment is based on number of words.
Personal Experience	Measures the rate of use of words in Linguistic Inquiry and Word Count (LIWC) categories "I", "We", "Family", and "Friends" [9]
Readability	How readable a comment is according to the SMOG standard index of reading grade level [8,20]
Recommendation	How many recommendation votes a comment has received
User History Criteria	
User Comment Rate	The average number of comments per month a user has written.
User Comment Length	The average comment length score for a user across their entire history.
User Personal Experience	The average personal experience score for a user's comments across their entire history
User Picks	The average rate at which a user's comments are selected as NYT Picks
User Readability	The average readability score for a user's comments across their entire history
User Recommendation	The average recommendation score for a user's comments across their entire history

Table 1. Scoring criteria for comments.

starting point for exploring the rankings and making the final judgment about quality. For both models, weights for the prediction were similar. Readability, article relevance, and user pick history were positively correlated with being an NYT Pick while the conversational relevance and length of comments were negatively correlated with picks. We used the SVM model outputs to create the default ranking.

Other weighting presets were creating using heuristics informed by our interviews with domain experts. We identified the following additional presets that could be useful in different commenting situations:

- **Default:** This preset is for finding generally high-quality comments and was trained using the NYT picks dataset as described above;
- **Personal story:** This sorting favors comments containing personal anecdotes. It is a combination of personal experience and length of comments;
- **Unexpected:** This tries to find short, unexpected comments. This is a combination of short length and high user reputation such as user picks and user recommendation.
- **Best user:** This ranking considers only the user reputation to find a comment written by reliable users.

Interaction and Visual Design

The CommentIQ system is composed of four interface components: article, overview visualization, custom ranking widget, and list of comments. The user can get an overview of comments from the visualization and filter based on making direct selection lassos on the overview visualization (**DG3**). The custom ranking widget provides a way to customize the ranking for one's needs (**DG1**). This section presents the design of these in detail. Figure 2 shows the final interface that was evaluated after soliciting feedback on intermediate designs. The final prototype is online at: http://moderator.comment-iq.com/#/demo

Customizable Ranking View

The goal of the customizable ranking component is to make the custom ranking more intuitive and easy to adjust (**DG1**). The interface is designed for TCMs with various skill levels and goals. At its core is the preset drop-down where the TCM can quickly select weightings for previously identified scenarios, or create a new weighting preset to meet an emerging recognized need. Scoring by different criteria as well as letting the user control the weighting of those scores allows for greater flexibility in the range of contexts where the tool can be employed (**DG2**).

Since our presets cannot cover all of the tasks, we provide a customization so the user can change the weightings of twelve criteria. For example, by giving higher weight to recent posting activity of a commenter, and to user recommendation score, we can surface comments by very active community members with a good reputation.

The different weights are presented with a stacked bar chart, inspired by the multi-attribute ranking visualization by Gratzl et al [15]. It provides a visual signature for each preset as well as feedback to the user about the current weighting distribution.

Overview Visualizations

The goal of the overview visualizations is to show a visual overview of comments according to different scheme, so that moderators can ideally select representative comments along different criteria such as locality, time, or based on quality dimensions (**DG3**). These views can also act as visual filters, where users can select an interesting subset of comments. For example, a user can lasso comments along the west coast on the map view, and only comments from that region will be shown in the comments list. We provide three types of overview explained next.

Map View: This view shows the location of users. We geocoded the location reported by users as free text metadata using the MapQuest Open Geocoding API[2]. This enables the selection of comments from specific geographical regions. Because the location is provided by users as free text, the locations are at various granularities (e.g. state, or city). We applied a force layout algorithm to prevent dots in the same area (e.g. city) from overlapping severely. As a result the map view is suggestive of locations without always being precise.

Commentplot: This view shows scatterplots of criteria scores for all comments. This was intended to provide a quick selection of comments across certain criteria. The axes of the scatterplots were left vague such as "lower" or "higher" because based on design feedback we determined that the relative distribution was more important to show than the absolute score on the overview. These plots show score distributions as well as highlight picked comments to provide feedback to the moderator about where in the score space they have identified quality contributions (**DG3**).

