The Role of New York Times Picks in Comment Quality and Engagement

Yixue Wang
Northwestern University
yixue.wang@u.northwestern.edu

Nicholas Diakopoulos
Northwestern University
nad@northwestern.edu

Abstract

While various methods can be used to maintain online discussion quality, one aspect that is underexplored is the role of highlights from professional moderators. In this work, we look at the impact of New York Times (NYT) Picks. We present an analysis of more than 13 million NYT comments, examining the quality and frequency of commenting on the site in response to NYT Picks. The findings offer evidence that NYT Picks are associated with an increase in the quality of first-time receivers' next approved comment, as well as the commenting frequency during commenters' early tenure on the site. The quality boost associated with receiving a Pick attenuates after subsequent picks and diminishes over time as the user continues commenting, but is still higher than commenters who don't receive Picks. The findings also indicate that exposure to Pick badges is positively correlated with subsequent higher quality approved comments, albeit to a lesser extent.

1. Introduction

According to a 2016 survey of 1,011 respondents in the U.S., 55% of Americans had left an online news comment, and 77.9% had read news comments at some point [1]. Ideally, comment sections provide a public space and community for users to interact with news and issues, express their opinions, share information, and learn about others' views in a way that supports democratic deliberation [2, 1, 3].

Although that ideal is difficult to reach in practice (e.g. due to incivility, obscenity, or harassment), some news organizations do manage to create a useful and high-quality commenting experience. Approaches include the moderation and interaction of staff, the use of community-based moderation to help flag and report problematic comments, and the development of new technologies that can contribute to semi-automated moderation [4]. At the *New York Times* (NYT), comments are pre-screened by professional moderators and editors before they are published to the comment section, helping to set standards for thoughtful, lively, and civil commentary. In parallel to filtering improper comments, the NYT also moderates its comment sections by highlighting high-quality comments, judged by professional moderators as "the most interesting or thoughtful", and gives these comments a Times Pick badge in the user interface [5]. While one survey found that almost half of commenters would like news organizations to highlight quality comments [6], neither moderators nor researchers fully understand the behavioral impact of such a moderation strategy.

This study analyzes whether this moderation strategy is positively correlated with users' aggregate comment quality and engagement. To enable this analysis, we develop and validate a machine learning model to measure comment quality based on NYT editorial standards across a set of more than 13M approved and published NYT comments. Our results underscore the potential for NYT Picks to encourage a higher quality of communication in online comment communities and draw attention to the opportunity for news organizations to adopt the moderation strategy of highlighting professionally selected high-quality comments to improve overall community quality.

2. Related Work

In this section, we motivate our specific research questions on the impact of NYT Picks using literature on social feedback, social norms, and social control.

In social learning theory, there are four stages in the process of learning: attention (i.e. exposure to and salience of behavior), retention (i.e. memory), reproduction (i.e. ability to perform behavior), and motivation (i.e. will to perform behavior) [7]. People learn through observation of models (i.e. other people) who provide examples of behaviors, and then they repeat those behaviors once their own behaviors' consequences are rewarded (positive reinforcement). They also learn by observing the consequences of others' behaviors

URI: https://hdl.handle.net/10125/70970
978-0-9981331-4-0

(vicarious reinforcement). In online communities, Kiesler et al. suggest that people learn social norms in three ways: observation of behaviors and their consequences, seeing instructions or codes of conduct, and behaving and directly receiving feedback [8].

The goal of feedback is "to reduce discrepancies between current understandings and performance and a goal" [9]. Previous studies indicate that feedback, serving as a form of formative assessment, can impact performance through self-efficacy levels, behavioral rewards, and a self-regulatory system [10]. NYT Picks can work as a feedback mechanism by rewarding positive behaviors with public recognition for the individual [11]. We elaborate how receiving a Pick could work as a form of feedback, and whether there is correlation between receiving Picks and commenting behaviors and engagement in Section 2.1 and 2.2.

