The Editor's Eye: Curation and Comment Relevance on the New York Times

Nicholas Diakopoulos
University of Maryland, College of Journalism
nad@umd.edu

ABSTRACT
The journalistic curation of social media content from platforms like Facebook and YouTube or from commenting systems is underscored by an imperative for publishing accurate and quality content. This work explores the manifestation of editorial quality criteria in comments that have been curated and selected on the New York Times website as "NYT Picks." The relationship between comment selection and comment relevance is examined through the analysis of 331,785 comments, including 12,542 editor's selections. A robust association between editorial selection and article relevance or conversational relevance was found. The results are discussed in terms of their implications for reducing journalistic curatorial work load, or scaling the ability to examine more comments for editorial selection, as well as how end-user commenting experiences might be improved.

Author Keywords
News comments; comment curation; computational journalism

ACM Classification Keywords
H.5.3 [Group and Organization Interfaces]: Computer-supported cooperative work.

General Terms
Human Factors; Design

INTRODUCTION
In a sense journalism has always incorporated an aspect of curation: selecting, organizing, and presenting information according to newsworthiness, public interest, or other editorial criteria. More recently, with the launch of services like Facebook Newswire[1], and tools like Storify, the curatorial role of journalists has expanded into the realm of social media, including everything from selecting and verifying user-generated content, to managing and moderating communities and comments on news sites [2].

A strong normative goal of providing accurate and verified information to the public underscores the curation of social media by journalists as a way to promote information quality [4,10], a desire not unlike that expressed in scientific curation communities as well [15]. Fears of comment quality in online news comment discourse [1,5] suggest a desire to improve the quality of the experience for end users both to create a safe space for civic dialogue as well as to create loyalty, host a forum for feedback and tips on news, and increase time on site for news customers [8]. Curating content is important for online communities because it helps set the tone for a site and may in turn improve the quality of subsequent comments on a site [16]. Likewise, unchecked comments that are rife with incivility expose the potential for polarizing risk perceptions [1,11]. By curating comments, journalists model and signal normative commenting behavior, thus making it more salient and visible to others in the community [9], and providing a reward for good behavior through the increased exposure and prominence of a selected comment.

In this paper, I examine the nature of comment curation on the New York Times, in particular as it relates to factors that may enable increased scale for the identification of what are termed "NYT Picks", editor's selections of the "most interesting and thoughtful" comments. While there has been some recent work on computational methods to help moderate and filter out uncivil or profane comments [14], here I instead focus on the other end of the quality spectrum: how to identify *high quality* comments that would be more likely to be selected by an editor. In comparison to related work that has examined different social voting methods for surfacing high quality contributions [18], in this work I focus on using intrinsic features of the content itself to signal quality. Could new computational tools be used to reduce the amount of time journalists need to spend doing this curatorial work, to identify worthy but overlooked contributions, or to scale their ability to consider more content?

Editorial standards for what gets filtered out, or even what constitutes a valid contribution that might be elevated and emphasized often revolve around the relevance of the comment [19]. Off topic or predictable comments have been shown to negatively affect perceptions of quality in online news comments [5].

CSCW '15, March 14 – 18, 2015, Vancouver, BC, Canada

ACM 978-1-4503-2922-4/15/03□$15.00
http://dx.doi.org/10.1145/2675133.2675160

Figure 1. The NYT commenting system, showing the NYT Picks filter and badge.

Other research has found that relevance correlates to "helpfulness" in Amazon product reviews [13], and social votes on Digg [7] and Slashdot [17]. This paper contributes a deeper look at relevance as a correlate to quality and editorial selection, including both a comment's relevance to a news article and its relevance to the broader conversation taking place in the comment thread. The findings are discussed in terms of implications for the optimization and scaling of editorial curation work in the domain of news comments.

STUDY OF NEWS COMMENTING

Platform of Study

Unlike many other online news sites, the New York Times pre-moderates all comments posted on nytimes.com, ensuring a generally high level of discourse that is largely free from obscenity, personal attacks, or other spam. Their policy states that, "most comments will be posted if they are on-topic and not abusive" [19]. The New York Times also has a feature called NYT Picks which represent a professional editorial selection of "comments that represent a range of views and are judged the most interesting and thoughtful." Like many other commenting systems users can vote up a comment by recommending it. Comments are sorted by the newest first by default, but can also be sorted by oldest first, or filtered according to "Readers' Picks" (i.e. high recommendation scores) or by the "NYT Picks". Figure 1 shows the filter interface and the additional attention-getting badge that NYT Picks receive on the site.

