Picking the NYT Picks: Editorial Criteria and Automation in the Curation of Online News Comments

Nicholas Diakopoulos

Journalists have a propensity to select online comments for publication according to editorial conceptions of quality content. This work considers various criteria for identifying quality user contributions for publication, evaluates how these criteria manifest in New York Times "Picks" comments, and operationalizes three such criteria computationally. Results indicate that many of the criteria enumerated from the literature do manifest in NYT "Picks" comments more so than non-selected comments, that most criteria are adequately rated by untrained non-professionals, and that relatively simple algorithms can be used to automatically assess some of these criteria. Implications for future online commenting experiences are discussed.

Introduction

The role of online comments on news sites is becoming an increasingly contentious subject as publishers are beginning to challenge the conventional wisdom of providing a space for commentary in response to articles, playing out the tension between the open and participatory nature of user-generated content (UGC) and the norms and goals of professional journalists seeking to control content (Lewis, 2012). Concerns over UGC, and specifically of online comments, by professionals often reflect the potentially damaging effects of low-quality content, such as defamation, brand damage, abusive comments, or injured relationships to community sources (Canter, 2013; Diakopoulos & Naaman, 2011) and many journalists continue to maintain an aloofness and disinterest in engaging with online discussion spaces (Meyer & Carey, 2013). Recently several prominent sites like Re/Code, Popular Science, and Reuters have altogether moved away from having comments on their sites.

Yet there is recognition amongst journalists that audience members can be quite knowledgeable on certain topics; editors see their own role as moderators, filtering user contributions for quality (Hermida & Thurman, 2008). Professionals have thus become more comfortable with employing UGC by imposing their pre-existing news selection processes and styles (Harrison, 2010). A survey of 219 news professionals in Britain found that 90% agreed that journalists' role should be to "filter good information from bad—not to publish anything we get" (Singer, 2010, p. 137). Still, the additional

commercial value of having more content, and the normative goal of facilitating civic discourse come into tension with staffing constraints and the time and resource heavy approach towards engagement and moderation (Harrison, 2010; Singer, 2010; Usher, 2014). This dilemma between a desire for control and maintenance of quality on one hand, and the realities of curation, moderation, and economics that are needed in order to achieve this quality on the other hand, is not easily resolved. Nonetheless, this paper seeks to make progress by first identifying journalistic criteria that are being applied in online comment selection, and then exploring the possibilities for implementing those criteria in computational algorithms that may enable the scalability of comment moderation via automation.

In particular this paper first examines the existence of various editorial criteria that may be applied in the selection of high quality comments by the New York Times as "NYT Picks", and then investigates the extent to which a subset of these criteria can be operationalized computationally in an effort to identify quality comments at scale. Thus this paper seeks to explore the possibilities for maintaining professional journalistic goals and editorial criteria for selecting online comments via automation. This paper contributes (1) a review of the literature on editorial criteria applied to selecting user contributions in different contexts, (2) a crowdsourcing experiment showing that these criteria are manifest in online comments at the New York Times chosen as "NYT Picks", and (3) an examination and validation of computational operationalizations of three of these criteria. This work posits an extension of the notion of robot journalism currently explored in tasks of reporting, writing, and data monitoring (Broussard, 2014; Carlson, 2014; Shearer & Simon, 2014) to the process of comment moderation. The implications of editorial support algorithms in comment moderation for the end-user experience, and for explicit embedding of journalistic criteria into technologies are discussed.

Literature Review

Journalists' concerns over discourse quality and their urge to apply quality criteria to shape and enhance that discourse are not unfounded. Recent studies have shown the potentially detrimental effect of unchecked and uncivil comments, such as polarized risk perceptions of content (Anderson et al., 2014) as well as the prevalence of incivility in online news discourse (Coe, Kenski, & Rains, 2014) and the role that anonymity might play in the quality of discourse that emerges (Santana, 2014). One approach to improving discourse quality that has mounting evidence of effectiveness is to signal norms and expectations for behavior (Jomini Stroud, Scacco, Muddiman, & Curry 2014; Manosevitch, Steinfeld, & Lev-On, 2014; Sukumaran & Vezich, 2011). By modeling and signaling expected behavior and tone in comments, and by cueing users in various ways, user contributions can be modulated in the direction of higher quality discourse. For instance, by having a reporter engage in a news outlet's comment threads on Facebook, (Jomini Stroud et al., 2014) found lower levels of incivility and a greater use of evidence in comments. An experiment by (Manosevitch et al., 2014) found that sticky textual reminder cues within a discourse about Israeli security policy promoted quality of deliberation with respect to issue relevance, expressed opinions, and supporting arguments. Another experiment by (Sukumaran & Vezich, 2011) showed

that thoughtfulness cues in comments led to participants contributing longer comments, spending more time writing those comments, and provided more issue relevant contributions.

Although not widespread, outlets like The Washington Post have dabbled in using a “Post Recommended” badge for outstanding comments, and the focus of this research, The New York Times, has a feature called “NYT Picks” that serves to highlight professionally curated comments. To the extent that new techniques become available to scale up the selection of such high quality material, highlighting that material on a site may be a way to signal expectations, create cues for behavior, and create a virtuous feedback loop for the development of more meaningful and high quality discourse. Although the effects of such cues are not evaluated in the current study, the work explores possible algorithmic approaches to identifying high quality comments that may enable such a strategy at scale by a news outlet in the future. Next, relevant editorial criteria that have been applied to selecting user comments in various journalistic contexts are reviewed.