Social control theory further highlights how formal and informal actors can enforce social norms in news comment spaces via affirming or sanctioning controls through direct or indirect actions [12]. Professional moderators, as formal social actors, can reinforce positive social norms through affirmative direct (i.e., selecting a pick) and indirect (i.e., showing the Pick badge in the interface) social control. Highlighting social norms can change people's commenting behaviors by making descriptive norms more salient and helping convey to users that good behavior is prevalent [8]. Seeing the consequences of the behavior can make the social norms more learnable [13], thus impacting commenting behaviors [14]. We further discuss whether a positive correlation exists between comment quality and indirect forms of social control in Section 2.3.

2.1. Feedback and Behavioral Change

In the context of online communities, different forms of social interaction can work as positive feedback and may motivate desirable behavioral changes. For instance, the implementation of badges, as a form of positive and public feedback from the community, leads to a behavior change to pursue badges on Stack Overflow [15], which confirms that feedback can reduce the gap between current behaviors and desirable behaviors.

On the other hand, some positive community feedback, quantified by the proportion of up-votes, was found not to encourage rewarded authors to improve the quality of their posts in online news communities [16]. However, research also indicates that formal feedback, such as via awarded points or stars, is more effective than informal feedback, such as informal comments, in helping people adopt desirable behaviors [8]. In addition, feedback from users with more authority was found to have more influence on users' behaviors [17].

Previous work has suggested that highlighting high-quality comments by professionals with high authority, as a form of formal positive public feedback, could create a feedback loop for the development of more meaningful and high-quality discourse [18] and help receivers better learn the desirable behavior. Therefore, we propose the following research question:

RQ1: *Do commenters receiving NYT Picks improve their future comment quality according to professional editorial standards?*

2.2. Feedback and Engagement

Feedback can not only motivate people to adopt desirable behaviors, but it can also prompt people's engagement with online communities: positive feedback from peers was shown to increase newcomers' general work motivation [19] and encourage them to become more active [20] on Wikipedia; receiving a response to one's first posted message motivated ongoing contributions from newcomers [21]; positive numeric ratings prompted newcomers to return faster to post a second comment [21]; and community responses, including direct messaging and direct reply, were correlated with an increased likelihood of posting again and an increase in people's total number of contributions [22, 23]. Whether they receive feedback is a predictor of whether a newcomer increases their sharing when they are initially inclined to contribute [24]. Community feedback, such as introductions referencing previous group participation in conversations, increases reply counts as well [24]. For newcomers, positive community feedback can signal community acceptance and can lead to more engagement in the future. While acknowledging that there may be other fruitful operationalizations of the construct, we define engagement in this study as commenting frequency. Therefore, we propose the following research question:

RQ2: *Do commenters receiving NYT Picks increase their future commenting frequency?*

2.3. Social Norms and Observers' Learning

In commenting communities, people exposed to highly thoughtful behaviors will be more thoughtful in their own comment behaviors [25]. People exposed to a civil discussion will be more civil in their comment behaviors and more willing to participate in discussions compared to those exposed to uncivil discussions [26]. Trolling behaviors can also stem more from negative

discussion context and mood than from an individual's history of trolling [27]. After the exclusion of interaction with trolls and spammers, comment quality improves [28].

NYT Picks may be a way to signal desirable expectations and create cues for behavior from moderators. As a hint of social norms in the comment section, they may therefore help users learn what makes a good comment and may motivate them to write higher quality comments in the future in order to receive a NYT Pick selection and badge. Therefore, we propose the following research question:

RQ3: *Do commenters observing NYT Picks improve their comment quality according to professional editorial standards?*

3. Measurement of Comment Quality

In this section, we focus on developing a methodology for quantifying comment quality based on NYT editorial standards. We begin by introducing the system and the dataset we used in the study: the *New York Times* comment section and NYT Picks. We then describe how we built machine learning models to measure comment quality. We use NYT Picks for building models to quantify comment quality because prior work has shown that NYT Picks comments correlate well with many dimensions of editorial quality [18]. As a result, we use NYT Picks for learning editorial judgment and for building a quality scoring model to study our research questions at scale.