Data Collection

Both comment and article full text data were collected and analyzed for the New York Times website (nytimes.com). The Times' Community API[2] was first used to collect comments posted between 1/15/2014 and 4/15/2014. Data collection included full text of all comments as well as relevant metadata such whether it was an editor's selection, the recommended score, and what article the comment was directed at. The NewsCred API[3] was then used to collect full text for article content since this data was not available from the Times API directly. Within the time period, 2,815 articles had full text via the NewsCred API. In the end, 331,785 comments were collected which were directed at one of these 2,815 articles within the collection period. On average there were 118 comments per article (SD=202, Median=32) and a total of 12,542 comments that were an editor's selection.

Data Processing

Full text for the comments and articles was tokenized, normalized, stop-word filtered, and stemmed. Unigrams that occurred 10 or more times across all comments constituted a vocabulary of 22,837 terms that was then used to define a feature vector to describe each comment and article. The reason for limiting the vocabulary was to reduce the influence of spurious and infrequent terms and so to make the feature vectors more robust. Each dimension of the feature vector was then calculated as the term-frequency inverse document frequency (tf-idf) for the term, where each comment was considered a document and term frequencies reflect the stemmed version of tokens. These vectors were then normalized to unit length so that they could be used directly in cosine similarity calculations.

For each comment in the dataset two metrics of relevance were computed: (1) *article relevance*, and (2) *conversational relevance*. Article relevance was computed by taking the cosine similarity score or dot product of the respective normalized feature vectors for a comment and the article to which it is attached. The higher the cosine similarity score, the more similar the comment is to the article. The notion of conversational relevance measures how similar a comment is to other comments on the same article. Only those articles with 10 or more comments were considered in order to ensure that there was enough of a discussion to produce robust feature vectors. To measure conversational relevance, for each article's comments a centroid feature vector was created representing the text of all of the comments on the article that were posted before a given comment. This represents the terms used across the thread up to that point in time. Then, for each comment in the thread its cosine similarity to this centroid representation was calculated in order to measure the comment's conversational relevance.

Results and Analysis

The results presented here address the overriding question of whether the measures of relevance described above could help editors in the process of curating comments. Do simple measures of article or conversational relevance correlate to editor's selections? And can comments be re-ranked to make the curation process more efficient and scalable?

Let us first consider how article relevance relates to whether a comment is selected by an editor or not. The average comment to article similarity for comments selected by editors is 0.183 (SD=0.113), whereas for comments not selected by editors it is 0.125 (SD = 0.106). A t-test between the distributions is highly significant (t = 59.2, p = 0), indicating that those comments that editors do

[2] http://developer.nytimes.com/
[3] http://newscred.com/developer/docs

Figure 2. Article relevance score binned in 0.01 increments versus rate of editor's selections for that bin.

Figure 3. Article relevance score binned in 0.01 increments versus number of editor's selections in that bin.

select tend to be much more on-topic or similar and relevant to the article to which they refer.

The article relevance of the comment is positively associated with a higher chance of it being selected by an editor. Figure 2 shows that increasing levels of article relevance are correlated to higher rates of selection by editors (Pearson $r = 0.92$, $p < 0.001$). A comment with a 0.05 article relevance score has approximately a 2% chance of being selected whereas a comment with a 0.40 article relevance score has closer to an 8% chance.

Figure 3 shows a peak in the volume of editor's selections around an article relevance of about 0.125. Far fewer comments are selected with scores below 0.02. But at the same time, the total volume of comments with low article relevance is very high. Whereas only 1.98% of editor's selections have an article relevance less than 0.02, this represents 12.9% of all comments in the dataset. This suggests a clear opportunity for optimizing professional curator's time: by filtering away comments with extremely low article relevance (i.e. < 0.02) more than one eighth of the comments would not need to be read, in exchange for missing only about 2% of comments that would have been deemed worthy.

Professional editor selections also correspond to the notion of conversational relevance that was computed. The average conversational relevance for comments selected by editors is 0.315 (SD=0.130), whereas for comments not selected by editors it is 0.225 (SD = 0.133). A t-test between the distributions is highly significant ($t = 72.1$, $p = 0$), indicating that those comments that editors do select tend to be much more on-topic and relevant to the other comments in the thread. As with article relevance, there is a very strong correlation between the conversational relevance and the rate of editor's selections (Pearson $r = 0.99$, $p < 0.001$).