Temporal View: This view shows aggregated scores according to chronological ordering. It can be used for the selection of comments in a specific time window. Also, the change of scores over time such as decreasing article relevance and increasing conversation relevance may be used to make editorial decisions, such as determining when to close the comment functionality on an article.

Brushing and linking between the views is enabled. For example, comment moderators can select the comments from east coast area with long length and high personal story score, and from specific time to find an informative personal anecdote for breaking news for that region.

Learning from User Feedback

Even though comments may be selected by editors for various reasons, the comment data from NYT is annotated simply as a 'Pick' regardless of any of a number of nuanced

[2]http://developer.mapquest.com/web/products/open/geocoding-service

reasons for why it may have been selected. Comments could be selected based on multiple qualifications, such as the presence of personal anecdotes, informative content, or a short and unexpected viewpoint, for example. In some contexts, one feature may have a positive correlation with quality while for others that same feature would have a negative correlation. For example, a comment might be selected either for being short and unexpected, or instead for being long and informative. It may be difficult for a classifier to generalize over the comment length feature.

To address this limitation of the data we have available, we designed CommentIQ to have a feedback loop (**DG4**). When a moderator designates a comment as high quality, they are prompted for information about why it was selected. Along with some predefined options such as "well-written", "informative", "personal experience", "critical", and "humorous" derived from our literature review [9], the moderator can provide free text rationale. When operating at scale the intent is that we could gather selection rationale and correlate the scores and features of those selections to different commenting contexts (e.g. breaking news, different topics or niches). The additional burden of tagging is acceptable because, according to our interviews, only about 1 in 20 comments is selected.

FIELD EVALUATION

We conducted an exploratory evaluation of CommentIQ to gain an understanding of whether and how the tool was helpful to moderators looking to find high quality comments on news articles. We probed users on the tool's utility and asked them to accomplish certain tasks to ensure that they had fully explored system functionality. We piloted the study with a data journalist in order to iterate on the interaction design and UI labeling before we expanded th e study to our target domain users. Based on pilot feedback, we adjusted criteria labels and adjusted the weighting UI to allow for negative weights.

Evaluation Design

We wanted to evaluate whether our approach could be used to enhance current practices as well to know if particular features of CommentIQ were particularly useful or still lacking. We oriented our evaluation according to:

- **Criteria:** Were criteria and the meaning of weights easy to understand? Were they useful in different editorial contexts? Were any criteria missing?
- **Presets:** Was the goal and utility of presets clear to users? Does the algorithm produce results as expected by users? What might be useful presets to add?
- **Rank tuning:** Were users able to create their own custom ranking for their own goals?
- **Overview:** How does the overview and filter approach change the moderation process?

To assess these goals we conducted an in-field evaluation of CommentIQ with domain experts. The following reports our procedure and findings.

Procedure

The study was conducted on a laptop at the place of work of each participant, usually in an area adjacent to the newsroom. After consenting to be in the study, the CommentIQ interface was demonstrated to the participant.