3.1. Data Collection and Characterization

In order to address our research questions, we collected data via The New York Times Community API. In total, we collected 13,213,626 approved comments (rejected comments are not included in the API's data) from December 19, 2007, to October 1, 2015, of which 326,901 (2.5%) were marked as NYT Picks. Due to this data limitation, we focus our research questions only on *approved* comments, a limitation which puts bounds on the scope of our claims and which we further elaborate upon in Section 5.2. We were unable to collect the most recent 4 years of comments because the API has been out of service being re-engineered until very recently. The comments collected were made in response to 221,740 articles (mean=59.6 comments per article; median = 10 comments per article). The full dataset is comprised of comment information including comment content and timestamps for when a comment was posted by a user, approved by NYT moderators, and selected as a Pick, the number of replies and recommendations by others for every comment, article information (article URL and article ID), and user information (user ID and display name).

In total, there were 1,201,646 unique users in this dataset. Overall, users commented 11 times on average (median = 1). We found that 55% of users (661,286) commented only once, while 45% of users (540,360) commented more than once. These repeat users made an average of 23.2 comments (Median = 4). There were 117,701 users (9.8%) who wrote at least one Pick, while 1,083,945 users (90.2%) never had a comment selected as a Pick. Users who received any Picks were selected by NYT 2.8 times on average (Median = 1).

3.2. Measuring Comment Quality at Scale

In order to address our research questions, and in light of the scale of the data collected, we developed a reliable automated method to measure comment quality. Specifically, we define a comment's quality score as the probability of being selected as a NYT Pick, and used a machine learning model to predict this score based on a variety of comment features. We validated model outputs against human ratings of comment quality collected on Amazon Mechanical Turk (AMT).

3.2.1. Building Models In order to quantify various aspects of comment quality, we calculated different features using the comment text. Many of these factors are motivated by prior work examining comment quality criteria in online news comments [18], however the emphasis in choosing these features is on achieving *predictive* accuracy for comment quality as a dependent variable, rather than on theoretical motivations for features.

The features in the model include: word count (i.e., comment length or brevity) [5], lexical diversity (i.e., the ratio of unique words to the total number of words in a text) [29], entropy (a measure capturing the information density reflected in word choices), personal experience score (i.e., the ratio of words in Linguistic Inquiry and Word Count [LIWC] categories "I", "We", "Family", and "Friends" in text) [5], readability (i.e., the SMOG standard index of reading grade level) [30, 5], sentiment [31], including negativity, positivity, and polarity scores, as well as subjectivity scores (i.e., textual indicators of personal feelings, views, or beliefs). We also utilized the Perspective API (version 4 released on August 23, 2017) to extract various quality dimensions of comments. Perspective API [32] is a model trained on data from hundreds of thousands of comments labeled by human moderators to measure dimensions including

toxicity (related to impoliteness [33]), attacks (on the author or commenter) (i.e., incivility [33]), incoherence, being inflammatory, likelihood of rejection (i.e., overall measure of the likelihood for the comment to be rejected according to NYT's moderation), obscenity, spam, and being unsubstantial (i.e., short length [5]). Since September 2016, these scores have been used by NYT to expand their moderation capacity; however, because this happened after the last date reflected in our dataset, there is no risk of feedback into our model. We computed features of comment content using both a term frequency-inverse document frequency (TF-IDF) vector (unigrams, bigrams and trigrams are included) and a document embedding vector with a size of 1000 for every comment (Doc2vec vectors were calculated using gensim.models.doc2vec). In addition to the scores derived from the comment text, we also included features based on the number of recommendations and the number of replies to a comment in order to capture how the community reacted.

We then built machine learning models to predict whether a comment was picked as a NYT Pick using a random sample of 100,000 comments (2,456 NYT Picks) and with a feature vector combining all the features described above. We experimented with several machine learning algorithms (using the sklearn package), including Linear Support Vector Machine, Linear SVM with SGD training, Random Forest, and Logistic Regression. Due to the nature of the imbalanced data (i.e., there are far more non-Pick comments than Picks in the dataset), we also built bagging classifiers on top of these algorithms [34], using 10 base models on the same amount of Picks and a random sample of non-Pick comments from the original training dataset. We tested the performance of the four bagged models on a held-out random sample of 100,000 comments. Logistic Regression achieved the highest AUC score of 0.832 while Random Forest achieved the lowest AUC score of 0.765. Linear SVM had an AUC of 0.781 and SGD had an AUC of 0.822. As described next, we then evaluated the classifiers against quality scores produced by human raters, which helps to further establish the ecological validity of the model we used.