Previous research has found a strong correlation between when a comment was posted and how many social recommendations it receives [7], or how much moderation attention it receives [12]. A similar result was found here for editor's selections. Time was measured here as the elapsed time of a comment since the first comment on an article. This was necessary since sometimes article timestamps are updated to reflect edits and measuring time with respect to the article publication time was too noisy. There was a slight negative correlation between elapsed time and whether the comment was an editor's selection (Spearman $rho = -0.048$, $p = 0$). Thus, there are less editor's selections later in the conversation. Also there was an overall negative correlation between elapsed time and article relevance (Spearman $rho = -0.035$, $p = 0$) indicating that comments later in the conversation tend to become less relevant.

Figure 4 shows more detail on the temporal relationship between elapsed time and article relevance within the first 48 hours, when 96.1% of comments are made. Comments made in the first hour have a distinctly higher article relevance than in the immediately subsequent hours. But after about 18 hours the average article relevance begins increasing again up to hour 48 (Spearman $rho = 0.012$, $p =$

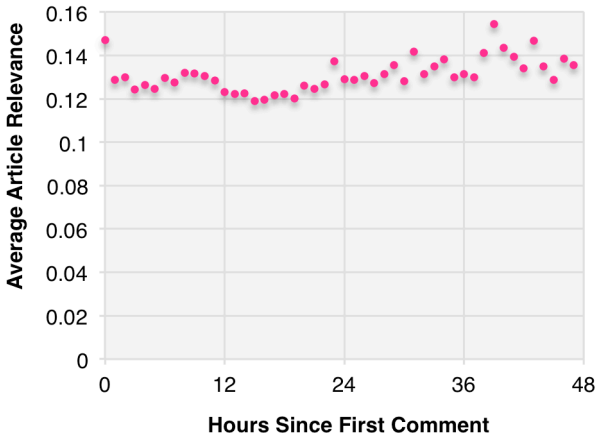

Figure 4. Average article relevance versus time binned into 1 hour increments up to 48 hours.

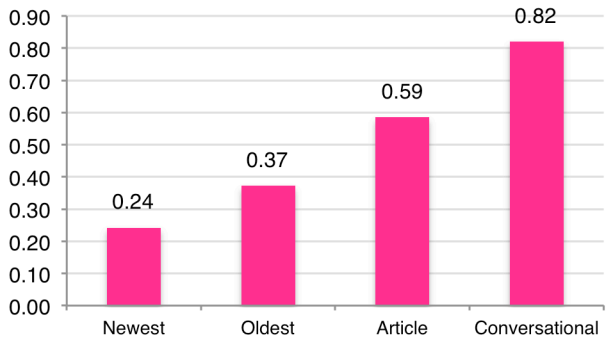

Figure 5. Average number of editor's selections in the top 10 comments when ranked by newest, oldest, article relevance, or conversational relevance.

0). However 83.5% of editors selections are for comments made in hours 2 to 18, and only 7.7% are for comments made in hours 18 to 48. So despite an overall lower likelihood for comment relevance as time progresses in the long tail, these results suggest there are interesting patterns within the first hours and days of an article. Editors might fruitfully direct attention to comments made during the first hour, or between hours 18 and 48 in order to find more relevant comments that may be worthy of editorial selection.

Finally, let us consider here how re-ranking comments by article or conversational relevance compares to the default ranking options of "newest" and "oldest" as this will help assess how other forms of ranking may boost the ability of a curator to find worthy contributions. To measure this the average number of editor's selections in the Top 10 comments was computed for rankings according to these various measures. These results are shown in Figure 5 and indicate that sorting by the newest comments surfaces the fewest editor's selections but that sorting by article relevance or conversational relevance can considerably increase the chances that a curator would quickly see a comment that was worthwhile for them to choose as an editor's selection. Practically speaking, since the conversational relevance score is only computed after there are 10 comments on an article, the article relevance could be used before there are 10 comments and after that the ranking could be switched to the conversational relevance.

DISCUSSION AND IMPLICATIONS

The findings presented here indicate that editors do tend to select comments that are both more on-topic and relevant to the base article, as well as to the other comments on the thread. Moreover it was found that comments made in the first hour in a conversation as well as from hours 18 to 48 of conversation tend to be more relevant. The results suggest that re-ranking comments by article or conversational relevance would allow editors to optimize their time in assessing comments for selection, while possibly surfacing comments in the discussion that might have been overlooked but are still deserving of highlight.

In addition to accelerating comment curation for editors, comment relevance could also be leveraged in the end-user experience of comments. For instance, comments that fall below a threshold of relevance to the article or conversation could be defaulted to a collapsed state in the user interface, so that they're still available but not as emphasized in terms of the screen real-estate they consume. This metric could be applied to entire threads of comments: if the average relevance of the comments responding to a comment is below some threshold then the whole thread could be rolled-up and condensed. Moreover, relevance scores could also be used as real-time feedback to comment authors as they are articulating their ideas as a way to signal to them if they are veering too far away from the conversation.