Editorial Criteria in Comment Selection

Editorial criteria can be applied in at least a couple different ways in moderating an online comment forum. Negative criteria encompass indicators that are used to exclude or otherwise de-emphasize comments from the discourse and have largely been used to buttress against incivilities, like ad hominem attacks, profanity, or other abusive behaviors (Coe et al., 2014) that may emerge in open forums. Technologies have been developed to help cope with the scale of online commenting sections and to aid in the automatic identification of personal insults, profanity, or other inappropriate content (Owseley Sood, Churchill, & Antin, 2012). Some of these techniques are baked into standard comment platforms like Disqus, or are available via third party plugins like KeepCon (http://keepcon.com/). On the other hand, positive editorial criteria can also be applied in an effort to elevate or highlight contributions that moderators determine are worthy. The focus on this paper is on these positive, inclusionary criteria that are applied by journalists in their efforts to editorially shape user-generate content.

In particular let us first review a number of studies in the literature that describe journalistic efforts to identify, curate, and highlight high quality contributions from the public across different forums such as traditional letters to the editor (Wahl-Jorgensen, 2001, 2002), online comments that were remediated for print publication (McElroy, 2013), on-air radio comments (Reader, 2007), as well as purely online comments (Diakopoulos, 2015). A set of 12 criteria that have been reported in the literature across these various contexts includes:

- **Argument Quality**. Reich indicates that argument quality is a dimension along which journalists select for comments (Reich, 2011). Although he does not elaborate on this we might interpret argument quality along the lines of validity in terms of whether a comment expresses a well-grounded and justifiable argument that warrants claims with evidence.

•**Criticality**. "Critical" has been noted as an attribute that was sought by producers at NPR looking to select letters to be read on-air (Reader, 2007). Such constructively critical comments are at times useful as they provide feedback that can lead to factual corrections (Reich, 2011).

•**Emotionality**. The study of traditional letters to the editor sections at newspapers has shown a predilection for "emotionally charged, personal stories of individuals" (Wahl-Jorgensen, 2001). A content analysis study by McElroy also found that from a sample of 309 printed reader comments (selected by editors from online comments), about 45% expressed either a positive or negative tone (McElroy, 2013).

•**Entertaining**. In the current competitive media environment comments and letters can also offer opportunities for readers to engage and be entertained and humored as they are exploring the discourse. In her study of editorial criteria applied to letters to the editor Wahl-Jorgensen found entertainment to be an important dimension, and as one of her interviewees remarked, "some people like a local newspaper basically because of the spiciness of the letters." (Wahl-Jorgensen, 2002). Moreover, studies of comment reading motivations show that a desire to be entertained is a substantial draw for some readers (Diakopoulos & Naaman, 2011).

•**Readability**. The "readability" or more specifically criteria related to the style, clarity, adherence to standard grammar, and degree to which a comment is well-articulated plays a substantial role in editorial selection; "Well written letters are better than poorly written letters" (Wahl-Jorgensen, 2002, p. 77). The degree to which letters were articulate and clear were factors used at NPR in selecting on-air letters (Reader, 2007).

•**Personal Experience**. Personal experiences and perspectives have been shown to be selected by journalists across a variety of contexts. A content analysis of NPR letters that had been selected to air found that 43% of those letters contained "personal observations or historical perspectives from listeners" (Reader, 2007). McElroy's study of online comments that had been selected for printing found that 72% of those selections offered a personal viewpoint (McElroy, 2013). Wahl-Jorgensen posited that the editor's ideology favored concrete personal experiences rather than drawing on abstract ideas (Wahl-Jorgensen, 2001). Broader scholarship relating to deliberation has shown the important role that personal experiences play in strengthening deliberative processes (Manosevitch & Walker, 2009).

•**Internal Coherence**. Oftentimes comments include discussions between multiple contributors who may ask questions or otherwise engage in debate and dialogue (Diakopoulos & Naaman, 2011; Hullman, Diakopoulos, Momeni, & Adar, 2015). However some editors have noted that it can be easier to select a comment for print publication if it is self-contained. It needs to make

sense on its own and so shouldn't refer directly to another comment or to too specific an element of the story (McElroy, 2013).

•**Thoughtfulness**. The degree to which a comment is thoughtful, substantive, and interesting in its expression also plays into editorial decisions in comment curation (McElroy, 2013; Reader, 2007).

•**Brevity**. Brevity is an editorial dimension that emerged to cope with the reality of newspaper production: limited space (Wahl-Jorgensen, 2002). Other media, like radio, have an analogous constraint on time (Reader, 2007). Aside from the limited resource of attention for readers, these other media-specific production constraints do not necessarily hold in the online space however.

Additionally there are editorial criteria that may apply not only to a single comment in isolation, but rather to the context of a set of comments or of a comment in relation to other media like a news article. Curation is not only about the selection of an individual comment, but also of how that comment relates to other selected contributions. The overall gestalt of the selections in a collection can be important. These criteria include:

•**Relevance**. Letters to the editor are often selected because they address issues or events that have already been put on the agenda by the news outlet. In terms of an editor selecting a letter, "Regular citizens' attempts at introducing their own topics to the agenda will almost invariably fail" (Wahl-Jorgensen, 2002, p. 73). More recently, the importance of relevance in editorial selection of online comments has been confirmed in a study of NYT Picks comments which showed that editors' selections were on average 46% more relevant to the news article they were referring to than non-editors' selections, according to a similarity score of word vectors between a comment and article (Diakopoulos, 2015).

•**Fairness**. Issues of fairness in representation of a debate arise when selecting letters to the editor (Reich, 2011), reflecting values about balanced representation of issues (Kovach & Rosenstiel, 2007). This can perhaps be generalized to a notion of diversity amongst perspectives, opinions, or other key demographic dimensions when representing the different voices on an issue.