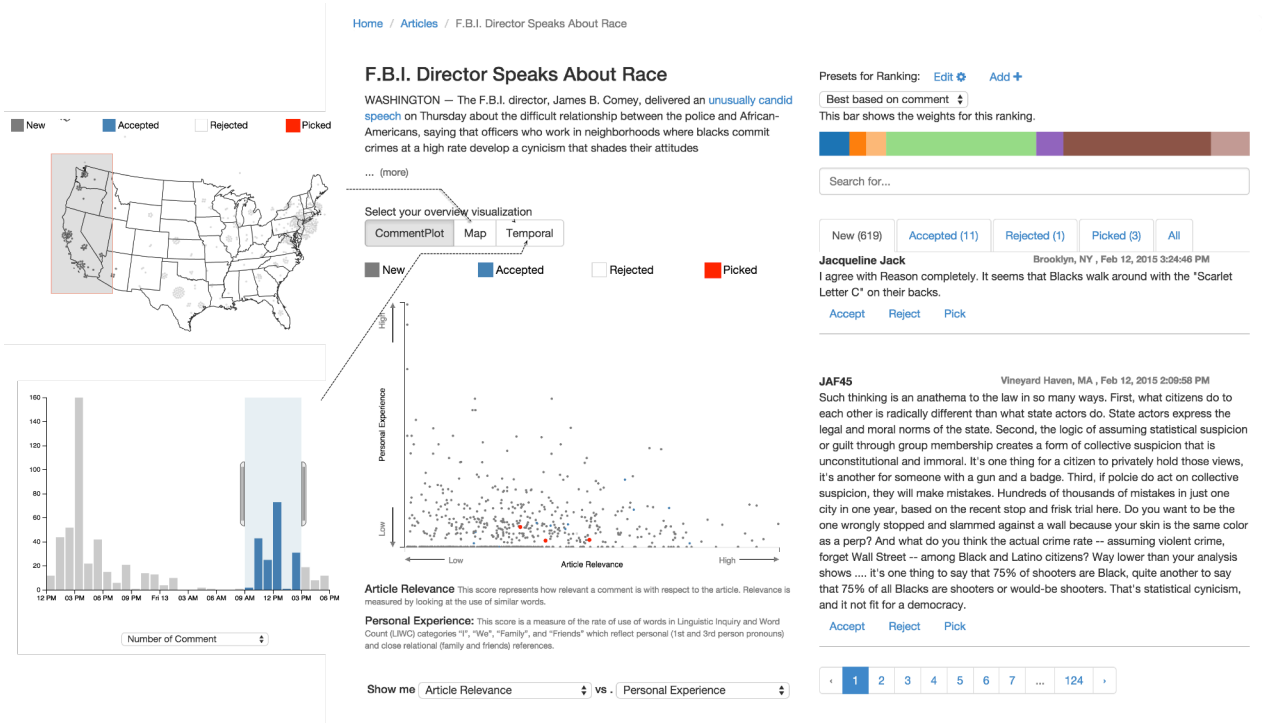

Figure 2. The CommentIQ UI showing toggleable visualizations such as scatterplot, map, and timeline (left) that enable overview and filtering of comments, as well as an adjustable ranking based on various weighted quality criteria (right).

All of the features of the interface were explained, such as how to use and filter on the overview visualizations, how to adjust weights or use the presets on the customized ranking, and the meaning of the various criteria used to score comments. Then the participants were given five minutes for free exploration so they could become acquainted with the system and ask questions if something was not clear. We asked participants to speak aloud as they used the tool. Then we asked users to conduct specific tasks such as using single or multiple criteria to adjust the rankings and weights, as well as using the commentplot and map for selection and filtering. Finally, users were asked to set a goal for target comments that they were interested to find, and then used CommentIQ to pursue finding those comments. After these tasks, we finished the session by asking for the user's general impression and opinion about the advantages and disadvantages of using system. The sessions were audio recorded and transcribed for analysis. Later we summarized important quotes to pull out themes that emerged. Sessions lasted 33 to 75 minutes (M=58).

Content

We made use of three datasets for our evaluation: one for demo purposes, and two for the journalists to interact with. We selected NYT news articles from a range of topics and with varying numbers of comments collected via the NYT Community API. One article, "What Is the Next 'Next Silicon Valley'?" (147 comments), was selected because we thought that it would elicit comments from different geographies that might be interesting for the moderators to consider on the map visualization. Another article, "Who Spewed That Abuse? Anonymous Yik Yak App Isn't Telling" (848 comments), was chosen because we thought that it would elicit comments that included personal stories or perspectives, and it had a much larger and more challenging number of comments. Finally, we chose "F.B.I. Director Speaks About Race" (634 comments) as a demo, since race relations are a hot topic in the U.S. and we thought that comments here might reflect different perspectives that the moderators might want to highlight.