3.2.2. Validity of Model In order to test our models' performance against actual human perceptions of comment quality, we validated our models' scores with quality ratings of comments collected via Amazon Mechanical Turk (AMT). Crowd-sourced ratings of content have been shown to effectively reproduce expert ratings when the quality of work is assured through careful screening and quality control mechanisms [35]. Therefore, AMT workers were screened for qualifications including their historical approval rate ($> 98\%$) and the number of HITs (i.e., Human Intelligence Tasks that AMT workers can work on, submit and collect payment for completing) approved (> 1000). Since comments from NYT may involve U.S. domestic politics and cultural context, we also required AMT workers' location to be the United States to ensure they had the cultural knowledge to rate comments more accurately. We determined the offered wage by calculating the median working time in a pilot task (median = 76s). For the data presented here, workers were paid 0.17 USD for each HIT they finished in order to be above US federal minimum wage of $7.25 per hour according to the pilot task median completion time. 100 Picks and 100 non-Picks were randomly selected from the test dataset for this evaluation so the sample dataset had no reliance on the models' scores. In total, 177 workers were recruited to produce five independent ratings of quality for each comment in the sample.

In each AMT assignment, we were motivated by the NYT commenting guidelines and described quality comments as "usually articulate and clear, well-informed, interesting, thoughtful, and relevant. Other criteria may include: having a well-reasoned argument or expressing first-hand knowledge." In addition to asking workers to rate the comment from low to high-quality on a scale from 1 (low) to 5 (high), we implemented quality control mechanisms to bolster the reliability of ratings [35] by asking them two other questions as checks to ensure they had thought about the comment and to help determine whether or not a rating would be included in our study [36]. Workers needed to provide 2-3 keywords that summarized the comment topically (Q1), along with a short explanation of their rating (Q2). We also used input validation on the page to suggest workers provided more details when the length of their answers for either Q1 or Q2 was too short (i.e., $<$ 6 characters). We also manually excluded data from an assignment when (1) keywords in Q1 did not topically describe the comment (e.g., "interesting" or just copy and pasted directly from a comment), (2) reasons in Q2 were not relevant to the scores (e.g., "okay", "interesting" were not deemed sufficient), (3) reasons provided in (Q2) were too emotional/personal/partisan, (i.e., "uninformed liberal"), or (4) if the working time of a task was substantially lower than others' working time (i.e., lower than the mean minus one standard deviation). If more than half of a worker's HITs were excluded based on the previous criteria, we assumed that the worker didn't fully understand the task according to our instructions, and thus we excluded all ratings they submitted. Regardless of exclusions all workers were

paid, but the excluded HIT was reposted to new workers. These various quality controls resulted in the exclusion of 347 ratings; however, we eventually collected 1,000 approved ratings. On average, approved HITs were finished in 137 seconds (Median = 101s). Because of the discrepancy in the final median completion time and the completion time based on our small pilot that did not include all quality checks, the wage estimate was low. In the future, we recommend running multiple pilots with all quality checks to get a better estimate of task completion time to ensure fairer wages.

Krippendorff's alpha is a reliability measurement of intersubjective agreement among independent ratings [37]. After rejecting and reviewing all the assignments based on the criteria described above, we calculated Krippendorff's alpha as 0.50 for the approved ratings, which indicates a moderate agreement among five independent workers. The average rating for NYT Picks was 3.80 compared to a relatively lower rating of 3.10 for non-Pick comments ($t(998) = 9.69, p < 0.0001$). To reduce variance in the ratings, we filtered the highest and lowest rating from each set of five ratings. We then found that Krippendorff's alpha increased to 0.77, which indicates a substantial agreement among the remaining three workers. The average rating for Picks was then 3.88 ($IQR = (3, 5), s.d. = 1.05$), and the average rating for non-Pick comments was 3.10 for non-Pick comments ($IQR = (2, 4), s.d. = 1.24, t(598) = 9.83, p < 0.0001$). The higher the alpha is, the more reliable the data is in terms of coherence amongst different coders. Since we had a moderate agreement while using all ratings and a substantial agreement when filtering the highest and lowest ratings, the human ratings collected reflect a reliable assessment of quality.