•**Novelty**. Measures of uniqueness or novelty amongst other contributions are likewise dependent on the overarching context of discourse both on a specific article and amongst different articles, perhaps even across multiple media outlets that are addressing similar news events (McElroy, 2013; Reich, 2011). A key difficulty in operationalization is whether novelty is meant in a local, contingent, or personal reference, or if uniqueness should be understood amongst a knowledge community or even more globally.

In the current study, the focus is on the first nine of these criteria (argument quality,

criticality, emotionality, entertaining, readability, personal experience, internal coherence, thoughtfulness, and brevity) using crowdsourcing and automated content analysis techniques. The last three criteria, including relevance, fairness, and novelty are much more challenging to measure due to their reliance on wider context and relationships amongst media. There has been some previous work on these criteria in related domains. For instance, Diakopoulos has previously shown the importance of relevance in editorial selections of comments (Diakopoulos, 2015) and other research on news articles has considered algorithmic approaches towards selecting, for instance, politically diverse article sets (Munson, Zhou, & Resnick, 2009). However, in the current study the focus is on the initial nine criteria, while the development of more complex methods for the manual content analysis or automated measurement and assessment of the three criteria that are contingent on broader contexts is left for future work. Next a study is presented to examine how the nine criteria apply in the context of online news comments.

Study

Here I consider a specific news site, the New York Times, and the editorial criteria that may manifest in the comments that are published there. In particular, the New York Times has a feature called “NYT Picks” which are a professionally curated set of “the most interesting and thoughtful” comments.(1) These comments are made available in a filtered tab within the interface that sets them apart and labels them as “NYT Picks”. The New York Times pre-moderates all comments on the site, meaning that no comment is published without it first being read by a moderator. This process ensures a generally high quality level for comments since the negative criteria for comment exclusion such as obscenity, personal attacks, or other spam have already been applied.

In particular the manifestation of the nine criteria articulated above is studied in New York Times’ comments, comparing “NYT Picks” comments to non-“NYT Picks” comments. The research questions driving the study are:

RQ1: Do “NYT Picks” comments reflect the positive editorial criteria that have been identified in the literature?

RQ2: Can algorithmic approaches to assessing these editorial criteria be developed?

In order to answer these questions crowdsourced ratings of the various editorial criteria were gathered and automated techniques were used to calculate some metrics, as described next.

Data Collection

Comment data was collected programmatically via the New York Times Community API2 which makes available all of the comments that are published on the site. (2) All comments made in the month of October 2014 were gathered (224,382 in total, including

5,174 “NYT Picks”), including full text of each comment as well as relevant metadata such as whether the comment had been selected by an editor or moderator as an NYT Pick. Data was stored in a MySQL database for further analysis. From the 224,382 comments collected 500 were randomly sampled (250 each from “NYT Picks” and non-“NYT Picks”) in order to arrive at a manageable sample size for the crowdsourcing task described next.

Crowdsourced Ratings

Human ratings of eight of the nine criteria under study (excluding brevity as it is easiest to measure directly and automatically based on text length) were gathered via crowdsourcing. Each of the criteria was rating on a scale from one to five (See Appendix A for the instrument). Ratings were collected using Amazon Mechanical Turk (AMT), a crowdsourcing platform that allows contributors to complete “micro tasks” for small amounts of money. Three independent workers rated each of the 500 comments selected for the study along each of the eight dimensions. The three ratings were averaged to arrive at a final aggregate rating for each comment. Workers were paid 15 cents for each set of eight ratings that they completed for a given comment—a reasonable wage that was determined by considering the average amount of time taken to complete the task in a pilot.

In order to improve the validity of crowdsourced data collection several steps were taken. Studies have shown that it is beneficial to integrate “checks” into the tasks, which force the worker to attend to the content being tagged or rated (Kittur, Chi, & Suh, 2009). In the task, users were asked to supply three keywords that described the content of the comment, thus cueing raters to more deeply process and understand the content of each comment. Amazon also makes available various filters that allow task requesters to restrict who is allowed to complete a task. Workers were restricted to only those who have a reliable history (more than 98% tasks approved) and a substantial history (more than 1,000 tasks completed). Moreover, as cultural context and language ability may be important for interpretation and introduce additional confounds in the reliability of content analysis (Riffe, Lacy, & Fico, 2005) workers were limited to only those that have accounts in the United States or Canada. In the end 1,500 ratings from 89 different workers were collected in about 10 hours.

Despite the lack of training of coders and the wide array of personal levels of knowledge or bias that 89 different coders might have, we find slight to moderate levels of inter-rater reliability. In particular Krippendorff’s alpha was measured for each of the eight crowdsourced criteria using a standard interval measure distance function (Artstein & Poesio, 2008) and find alphas indicating slight to moderate levels of agreement amongst the three raters for all criteria except entertainment; (Argument = 0.32, Criticality = 0.24, Emotionality = 0.16, Entertaining = 0.01, Internal Coherence = 0.20, Personal Experience = 0.22, Readability = 0.21, Thoughtfulness = 0.28). The lack of any reliable signal for the entertaining ratings indicate that individual and subjective personal notions of whether a comment is entertaining may outweigh any underlying generalizable construct as the ratings task was currently framed and defined. The other Krippendorff

alphas suggest that we may draw tentative conclusions based on these ratings. Ratings in the analyses are averaged in order to mitigate noise or subjectivities from the different raters. Moreover, strong claims are not made based on the absolute ratings of any measures, and only general tendencies of aggregate comparisons are considered. As the results will show in the next section, even these relatively blunt measures allow us to expose statistically significant differences between "NYT Picks" comments and non-"NYT Picks" comments.