Participants

We evaluated CommentIQ with working professional journalists who have direct responsibilities for comment moderation as part of their duties. Such professionals have knowledge of the real challenges in comment moderation, and the workflows and editorial criteria associated with evaluating online news comments. Combined with the in situ setting in which the study took place, this allows for more ecological validity to reflect on how the tool may be useful in practice. We recruited participants by soliciting industry contacts and asking for referrals. In all, we recruited seven participants (five male, two female) from local (Baltimore Sun) and topical (Wall Street Journal) as well as national (New York Times, Washington Post) outlets. The moderation workflows of the different outlets provided some variability and diversity for the evaluation.

Our participants come from some of the most respected news outlets in the U.S. and are leading-edge practitioners. Their titles and roles include community moderators, community manager, a director of audience engagement development, and a director of digital news projects. Table 1 shows affiliations and related work experience. We will refer to them as 'P1' or 'P2' etc. when citing their comments in the rest of the paper.

Findings

In general participants were positive about the approach and capabilities of CommentIQ. P4 stated that CommentIQ is a great improvement to his existing interface. P3 stated that it is a much more sophisticated and powerful tool than anything she has ever used for comment moderation. The ability to find personal anecdotes quickly and to select comments based on geographic regions received especially positive responses. This reaction was observed from outlets with both pre-moderation and post-moderation processes. Currently, most outlets are just filtering out bad comments, but this is due to resource constraints rather than resistance to selecting and promoting high-quality ones. Participants anticipated that CommentIQ would enable them to find high-quality comments, which is currently not supported by tools that assist in removing low quality comments.

Utility of Analytic Criteria

Participants were able to understand the meaning of the various criteria provided and use multiple criteria to find interesting comments for their goals. Article relevance was used frequently as a comment quality qualification: "I am particularly interested in comments that are on topic," noted P2. And P7 suggested that conversational relevance was also of importance. P1 made an argument that a lower article relevance might also at times be interesting because it is not repeating the article but offering a fresh viewpoint.

The ability to rank comments based on their reflection of personal stories and experiences was met enthusiastically by participants (P1, P3, P4, P5, P6, P7). P3 noted that, "for me personal experience is what makes a productive comments section," underscoring the interest that journalists have in surfacing personal experiences that are

ID	Organization	Field-experience (in years)	Workflow
P1	Washington Post	10	Post-moderation
P2		1	
P3		4	
P4	New York Times	4	Pre-moderation
P5		7	
P6	Wall Street Journal	4	Post-moderation
P7	Baltimore Sun	7	Post-moderation

Table 2. Experience and affiliations of participants.

reported via comments. Initial use of the personal experience score in CommentIQ was hindered because the score counts the frequency of words such as 'I' or 'we', resulting in short one line sentences with 'I' ranked higher. However participants quickly developed various strategies to remedy this, such as selecting comments of a certain length from the CommentPlot (P1, P6) or giving a higher weight to the 'Length' criterion to customize the ranking (P5). For instance, P5 was able to successfully create his own ranking for personal stories. He gave a higher weight to the personal experience criterion, and to the user picks and user recommendation score to select good people. P3 was also able to find good comments for reporters to follow-up on using a combination of the map and a ranking emphasizing conversational relevance and user reputation.

The readability score of comments received mixed reviews. Many participants suggested readability was an important measure, but the result of the current algorithm did not match expectations (P1, P4, P6) as the label did not match their perception of the results they saw. The readability was computed as the SMOG index [21], which is a measure of the grade-level difficulty of the text. However, the score requires further refinement to represent general readability in editorial contexts. As readability features, P1 and P4 suggested the proper use of paragraphs in a comment would make long comments more readable. Sometimes bad comments use meaningless sophisticated language without any hierarchy, resulting in a dense block of text. P6 suggested similar criteria, such as use of 'long ellipsis' or weird punctuation or spacing as features to consider.