We used the Spearman rank-order correlation coefficient to measure the relationship between the human ratings and quality scores produced by the machine learning models since they are not normally distributed. The scores produced by the bagged Random Forest (RF) model correlated best with normalized average human ratings (scale from 0 to 1). The Spearman correlation coefficient was 0.58 ($p < 0.0001$), demonstrating additional ecological validity of the model with respect to how untrained raters evaluated the comments.

In the rest of this study, we used the quality scores produced by the bagged Random Forest model since the scores correlated best with these human ratings. Across all NYT Picks in the test dataset, the average quality score from the bagged RF model was 0.5044 (median = 0.4805), whereas for non-Pick comments the average score was 0.3552 (median = 0.3310), indicating that the bagged RF model was successful at differentiating quality between the two groups ($t(99998) = 58.71, p < 0.0001; U = 5.8 \times 10^7, p < 0.0001$).

4. Findings

After scoring all the approved comments, we are able to understand in aggregate how comment quality varies under different situations, including receiving NYT Picks corresponding to RQ1 and RQ2, as well as observing others' NYT Picks corresponding to RQ3. Overall, the quality score ranges from 0.1115 to 0.9830, with an average of 0.3560 and a median of 0.3360. The distribution of quality scores for all approved comments is skewed. Many of the comments tend to have lower comment quality scores, while relatively fewer achieve high comment quality scores according to our model. Across all NYT Picks, the average quality score is 0.4993 (Median = 0.4745), whereas for non-Pick comments the average score is 0.3523 (Median = 0.3335). In the following analyses, we report median comment quality scores in order to be more robust to the skew in the distribution.

4.1. Receiving NYT Picks

The following subsections address RQ1 and RQ2, which relate to the impact of receiving NYT Picks. We focus on users who receive at least one NYT Pick (N= 117,701) and analyze the correlation between NYT Picks and user commenting quality and frequency.

4.1.1. Do commenters receiving NYT Picks improve their future comment quality? Here we are interested in how comment quality evolves for users over time with respect to the Picks they receive. We first examine this evolution by looking at the median comment quality across users' commenting history. To do this, we plot the order within each user's comment history and the median comment quality score for all comments of that order (See Figure 1). We see that the median comment quality tends to be highest for the first approved comment that users write. The median comment quality then decreases as the commenters interact more with the comment section. We observe a similar decreasing trend regardless of the number of comments a user has made (i.e., frequent commenters exhibit the same trend as less frequent commenters).

To assess the impact of receiving positive professional feedback, we compare all approved comments before and after a user's first Pick to evaluate how quality may have changed for users who receive at least one Pick (N=117,701).

To reduce the possible impact from confounders that

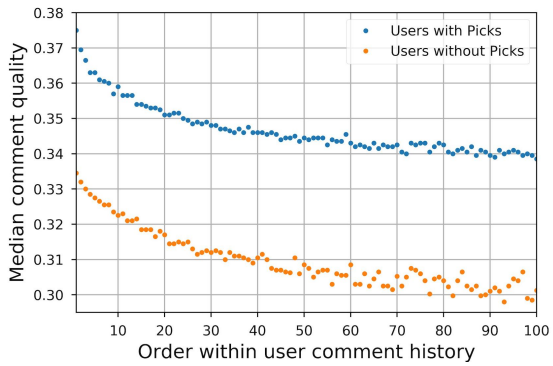

Figure 1. User comment quality in users' history. For every order, the data contains all commenters who commented at the time.