Automatically Computed Ratings

A longer term research objective, beyond the scope of this paper, is to computationally operationalize all of the above articulated editorial criteria so that they can be automatically applied at scale to help moderators cope with ever increasing numbers of online comments. For now let us set our sights more modestly and consider the computation of three of the nine criteria (Brevity, Readability, and Personal Experience) using relatively simple either off-the-shelf metrics or metrics derived from readily available linguistic resources. The computational operationalization of the remaining six criteria are left for future work.

For brevity, a subjective crowdsourced rating was not collected since the length of a comment can be easily and precisely measured computationally. Using standard natural language processing techniques (Bird, Loper, & Klein, 2009) the full text of each comment in the study is tokenized based on white space and the number of resulting word tokens in the comment is counted. This count becomes the brevity score.

For readability, the crowdsourced ratings are still interesting and useful as subjective assessments of clarity and grammar, but we can also explore a range of readability metrics that have been used in educational settings to automatically score the difficulty of texts including the Flesh-Kincaid score, Gunning-Fog score, Coleman-Liau Index, Automated Readability Index, and SMOG index (McLaughlin, 1969). These scores all attempt to estimate the number of years of schooling that would be needed in order to understand a text and are implemented in open source code that was leveraged.3 Here results are reported using the SMOG index as it was shown to have the highest Pearson correlation *(r* = 0.40) with the crowdsourced ratings. The SMOG index thus becomes the Readability score.

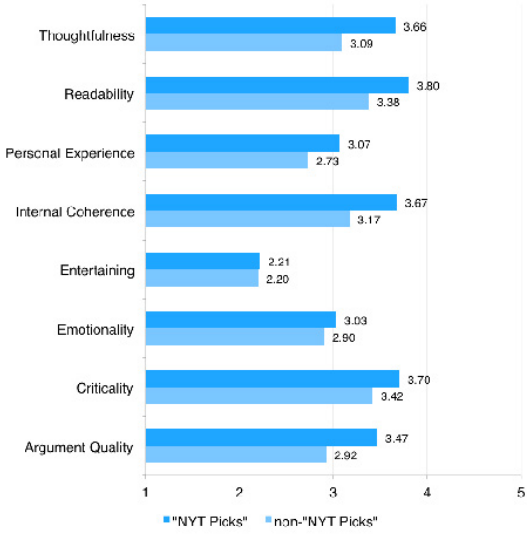
Finally, a new computational operationalization was developed that attempts to score comments based on the degree to which they share personal stories or experiences. The text analysis dictionary LIWC[4] (Linguistic Inquiry and Word Count) was utilized. LIWC is a linguistic resource that is often used in computerized text analysis and has been validated as a way to measure psychologically meaningful constructs by counting word usage in various categories that are defined by dictionaries (Tausczik & Pennebaker, 2010). It was hypothesized that comments which express personal experiences will use more words in LIWC categories "I", "We", "Family", and "Friends" as such terms would reflect personal (first and third person pronouns) and close relational (i.e. family and friends) experiences. Helpfully, the LIWC dictionary also includes colloquial expressions

(e.g. “gf” for “girlfriend” is included in the dictionary), which is well suited to our content domain of casual online communication. The combined dictionary comprises 126 words or word stems (e.g. “acquainta” is the stem of both “acquaintance” and “acquaintances”). Because of the word stems used in the dictionary the Porter stemming algorithm (Porter, 1980) is used in processing the comment text, which is implemented as part of the Natural Language Toolkit (NLTK) (Bird, Loper, & Klein, 2009), and which translates any word into its root stem allowing us to look it up in the dictionary. Each comment is scored by computing the number of stemmed tokens from the comment’s text that are contained in the dictionary, divided by the total number of tokens (i.e. words) in the comment. Thus the score is normalized for the length of the comment. This normalized value becomes the Personal Experience score.

Results

The results presented here address the primary research question of how “NYT Picks” do or do not manifest the various editorial criteria identified in the literature. First let us consider the eight criteria that were rated by crowd workers and compare the ratings for “NYT Picks” and non-“NYT Picks” comments (See Figure 1).

Figure 1. Average ratings of “NYT Picks” and non-”NYT Picks” comments for each editorial criteriathat was rated by the crowdworkers.



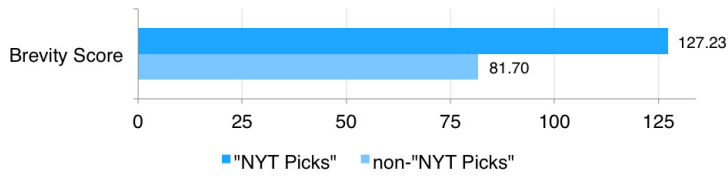
The results indicate that for six of the eight criteria that were crowdsourced (Argument Quality, Criticality, Internal Coherence, Personal Experience, Readability, and

Thoughtfulness) comments that were “NYT Picks” were rated significantly higher than comments that were non-“NYT Picks” (t-tests, $p < .0001$ in all cases). Ratings were anywhere from 0.29 higher on average for Criticality to 0.57 higher on average for Thoughtfulness. In the case of Emotionality a t-test for significance of difference in the means suggests weak evidence that “NYT Picks” were rated higher in terms of their emotion ($t(499) = -1.74$, $p = 0.08$). Finally, the criterion of Entertaining exhibited no statistically significant difference between “NYT Picks” and non-“NYT Picks”, though this is unsurprising as those ratings were not reliable according to the Krippendorff’s alpha that was computed and reported above.