Presentation and Visualization

The map view received a lot of positive feedback as a key benefit of the system, with many potential usage scenarios suggested. For example, P1 could find interesting comments from a specific region using the map view. P4 suggested its use to find personal stories via comments from the geographic area where the news is coming from. P6 suggested that the map view could be useful for sport articles, where people are often passionate about their home or school team. Participants compared opinions of Silicon Valley with Florida in the article about Florida being the next Silicon Valley (P1). Also, some users tried to compare conservative states with progressive states, using geography as a proxy for political perspective (P7) and to look for geographic diversity (P5). It was suggested that the granularity of the map might be made flexible to suit different demands: the local outlet wanted a more local map, while national outlets also wanted a global map (P3, P5, P6), with mechanisms to quickly select sub-regions (e.g., a state) using a single click. Geographic information could be improved in the future by using IP geocoding or by extracting location mentions from comments themselves rather than just use metadata.

In the tasks that used CommentPlots, people could select different criteria to get an insight from the scatterplots. Many people used personal experience (P4, P5) and readability (P4) to find interesting comments. P1 successfully used highly relevant but short comments for reporters to quickly find a quote to dress up their story with reactions from users. This shows the adaptability of our approach of leveraging various criteria to accommodate different use contexts.

Elaborate usage scenarios for the temporal view were also suggested. P5 suggested its use for breaking news items, where the ability to find comments from a certain time window can be useful to detect a change of tone or information. Because people do not want to read comments about out-dated news, moderators can put their limited resources on more recent ones. P1 suggested the use of temporal information for getting changing responses of readers as sports events evolve. For example, when there is a new score for an ongoing football game, the sentiment of readers might change accordingly. P6 suggested that the temporal view can be useful when multiple moderators are working together in coordination. P7 articulated that the first wave of comments might contain more personal anecdotes of the event and the last wave of comments will contain more diverse viewpoints about the issues.

The presentation of comments in the ranking view was straightforward but several participants made it clear that threading information should have been reflected there (we lacked this information which was not available from the NYT API at the time of collection). A comment's context within the discussion can be an important element to a moderator's decision as P1, P3, and P4 indicated.

Journalistic Sourcing and Other Use Cases

Several participants emphasized the idea that quality comments were just an entry point for the even more important journalistic goal of identifying potential sources. There was interest in understanding user reputation (P3), repeat commenters and commenting frequency in general (P6, P7), history of negative behaviors like bans, profanity use, or flag history (P3, P6), expertise (P3), and better search access to user comment history. We found that user activity scores, such as average user picks or average user recommendations were frequently combined with comment-based features as a strategy to find reliable people (P5, P6). P7 suggested it would be useful to identify thoughtful users and their expertise and use these to produce future story and article ideas. Our participants thus saw the process of comment moderation not only in terms of the content they were evaluating, but as the potential to identify and engage with potentially new sources. Better user profiling based on past commenting history could thus open up opportunities for new reporting practices.

Some participants expressed that the current tools for comment moderation, which are usually based on chronological or recommendation-based scores, are sub-optimal (P6, P5). Some features of CommentIQ, such as the map-based view and the sorting presets feature may also

have value if presented directly to online commenters (P3, P5). P3 stated that, as a community manager, she wants readers to stick around, but the current comment list they use might scare them off. Given the ability to sort comments according to quality might make them stay longer. P5 articulated that he could program some presets for reporters so that reporters could use the intelligent sorts to find interesting content that may inform them. CommentIQ would have utility for longer enterprise stories where there are a surfeit of comments as well as time, resources, and interest in doing follow-up stories (P2).

DISCUSSION AND FUTURE WORK

Our evaluation of CommentIQ along the dimensions of scoring criteria, weighting presets, easy tuning, and overview visualizations showed that the visual analytic system that we designed was productive and useful for comment moderators. Users were able to effectively understand the criteria, compose them into conjunctions of scores that were editorially meaningful, and tune the results to provide access to higher quality comments of editorial interest. The visualizations proved useful for providing an overview of the comment score distribution and helped orient users towards comments according to geography, analytic scores, and temporal characteristics. The evaluation largely validates our design goals **DG1**, **DG2**, and **DG3**. Assessing **DG4** would require a longer-term deployment to study whether the approach can yield enough feedback data to train better models, and is left for future work.