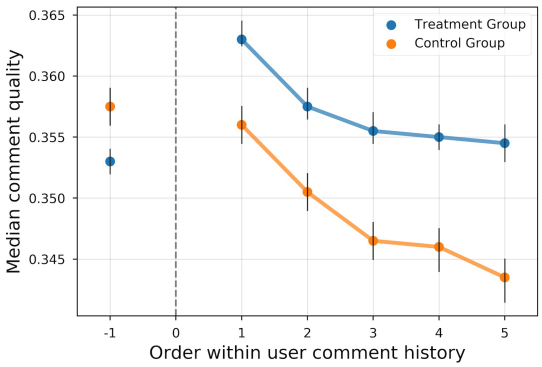

Figure 2. Median quality scores for each of one comment before and five comments in user history after the first Pick with 95% CIs highlighted in dark gray lines. Order zero corresponds to the (matched) Picked comment. For every order, the data contains all commenters who commented at that order.

could affect both the treatment (i.e., selected as Picks) and the outcome (i.e., comment quality), we match Pick comments with other comments from commenters who never receive any Picks. Since we define a comment's quality score as the probability of being selected as a Pick, we are able to use the comment quality score to do propensity score matching between Picks (i.e., treatment group) with comments of the same comment quality score that were not picked (i.e., control group). The control group consists of users who might have received a Pick based on the quality of a comment but didn't, drawn from the set of commenters who never received any Picks. We used an exact matching approach [38].

In order to observe the trend of how receiving Picks influences the between-subject comment quality distributions, we plot one approved comment before and five approved comments after the (matched) Picks from both the control and the treatment group in Figure 2. Order zero is the (matched) Pick comment, and we plot a dashed line to compare the comment quality before and after receiving Picks. We do not show the (matched) quality value in the figure because the value is too large for the scale of the chart (median = 0.4590), and if included would make it more difficult to compare the trends before and after receiving the Pick.

After commenters receive a first Pick, the quality distribution of the next comment after the Pick increases (before: 0.3530, after: 0.3630, $U = 3.6 \times 10^9, p < 0.0001$, common language effect size = 0.5209), but then steadily decreases until the sixth comment after the Pick when it reaches a level not different from the quality before the Pick ($U = 2.5 \times 10^9, p = 0.3140$). Past this point, the quality does not appear to dip substantially below the quality of the comment just before the Pick.

When we compare the trend for Pick receivers to the control group in Figure 2, we can see that the control group exhibits a small decrease in quality after receiving matched Picks, and a larger dropoff in quality in comparison to the treatment group. This finding helps demonstrate the relationship between receiving a Pick and the elevation of future comment quality.

We repeat this analysis for approved comments before and after the second and third Picks in order to examine whether there is an additional increase in the quality distribution after receiving multiple instances of positive feedback in the form of Picks. We find that the boosts from subsequent Picks are lower than from the first Pick. We observe lower boosts using Mann-Whitney's U test after the 2nd (before: 0.3680, after: 0.3703, $U = 7.1 \times 10^8, p = 0.11$, common language effect size = 0.5021) and 3rd Pick (before: 0.3655, after: 0.3695, $U = 2.5 \times 10^8, p = 0.0002$, common language effect size = 0.5092). The comment quality also drops more quickly compared to after the first Pick, which suggests repetitive positive reinforcement loses influence over time, an effect in line with conditioning theory [39]. To further examine the potential relationship between receiving Picks and article topic, we compare comment quality differences before and after receiving Picks across the top five most frequently commented-on news sections on NYT (World, U.S., Opinion, Sports, and Arts [40]). Receivers from all topics exhibit trends similar to Figure 2 after receiving Picks.

In summary, our results suggest that there is a quality boost to approved comments after commenters receive the first Pick, if they choose to return to the comment section, but that the quality boost attenuates

after subsequent picks and diminishes over time as the user continues commenting.

4.1.2. Do commenters receiving NYT Picks increase their future commenting frequency? In our analysis, we define a comment interval as the time difference between when the current comment was created and when the next comment was created. The median comment interval across all comment intervals from users who never receive any Picks is 13.88 days, higher than the 0.97 days for users who receive at least one Pick ($U = 7.7 \times 10^{12}, p < 0.0001$), which implies that users who receive Picks are more frequent users. We also observe that users decrease their comment intervals as they interact more with the comment sections, which suggests that if users choose to return to the comment section, they return to the comment section more and more frequently over time (see Figure 3). We observe a similar decreasing trend in comment intervals regardless of the number of comments a user has made (i.e., frequent commenters exhibit the same trend as less frequent commenters).