The correlations amongst each pair of criteria show that several of the criteria are highly correlated. For instance, the Argument Quality ratings have a very high Pearson correlation to Internal Coherence ratings ($r = 0.67$), Readability ratings ($r = 0.67$), and to Thoughtfulness ratings ($r = 0.83$). Thoughtfulness and Readability ratings were also highly correlated ($r = 0.70$). The particularly high correlation coefficient for Argument Quality and Thoughtfulness ($r = 0.83$) indicates that these two criteria might be effectively condensed into one scale in future applications of this crowdsourcing task.

The Brevity score, which again is computed as the number of words in a comment, shows that “NYT Picks” comments used on average 127.2 words (SD=72.2), whereas non-“NYT Picks” used far fewer, only about 81.7 words on average (SD=67.4) (See Figure 2). This difference is statistically significant according to a t-test ($t(499) = -7.29$, $p = 1.26 \times 10^{-12}$). Thus “NYT-Picks” comments used about 56% more words per comment than did non-“NYT-Picks” comments, a result that suggests a reversal of the traditional editorial criteria of brevity reported in research on letters to the editor (Wahl-Jorgensen, 2002) or on-air letters (Reader, 2007) and a shift towards editors favoring longer contributions.

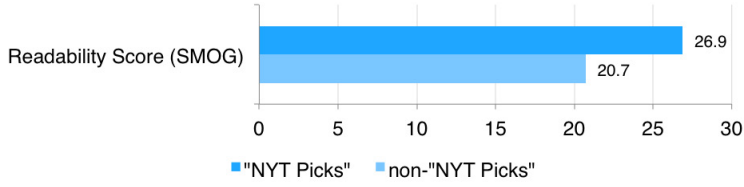
Figure 2. Average Brevity score for “NYT Picks” and non-”NYT Picks” comments, which reflects the number of words in a comment.



The Readability score, which again is the SMOG index or reading grade level of the text, also shows a difference where “NYT Picks” comments are higher (M = 26.9, SD = 8.1) than non-“NYT-Picks” comments (M = 20.7, SD = 9.0) (See Figure 3). This difference is statistically significant according to a t-test ($t(499) = -8.03$, $p = 7.19 \times 10^{-15}$). The Readability score also exhibited a high correlation to the readability ratings collected via crowdsourcing, providing validation of the automated score against human judgments of readability (Pearson’s $r = 0.40$, $p = 1.51 \times 10^{-20}$). These results suggest that the reading

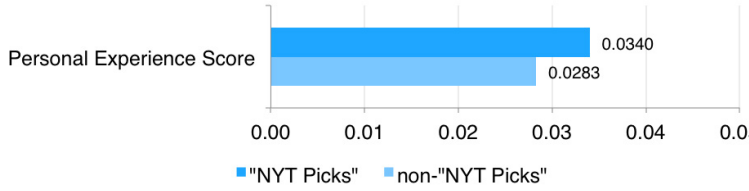
level needed to parse the comments of the New York Times whether selected by editors or not, is very high, requiring years of graduate education. The implications of this finding are considered further in the discussion below.

Figure 3. Average Readability score (SMOG index) for “NYT Picks” and non-”NYT Picks” comments, which reflects the grade level of the text of each comment.



The Personal Experience score, which represents the rate of usage of terms in a set of LIWC dictionaries similarly shows a difference between conditions where “NYT Picks” have a higher average score (M = 0.0340, SD = 0.032) than non-“NYT Picks” (M = 0.0283, SD = 0.033) (See Figure 4). This difference is statistically significant according to a t-test ($t(499) = -1.96$, $p = 0.050$). The Personal Experience score was also highly correlated to the personal experience ratings collected via crowdsourcing, providing validation for the construction of the new metric as a measurement of the rate of usage of terms across key LIWC dictionaries related to personal words and relationships (Pearson’s $r = 0.29$, $p = 3.75 \times 10^{-11}$). Both the Personal Experience score and the Readability score suggest that relatively simple techniques can be used to automatically assess comment texts along editorial criteria that have been shown (by the crowd sourced ratings) to relate to the selection of NYT Picks.

Figure 4. Average Personal Experience score for “NYT Picks” and non-”NYT Picks” comments, which reflects the rate of usage of terms in a set of LIWC dictionaries.



Discussion

The findings presented show that editorial selections as expressed as “NYT Picks” by the New York Times in their online comments do reflect many of the editorial criteria that have been articulated in the literature. With the exception of “entertaining” which was not reliably measured by the crowdsourced ratings apparatus, and with only a weak trend for the “emotionality” rating, the other ratings for argument quality, criticality, internal coherence, personal experience, readability, and thoughtfulness showed reliable and statistically significant differences in average ratings between comments that were “NYT Picks” and those that were not. Moreover, the measurement of a brevity score showed a strong difference in the length of comments that were selected as “NYT Picks”, though not in the direction that the literature suggests. Instead of brevity being a positive criterion, articulated in the literature as a way for journalists to manage limited space or time constraints, in the online space it becomes a negative criterion. Thus, editors at the *The New York Times* preferred longer comments for “NYT Picks”.

The results show that the online comment content of *The New York Times* reflects the application of various professional editorial criteria that have been articulated in other contexts of journalism, such as in selections of letters to the editor. These results mostly support previous observations of the continuity of professional journalistic values as they are carried into online spaces and applied to user-generated content (Harrison, 2010), with the exception of the brevity criterion. Online spaces obviously do not entail the same space constraints of print and thus we observe editorial criteria adapting to allow for longer content in comments online. Perhaps longer comments offer more ground for commenters to express quality and thoughtful arguments. At the same time, the limited resource online is now attention, and a consideration of this reversal of brevity as a selection criteria might be fruitfully pursued from an end-user perspective in future work: Do users prefer reading longer or shorter comments, and how does that interact with their experience of a meaningful discourse?