Our evaluation showed that CommentIQ supports a transformational change to the moderation process. One of the more surprising results from our evaluation was the extent to which comment moderators were ready to begin thinking of moderation not as a policing function, but as a first-class editorial position in the newsroom. As P5 explained, CommentIQ positively changed the moderation workflow by "*shifting moderating to a reporting research job.*" It changes the role of moderators to editorial knowledge work, because they now think in terms of the qualifications for comments they are looking to publish or use. CommentIQ allows moderators to set up hypotheses and run experiments on presets for the workflows that work well in different contexts and for different types of journalism. Moderators can then publish or share that knowledge with others in the form of custom ranking presets, either internally with other moderators or reporters, or indeed ultimately also with their readers.

By framing comment selection as something that is done to identify interesting content to be published, our users articulated new use cases for comments. Of particular interest is that participants wanted to understand more about the *people* who had written various comments, including aspects such as profession or background knowledge of the commenter. Treating the comments as content, and the commenters as potential journalistic sources, opens new possibilities for leveraging comments in journalistic practices. Future work on deriving user models or profiles based on past commenting behavior and content would allow journalists to tap into comments in new ways.

Through our evaluation we found that journalists (and journalism as a domain) require analytic solutions that place humans in a flexible sensemaking loop with the analytics. Visual analytics is thus a well-suited approach for design in this domain. Journalists do not want editorial decisions automatically made for them per se, but rather seek designs that enable and enhance their own decision-making functions so that they can adapt to new situations and contexts and apply human judgment to editorial decisions.

As an endeavor in value-sensitive design for journalists, CommentIQ necessarily embeds biases and institutional preferences in its representations and analytics. But as a visual analytics system, CommentIQ involves humans in the loop and can thus help offset such algorithmic biases. Moderators can tweak and tune the ranking by re-weighting factors to account for variable contingencies and contexts of work. CommentIQ also visualizes moderation choices within the interface, allowing moderators to get an overview of possible biases in their selections.

While analytic scores were selected to reflect editorial quality, any single score has its limitations. For example, readability was measured using the SMOG index, which likely reflects the educational background of the commenter and could thus be used to privilege certain educated voices. As discussed in the previous work [9] in more detail, each score must be used carefully while critically considering limitations. In a future system it would be interesting to examine the potential for different news organizations to plug in their own scores for measuring different dimensions of quality, thus enabling a wider operationalization of editorial interests than we have currently explored.

Future work in this domain might be oriented towards developing new and better quality scores, including dimensions such as novelty, criticality, and thoughtfulness, [9]. Additional work on natural language processing and analytics needs to develop such metrics to be understandable and useful to moderators. Based on our initial interviews, we identified the desire to maintain diversity and balance in the selected comments. While CommentIQ does provide this to some extent, moderators in our evaluation were looking for minority opinions in terms of sentiment, political thought, or even religious affiliation. CommentIQ can provide limited access to this, such as by proxying political thought to a map view, but the current system lacks the ability to show these dimensions in terms of semantic analysis such as political position. Future work could integrate such advanced analytics.

ACKNOWLEDGEMENTS

We thank our participants for their time and insights. And we thank the John S. and James L. Knight Foundation, which supported this work through a prototype fund grant.