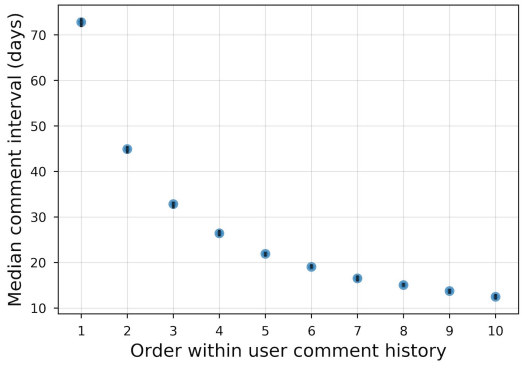

Figure 3. User comment interval for the first 10 comments in users' history with 95% CIs shown as dark gray lines. For every order, the data contains all commenters who commented at the time.

We find a general trend of comment intervals that decrease after the first comment after the Pick; therefore, we need to isolate whether this decrease is from the natural decrease as users engage more with the comment section (see Figure 3), or is a result of the reception of Picks. To assess this, we calculate the gradients of comment intervals for every user using the two-sided difference, which provides a more accurate approximation to the gradient than the one-sided difference approach. We compute the median of these gradients for users in the matched control and treatment group with at least three comments (i.e., two comment intervals) to calculate gradients (see Figure 4). Receiving Picks makes the gradients more negative than not receiving Picks (Mann-Whitney's U test, $p < 0.001$) for the first 2 comments, suggesting that receiving Picks may motivate commenters during their early tenure on the site (i.e., for their first 2 comments) to return to the section more quickly to make their next comment.

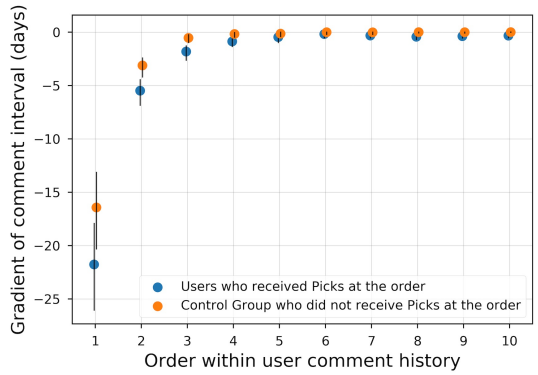

Figure 4. Median gradient on comment interval for the first 10 comments in users' history, matched by quality scores with users who didn't receive Picks at the order, with 95% CIs shown as dark gray lines.

4.2. Observing NYT Picks

In this section we focus on RQ3, which relates to the impact of users observing NYT Picks. In this section, we focus on articles that have at least one Pick (N = 42,724). 179,016 out of 221,740 articles in our dataset (80.7%) don't have any Picks. Among the 42,724 articles with Picks, the average number of Picks for each article is 7.65 (median = 5) with a maximum number of 218. We focus on comment replies as a way to understand the potential impact of Picks on observers' comment quality. We assume that commenters will read and be exposed to a comment before replying directly to it. To examine whether repliers may learn from the quality of the parent comments, we calculate the Spearman correlation between reply quality and parent comment quality and find it is positive (Spearman rho = 0.1220, $p < 0.0001, N = 1,769,840$). We compare the quality distribution for replies in response to Picks and in response to non-Picks to see whether reading a Pick is correlated with an increase in reply quality. The median quality score of replies in response to Picks is 0.2975 (N = 231,419), which is higher than replies in response to non-Picks (N = 1,538,421, median = 0.2885) ($U = 2.1 \times 10^{11}, p < 0.0001$, common language effect

size = 0.5207), suggesting that the Picks may have a positive association with increased comment quality in responses.

To further examine whether reply quality is simply associated with the generally higher quality of parent comments that are Picks or if the visibility of Pick badges may also play a role as suggested by previous studies [15], we build a generalized regression model (GLM) to predict reply quality using parent comment quality and the visibility of Pick badges. We also include the order of reply creation on the article and in user history (log-transformed since the distributions are skewed), and user first comment quality (which acts as a baseline before drop-off seen in Figure 1). We build a GLM using the gamma distribution and identity link function, since the Gamma distribution is for right-skewed continuous positive probability distributions (i.e., comment quality scores), and identity link results in the lowest AIC score. We observe that having a Pick parent badge has a small positive impact on the predicted reply quality; however, the parent comment quality has a much larger positive impact, indicating a greater influence from high-quality parent comments in comparison to Pick badges. The order of the comment in the article has a negative impact on the predicted comment quality (see Table 1).