The slight to moderate inter-rater reliability Krippendorff alphas for the crowdsourced ratings (except for entertaining), also indicate that non-professional human coders with little to no training were able to recognize and apply the various professional criteria used for editorial content selection. This suggests future opportunities for the design of commenting systems that corral and leverage ratings from a community towards the evaluation of comment quality along journalistically important dimensions. For instance, instead of post-moderation of comments only supporting the flagging of comments that break a negative criterion such as spam or incivility (Diakopoulos & Naaman, 2011) community members might also be tasked with tagging or rating comments according to the various positive editorial criteria that have been shown to correlate well with “NYT Picks”. Such a method has been shown to work well in the online Slashdot community (Lampe & Resnick, 2004), as well as in rating other forms of user-generated content like Yelp reviews (Bakhshi, Kanuparthy, & Shamma, 2015), and would represent a shift towards a networked gatekeeping model with the discussion space co-curated by community members (Barzilai-Nahon, 2008). By aligning the dimensions of evaluation of the community with journalistically recognized editorial criteria it may ease the adoption

and acceptance of such an approach by journalists.

The implementation of automated readability and personal experience scores and their validation via correlation to human crowdsourced judgments of those same dimensions offers an exciting new direction and initial demonstration of what may be possible in the future for automated techniques and algorithms that reliably assess journalistically important editorial criteria. This methodology effectively combines manual content analysis methods as a ground truth for assessing the validity of an automated technique (Lewis, Zamith, & Hermida, 2013).

One of the utilities that journalists have found for comments is to identify potential sources for follow-up stories (Hermida & Thurman, 2008). The personal experience score could further enable this by helping to identify comments that are more likely to express information and personal anecdotes that journalists might want to follow-up on as sources, amplifying the value of comments for them. Future work should strive to develop and validate more such automated metrics as in (Diakopoulos, 2015), for example by adapting techniques from computing and information science disciplines such as (Swapna Gottipati, 2012). This will be challenging work and will require not only considering the development of content metrics but also looking at social contexts and user histories as well as the relationships within sets of content in order to consider set-based criteria like novelty or fairness.

Automation of more editorial criteria will raise interesting questions for their deployment and use by journalists, including how algorithmically informed editorial decisions interact with professional norms of control (Lewis, 2012), or redefine labor practices and authority with respect to journalistic practices of reporting (Carlson, 2014; Young & Hermida, 2014), and in this case moderation. New end-user experiences will be enabled while simultaneously reducing the burden of moderation work for journalists, a key concern in the deployment of user generated content by newsrooms (Diakopoulos & Naaman, 2011; Singer, 2010). For instance, end-users might be provided a palette of criteria that they can use and control in order to rank the comments they view. Such an experience would obviate the need for journalists to assess the value of each and every comment, but still provide rankings according to journalistic values and norms for quality content, a value-sensitive design approach (Friedman, Kahn Jr., & Borning, 2006). This would also allow end-users to express different contingent interests in line with a variety of motivations for reading comments as suggested by (Diakopoulos & Naaman, 2011) and adapt their own view of the comments. Again, journalistic norms would set the stage in terms of what metrics are available for ranking and how they are defined computationally, but the end-user would be in control of driving their own experience within that framework.

Adaptability of the criteria that are used to rank comments may also enable different contingent views of the comments within the newsroom itself. For instance, personal experiences, though they may be of high interest as comments on some stories, may be less useful for stories where expertise or cognitive authority is more important. Given research that shows how sourcing practices vary according to different types of stories, based on factors such as proximity (Berkowitz & Beach, 1993) and time demands

(Boczkowski, 2010), an exciting area for future work would be to consider how editorial criteria for comments may also vary across story types or topics. We know that there are some story topics, such as those related to controversial or sensitive social topics that provoke more uncivil dialogue (Coe et al., 2014) and where journalists are apt to want to switch off comments altogether (Diakopoulos & Naaman, 2011). Algorithmic solutions should be sensitive to overgeneralizing across contexts and instead seek to empower users to adapt algorithms for different situations and contingencies.

The readability metric that was applied here shows that whether comments are selected by editors or not (though keep in mind that all NYT comments are pre-moderated), they have a uniformly high reading level, with selected comments being even higher. This raises questions of the broader accessibility and "entrance requirement" to online discourse (Wahl-Jorgensen, 2002, p. 76). A recent post by the New York Times' Public Editor, Margaret Sullivan (Sullivan, 2014) quotes executive editor Dean Baquet as saying, "I think of The Times reader as very well-educated, worldly and likely affluent." The possibility of embedding the editorial criteria of readability into algorithms thus gives the newsroom the power to deeply integrate such top-down perceptions of audience into a generalizable and highly scalable and systematic way to "enforce" the appeal of content to a well-educated audience that can write well-formed, grammatical, and perhaps even eloquent comments. It is here that notions of algorithmic accountability and transparency (Diakopoulos, 2014) become particularly relevant since, as these criteria become conscious and articulated in computer code, so too must the news organization begin to grapple with how to be transparent and indeed apply these criteria ethically. Is it categorically better to select comments that have a high readability? When might this come into tension and conflict with other criteria like fairness or diversity?