REFERENCES

1. Ashley A. Anderson, Dominique Brossard, Dietram A. Scheufele, Michael A. Xenos, and Peter Ladwig. 2014. The "Nasty Effect:" Online Incivility and Risk Perceptions of Emerging Technologies. *Journal of Computer-Mediated Communication* 19, 3: 373–387. http://doi.org/10.1111/jcc4.12009
2. Saeideh Bakhshi, Partha Kanuparthy, and David A. Shamma. 2015. Understanding Online Reviews: Funny, Cool or Useful? *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, ACM, 1270–1276. http://doi.org/10.1145/2675133.2675275
3. Dirk Brand and Brink Van Der Merwe. 2014. Comment classification for an online news domain. Retrieved January 18, 2015 from https://scholar.sun.ac.za/handle/10019.1/96148
4. M. Brehmer, S. Ingram, J. Stray, and T. Munzner. 2014. Overview: The Design, Adoption, and Analysis of a Visual Document Mining Tool for Investigative Journalists. *IEEE Transactions on Visualization and Computer Graphics* 20, 12: 2271–2280. http://doi.org/10.1109/TVCG.2014.2346431
5. Dan Brown. 2006. *Communicating Design: Developing Web Site Documentation for Design and Planning*. New Riders Publishing, Thousand Oaks, CA, USA.
6. Kevin Coe, Kate Kenski, and Stephen A. Rains. 2014. Online and Uncivil? Patterns and Determinants of Incivility in Newspaper Website Comments. *Journal of Communication* 64, 4: 658–679. http://doi.org/10.1111/jcom.12104
7. Kushal Dave. 2004. Flash Forums and ForumReader: Navigating a New Kind of Large-Scale Online. *Discussion,'' Proceedings of the ACM Conference on Computer Supported Cooperative Work*, 232–241.
8. Nicholas Diakopoulos. 2015. The Editor's Eye: Curation and Comment Relevance on the New York Times. *Proc. Conference on Computer Supported Cooperative Work (CSCW)*.
9. Nicholas Diakopoulos. 2015. Picking the NYT picks: Editorial criteria and automation in the curation of online news comments. *the official research journal of international symposium on online journalism* 5, 1.
10. Nicholas Diakopoulos, Stephen Cass, and Joshua Romero. 2014. Data-Driven Rankings: The Design and Development of the IEEE Top Programming Language News App. *In Proceedings of the Symposium on Computation + Journalism*.
11. Nicholas Diakopoulos, Munmun De Choudhury, and Mor Naaman. 2012. Finding and Assessing Social Media Information Sources in the Context of Journalism. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, 2451–2460. http://doi.org/10.1145/2207676.2208409
12. Nicholas Diakopoulos and Mor Naaman. 2011. Towards Quality Discourse in Online News Comments. *Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work*, ACM, 133–142. http://doi.org/10.1145/1958824.1958844
13. Bassey Etim. 2014. A Comment's Path to Publication. *The New York Times*. Retrieved August 8, 2015 from http://www.nytimes.com/times-insider/2014/04/17/a-comments-path-to-publication/
14. Siamak Faridani, Ephrat Bitton, Kimiko Ryokai, and Ken Goldberg. 2010. Opinion Space: A Scalable Tool for Browsing Online Comments. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, 1175–1184. http://doi.org/10.1145/1753326.1753502
15. Samuel Gratzl, Alexander Lex, Nils Gehlenborg, Hanspeter Pfister, and Marc Streit. 2013. LineUp: Visual Analysis of Multi-Attribute Rankings. *IEEE Transactions on Visualization and Computer Graphics* 19, 12: 2277–2286. http://doi.org/10.1109/TVCG.2013.173
16. Enamul Hoque and Giuseppe Carenini. 2015. ConVisIT: Interactive Topic Modeling for Exploring Asynchronous Online Conversations. *Proceedings of the 20th International Conference on Intelligent User Interfaces*, ACM, 169–180. http://doi.org/10.1145/2678025.2701370
17. Chiao-Fang Hsu, Elham Khabiri, and James Caverlee. 2009. Ranking Comments on the Social Web. *Proceedings of the 2009 International Conference on Computational Science and Engineering - Volume 04*, IEEE Computer Society, 90–97. http://doi.org/10.1109/CSE.2009.109
18. Karin Wahl Jorgensen. 2002. Understanding the Conditions for Public Discourse: four rules for selecting letters to the editor. *Journalism Studies* 3, 1. http://doi.org/10.1080/14616700120107347
19. Cliff A.C. Lampe, Erik Johnston, and Paul Resnick. 2007. Follow the Reader: Filtering Comments on Slashdot. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, 1253–1262. http://doi.org/10.1145/1240624.1240815
20. Cliff Lampe and Paul Resnick. 2004. Slash(Dot) and Burn: Distributed Moderation in a Large Online Conversation Space. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, 543–550. http://doi.org/10.1145/985692.985761
21. G. Harry Mc Laughlin. 1969. SMOG Grading-a New Readability Formula. *Journal of Reading* 12, 8: 639–646.
22. Annie Louis and Ani Nenkova. *What Makes Writing Great? First Experiments on Article Quality Prediction in the Science Journalism Domain*.
23. Kathleen McElroy. 2013. Where Old (Gatekeepers) Meets New (Media). *Journalism Practice* 7, 6: 755–771. http://doi.org/10.1080/17512786.2013.774117
24. Kevin K. Nam and Mark S. Ackerman. 2007. Arkose: Reusing Informal Information from Online Discussions. *Proceedings of the 2007 International ACM Conference*