The goal in the above would be to guide attention and emphasis by applying the editorial criteria of relevance to comments in the interface automatically. Yet we must also consider the limitations of such an automated approach, such as the possible censorship (or at least de-emphasis) of comments that may in fact be related but not detected by the algorithm as relevant due to differences in language use. Future work should also qualitatively examine low-relevance comments including the few there *were* selected by editors in order to better understand what might be lost if such comments were automatically de-emphasized, as well as what other ranking criteria might enable editors to filter for these still-interesting yet "irrelevant" (by the metric) comments.

A further limitation of the current study is that the data available from the NYT API only reflects those comments that were approved for the site. Thus we're not able to reason about editor's judgments amongst comments that were never posted. Future work should extend the current analysis of relevance to other commenting platforms, including those that do not use pre-moderation and which might contain more incivility.

Recent research has also shown how the editorial selections of "network gatekeepers" can diverge in interesting ways from professionals [3,6], suggesting that future work in this domain also examine the differences in editor's selections from "Readers' Picks" in news comments. Understanding where professional editorial criteria diverge or are the same as those of social recommendations could enable the design of commenting systems that better take advantage of both and lead to insights that help integrate professional and user-driven curation activities [15].

REFERENCES

1. Anderson, A.A., Brossard, D., Scheufele, D.A., Xenos, M.A., and Ladwig, P. The "Nasty Effect:" Online Incivility and Risk Perceptions of Emerging Technologies. *Journal of Computer-Mediated Communication 19*, 3 (2014).

2. Bakker, P. Mr. Gates Returns: Curation, community management and other new roles for journalists. *Journalism Studies*, (2014).

3. Boczkowski, P.J. and Mitchelstein, E. *The news gap: When the information preferences of the media and the public diverge*. MIT Press, 2013.

4. Diakopoulos, N., De Choudhury, M., and Naaman, M. Finding and Assessing Social Media Information Sources in the Context of Journalism. Proc. *Conference on Human Factors in Computing Systems (CHI)*, (2012).

5. Diakopoulos, N. and Naaman, M. Toward Quality Discourse in Online News Comments. *Proc Computer Supported Cooperative Work (CSCW)*, (2011).

6. Diakopoulos, N. and Zubiaga, A. Newsworthiness and Network Gatekeeping on Twitter: The Role of Social Deviance. *International Conference on Weblogs and Social Media (ICWSM)*, (2014).

7. Hsu, C.-F., Khabiri, E., and Caverlee, J. Ranking Comments on the Social Web. *Proc. International Conference on Computational Science and Engineering*, (2009).

8. Jomini Stroud, N., Muddiman, A., Scacco, J., and Curry, A. *Journalist Involvement in comment sections*. 2013.

9. Kiesler, S., Kraut, R.E., Resnick, P., and Kittur, A. Regulating Behavior in Online Communities. In *Building Successful Online Communities*. MIT Press, 2012.

10. Kovach, B. and Rosenstiel, T. *The Elements of Journalism: What Newspeople Should Know and the Public Should Expect*. Three Rivers Press, 2007.

11. LaBarre, S. Why We're Shutting Off Our Comments. *Popular Science*, 2013. http://www.popsci.com/science/article/2013-09/why-were-shutting-our-comments.

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

13. Otterbacher, J. Helpfulness in online communities: a measure of message quality. *Proc. Conference on Human Factors in Computing Systems (CHI)*, (2009).

14. Owseley Sood, S., Churchill, E.F., and Antin, J. Automatic Identification of personal insults on social news sites. *Journal of the American Society for Information Science and Technology (JASIST) 63*, 2 (2012).

15. Rotman, D., Procita, K., Hansen, D., Sims Parr, C., and Preece, J. Supporting Content Curation Communities: The Case of the Encyclopedia of Life. *JASIST 63*, 6 (2012), 1092–1107.

16. Sukumaran, A. and Vezich, S. Normative Influences on Thoughtful Online Participation. *Proc. Conference on Human Factors in Computing Systems (CHI)*, (2011).

17. Wanas, N., El-Saban, M., Ashour, H., and Ammar, W. Automatic Scoring of Online Discussion Posts. *Proc WICOW*, (2008).

18. Xu, A. and Bailey, B.P. A Reference-Based Scoring Model for Increasing the Findability of Promising Ideas in Innovation Pipelines. *Proc. Conference on Computer Supported Cooperative Work (CSCW)*, (2012).

19. Help: Comments & Readers' Reviews. http://www.nytimes.com/content/help/site/usercontent/usercontent.html.