Conclusions

In this paper the researcher has explored the manifestation of editorial criteria at play in the selection of comments as "NYT Picks" at the New York Times. The researcher first articulated a set of 12 factors identified in the literature as editorial criteria that have been employed by journalists for selecting user-contributed content in various contexts. A crowdsourcing experiment was then undertaken, which showed that for six of the criteria (Argument Quality, Criticality, Internal Coherence, Personal Experience, Readability, and Thoughtfulness) comments that were "NYT Picks" were rated significantly higher than comments that were non-"NYT Picks. Weaker evidence was found for a difference for the criterion of Emotionality, and no evidence was found for Entertaining as a criterion for selection. But while we find that NYT editors do appear to apply criteria that manifest along many of these dimensions, the results cannot categorically prove that Entertainment is not a criterion for selection under some circumstances. The study is limited to only the New York Times, and to the editorial criteria employed there. It may be that other news outlets would employ humor and entertainment as selection criteria. As such, future work should strive to repeat such a crowdsourcing experiment for other news outlets that also identify and highlight editorially selected comments.

The researcher has also articulated computational operationalizations of three criteria,

including Brevity, Readability, and Personal Experience. The results on Brevity show that journalists in the online space actually select for longer comments, rather than shorter comments as the literature suggested from studies of the print domain. The Readability and Personal Experience metrics show good correlations to the crowdsourced results for those same criteria lending validity to those operationalizations. These results thus suggest that automated technologies leveraging natural language processing might be further explored to computationally operationalize the other editorial criteria identified in this paper. Such developments in technology offer tremendous opportunity for empowering both end-users and journalists in finding new value in online comments, yet we must proceed with caution and consider algorithmic implementations that are adaptable to the myriad contexts encountered across the media.

References

Anderson, A. A., Brossard, D., Scheufele, D. A., Xenos, M. A., & Ladwig, P. (2014). The “Nasty Effect:” Online incivility and risk perceptions of emerging technologies. Journal of Computer-Mediated Communication, 19(3).

Artstein, R., & Poesio, M. (2008). Inter-coder agreement for computational linguistics. Computational Linguistics, 34(4), 555–596.

Bakhshi, S., Kanuparthy, P., & Shamma, D.A. (2015). Understanding online reviews: funny, cool or useful? In Proc. Computer Supported Cooperative Work (CSCW).

Barzilai-Nahon, K. (2008). Towards a theory of network gatekeeping: A framework for exploring information control. JASIST, 59(9), 1493–1512.

Berkowitz, D., & Beach W. D. (1993). News sources and news context: The effect of routine news, conflict, and proximity. Journalism Quarterly, 70, 4-12.

Bird, S., Loper, E., & Klein, E. (2009). Natural language processing with python. O’Reilly Media Inc.

Boczkowski, P. J. (2010). News at work: Imitation in an age of information abundance. Chicago, IL: University of Chicago Press.

Broussard, M. (2014). Artificial intelligence for investigative reporting: Using an expert system to enhance journalists’ ability to discover original public affairs stories. Digital Journalism. Advance online publication.

Canter, L. (2013). The misconception of online comment threads. Journalism Practice, 7(5), 604–619.

Carlson, M. (2014). The robotic reporter: Automated journalism and the redefinition of labor, compositional forms, and journalistic authority. Digital Journalism. Advance online publication.

Coe, K., Kenski, K., & Rains, S. a. (2014). Online and nncivil? Patterns and determinants of incivility in newspaper website comments. Journal of Communication, 64(4), 658–679

Diakopoulos, N., & Naaman, M. (2011). Toward quality discourse in online news comments. In Proc. Computer Supported Cooperative Work (CSCW).

Diakopoulos, N. (2014). Algorithmic accountability: Journalistic investigation of computational power structures. Digital Journalism. Advance online publication.

Diakopoulos, N. (2015). The editor’s eye: Curation and comment relevance on the New York Times. In Proc. Conference on Computer Supported Cooperative Work (CSCW).

Friedman, B., Kahn Jr., P. H., & Borning, A. (2006). Value sensitive design and information systems. In Human-computer interaction in management information systems (pp. 348–372).

Harrison, J. (2010). User-Generated content and gatekeeping at the BBC Hub. Journalism Studies, 11(2), 243–256.

Hermida, A., & Thurman, N. (2008). A clash of cultures: The integration of user-generated content within professional journalistic frameworks at British newspaper websites. Journalism Practice, 2(3), 343–356.

Hullman, J., Diakopoulos, N., Momeni, E., & Adar, E. (2015). Content, context, and critique: Commenting on a data visualization blog. In Proc. Computer Supported Cooperative Work (CSCW).

Jomini Stroud, N., Scacco, J. M., Muddiman, A., & Curry, A. L. (2014). Changing deliberative norms on news organizations' Facebook sites. Journal of Computer-Mediated Communication.

Kittur, A., Chi, E. H., & Suh, B. (2009). Crowdsourcing user studies with Mechanical Turk. In Proc. Conference on Human Factors in Computing Systems (CHI) (pp. 453–456).

Kovach, B., & Rosenstiel, T. (2007). The elements of journalism: What newspeople should know and the public should expect (2nd ed.). Three Rivers Press.

Lampe, C., & Resnick, P. (2004). Slash(dot) and burn: distributed moderation in a large online conversation space. In Proc. SIGCHI Conference on Human Factors in Computing Systems (CHI).

Lewis, S. C. (2012). The tension between professional control and open participation. Information, Communication & Society, 15(6), 836–866.

Lewis, S. C., Zamith, R., & Hermida, A. (2013). Content analysis in an era of big data: A hybrid approach to computational and manual methods. Journal of Broadcasting & Electronic Media, 57(1), 34–52.

Manosevitch, E., & Walker, D. (2009). Reader comments to online opinion journalism: A space of public deliberation. In Proc. International Symposium on Online Journalism.