on Supporting Group Work, ACM, 137–146. http://doi.org/10.1145/1316624.1316644
25. Emily Pitler and Ani Nenkova. 2008. Revisiting Readability: A Unified Framework for Predicting Text Quality. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, 186–195.
26. Vladimir V. Putin. 2013. What Putin Has to Say to Americans About Syria. *The New York Times*. Retrieved August 8, 2015 from http://www.nytimes.com/2013/09/12/opinion/putin-plea-for-caution-from-russia-on-syria.html
27. Bill Reader. 2007. Air Mail: NPR Sees "Community" in Letters From Listeners. *Journal of Broadcasting & Electronic Media* 51, 4: 651–669. http://doi.org/10.1080/08838150701626529
28. Raz Schwartz, Mor Naaman, and Rannie Teodoro. 2015. Editorial Algorithms: Using Social Media to Discover and Report Local News. *Ninth International AAAI Conference on Web and Social Media*.
29. M. Sedlmair, M. Meyer, and T. Munzner. 2012. Design Study Methodology: Reflections from the Trenches and the Stacks. *IEEE Transactions on Visualization and Computer Graphics* 18, 12: 2431–2440. http://doi.org/10.1109/TVCG.2012.213
30. Jane B. Singer. 2010. Quality Control. *Journalism Practice* 4, 2: 127–142. http://doi.org/10.1080/17512780903391979
31. Sara Owsley Sood, Elizabeth F. Churchill, and Judd Antin. 2012. Automatic identification of personal insults on social news sites. *Journal of the American Society for Information Science and Technology* 63, 2: 270–285. http://doi.org/10.1002/asi.21690
32. Natalie Jomini Stroud, Joshua M. Scacco, Ashley Muddiman, and Alexander L. Curry. 2015. Changing Deliberative Norms on News Organizations' Facebook Sites. *Journal of Computer-Mediated Communication* 20, 2: 188–203. http://doi.org/10.1111/jcc4.12104
33. Abhay Sukumaran, Stephanie Vezich, Melanie McHugh, and Clifford Nass. 2011. Normative Influences on Thoughtful Online Participation. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, 3401–3410. http://doi.org/10.1145/1978942.1979450
34. Nayer Wanas, Motaz El-Saban, Heba Ashour, and Waleed Ammar. 2008. Automatic Scoring of Online Discussion Posts. *Proceedings of the 2Nd ACM Workshop on Information Credibility on the Web*, ACM, 19–26. http://doi.org/10.1145/1458527.1458534
35. Kevin Wise, Brian Hamman, and Kjerstin Thorson. 2006. Moderation, Response Rate, and Message Interactivity: Features of Online Communities and Their Effects on Intent to Participate. *Journal of Computer-Mediated Communication* 12, 1: 24–41. http://doi.org/10.1111/j.1083-6101.2006.00313.x