Manosevitch, E., Steinfeld, N., & Lev-On, A. (2014). Promoting online deliberation quality: cognitive cues matter. Information, Communication & Society, 17(10), 1–19.

Matt Shearer, Basile Simon, C. G. (2014). Datastringer: easy dataset monitoring for journalists. In Proceedings Symposium on Computation + Journalism.

McElroy, K. (2013). Where old (gatekeepers) meets new (media). Journalism Practice, 7(6), 755–771.

McLaughlin, G. H. (1969). SMOG grading - A new readability formula. Journal of Reading, 12(8), 639–646.

Meyer, H. K., & Carey, M. C. (2013). In moderation: Examining how journalists' attitudes toward online comments affect the creation of community. Journalism Practice, 8(2), 1–16.

Munson, S. A., Zhou, D. X., & Resnick, P. (2009). Sidelines: An algorithm for increasing diversity in news and opinion aggregators. In Proc. International Conference on Weblogs and Social Media (ICWSM).

Owseley Sood, S., Churchill, E. F., & Antin, J. (2012). Automatic identification of personal insults on social news sites. Journal of the American Society for Information Science and Technology

(JASIST), 63(2).

Porter, M. F. (1980). An Algorithm for Suffix Stripping. Program, 14(3), 130–137.

Reader, B. (2007). Air mail: NPR sees “community” in letters from listeners. Journal of Broadcasting & Electronic Media, 51(4), 651–669.

Reich, Z. (2011). User comments: The transformation of participatory space. In J. B. Singer, A. Hermida, D. Domingo, A. Heinonen, S. Paulussen, T. Quandt, … M. Vujnovic (Eds.), Participatory Journalism.

Riffe, D., Lacy, S., & Fico, F. (2005). Analyzing media messages: Using quantitative content analysis in research (2nd ed.). Lawrence Erlbaum Associates.

Santana, A. D. (2014). Virtuous or vitriolic: The effect of anonymity on civility in online newspaper reader comment boards. Journalism Practice, 8(1), 18–33.

Singer, J. B. (2010). Quality control: Perceived effects of user-generated content on newsroom norms, values and routines. Journalism Practice, 4(2), 37–41.

Sukumaran, A., & Vezich, S. (2011). Normative influences on thoughtful online participation. In Proc. Conference on Human Factors in Computing Systems (CHI).

Sullivan, M. (2014). Pricey doughnuts, pricier homes, priced-out readers. New York Times.

Swapna Gottipati, J. J. (2012). Finding thoughtful comments from social media. In Proc. COLING.

Tausczik, Y. R., & Pennebaker, and J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. Journal of Language and Social Psychology, 29(1), 24–54.

Usher, N. (2014). Making news at the New York Times. University of Michigan Press.

Wahl-Jorgensen, K. (2001). Letters to the editor as a forum for public deliberation: Modes of publicity and democratic debate. Critical Studies in Media Communication, 18(3).

Wahl-Jorgensen, K. (2002). Understanding the conditions for public discourse: four rules for selecting letters to the editor. Journalism Studies, 3(1), 69–81.

Young, M. L., & Hermida, A. (2014). From Mr. and Mrs. outlier to central tendencies: Computational journalism and crime reporting at the Los Angeles Times. Digital Journalism. Advance online publication.

Endnotes

1. http://www.nytimes.com/content/help/site/usercontent/usercontent.html

2. http://developer.nytimes.com/

3. https://github.com/mikedawson/textstatistics-python

4. http://www.liwc.net/

Appendix A – Crowdsourced Ratings Instrument

Instructions

Below you will see a comment recently published on the New York Times in response to a news article.

> •Read the comment carefully and thoroughly. Make sure that you understand what it means.
>
> •Provide ratings for the comment in the embedded questionnaire below. Please be honest, there are no “right answers”.

Comment

<comment text shown here>

Please provide 2-3 keywords that summarize the comment:

<text entry box here>

To what extent is this comment amusing, entertaining, or humorous?

<Radio button Likert scale 1-5 with 1 labeled “not at all” and 5 labeled “a lot”>

To what extent does this comment express a well-grounded and justifiable argument of high quality?

<Radio button Likert scale 1-5 with 1 labeled “not at all” and 5 labeled “a lot”>

To what extent is this comment well-articulated, clear, and grammatical?

<Radio button Likert scale 1-5 with 1 labeled “not at all” and 5 labeled “a lot”>

To what extent does this comment express emotions such as happiness, sadness, surprise, fear, disgust, or anger?

<Radio button Likert scale 1-5 with 1 labeled “not at all” and 5 labeled “a lot”>

To what extent does this comment express a personal experience, story, or perspective?

<Radio button Likert scale 1-5 with 1 labeled “not at all” and 5 labeled “a lot”>

To what extent is this comment thoughtful and substantive in its content?

<Radio button Likert scale 1-5 with 1 labeled “not at all” and 5 labeled “a lot”>

To what extent does this comment make sense on its own even without the rest of the comment thread or article?

<Radio button Likert scale 1-5 with 1 labeled “not at all” and 5 labeled “a lot”>

To what extent does this comment express a critical perspective?

<Radio button Likert scale 1-5 with 1 labeled “not at all” and 5 labeled “a lot”>

Nicholas Diakopoulos is an Assistant Professor at the University of Maryland College of Journalism. His research is in computational and data journalism with an emphasis on algorithmic accountability, narrative data visualization, and social computing in the news. He received his Ph.D. in Computer Science from the School of Interactive Computing at Georgia Tech where he co-founded the program in Computational Journalism. Before UMD he worked as a researcher at Columbia University, Rutgers University, and CUNY studying the intersections of information science, innovation, and journalism.