Table 1. GLM summary predicting the current reply quality.

	Coef.	Std. Err.	z	$P >\| z \|$
Constant	0.2694	0.0004	633.5510	$< .0001$
Parent Pick	0.0016	0.0002	7.4048	$< .0001$
Parent Quality	0.0507	0.0004	118.6333	$< .0001$
Log order in article	-0.0057	0.0001	-97.6137	$< .0001$
Start Quality	0.1411	0.0006	250.5106	$< .0001$
Log order in user history	-0.0013	0.0002	7.4048	$< .0001$

Pseudo R^2 = 0.052; Num. obs. 1,769,840

5. Conclusions

In this paper, we develop and validate a machine learning model to quantify comment quality according to NYT editorial standards. Our analysis based on the model shows the positive relationship between NYT Picks and users' commenting behaviors. Our findings indicate that: (1) Picks are correlated with an improvement in first-time receivers' next approved comment quality, with the quality boost associated with receiving a Pick attenuating after subsequent picks; (2) receiving a Pick is associated with commenters early in their tenure on the site (i.e., within their first 2 approved comments) returning to the comment section more quickly to make their next comment; and (3) Exposure to Pick badges is associated with subsequently writing higher quality approved comments, though to a somewhat lesser degree compared to the quality of parent comments. Beyond these specific findings, the comment quality model we built could help inform future studies of online commenting communities.

Our findings support that there may exist social learning for both receivers and observers of NYT Picks, especially for newcomers and first-time receivers, and emphasize the importance of social feedback and learning for positive users' behaviors and engagement in online comment communities. Based on our findings, we think highlighting high-quality comments (i.e., Picks) is a sound moderation strategy for news organizations to signal quality debates with potential for positive outcomes. Picks are associated with receivers coming back to the comment section more frequently with higher comment quality, and are also associated with a small increase in observers' comment quality.

5.1. Design Opportunities

Informed by our findings, we consider design opportunities for online comment sections to experiment with the use of professional highlights as an approach to maintaining quality.

Design Opportunity 1: *Highlight high-quality comments from users who have not yet received Picks and who are in their early stages of commenting.*

Since our findings show that Picks are positively correlated with an improvement in the quality of receivers' next approved comment, with the correlation most prominent after the first Pick, we suggest that moderators prioritize highlighting comments from users who have never had a comment selected before. This may help encourage a wider range of users to write high-quality comments. Since receiving a Pick within the first two approved comments decreases receivers' comment intervals, we further suggest highlighting high-quality comments from newcomers to encourage them to return more quickly to the commenting section in the future. Providing an indication such as a badge for selected comments in the user interface may additionally spur reply comments to be slightly higher quality.

Design Opportunity 2: *Send notifications to users when they have a comment selected as a Pick.*

Email notifications are currently not sent when users receive Picks in the NYT comment sections. Considering our findings suggest a positive correlation between receiving Picks and future comment quality and frequency, we think it would be beneficial for comment sections to have a system to notify users when they are Picked. We expect that the correlations found in our study may be strengthened by increasing the salience to the user of having a comment selected as a Pick.

Design Opportunity 3: *Order the comments from high-quality to low-quality for users*.

Our findings indicate that repliers' comment quality is positively correlated with the parent comment quality. We suggest that the default order of comments in the interface could reflect higher quality at the top of the ranking since it may positively impact the quality of subsequent commenting and replies. Such an approach to ranking could be based on an automated quality evaluation model similar to the one we developed.

5.2. Limitations and Future Work

This study suggests several interesting directions for future work. Since our study only shows correlation rather than causality between NYT Picks and user comment quality and engagement, the direct impact of NYT Picks on user comment quality and engagement should be tested in future controlled experiments that might use click and trace data to more closely track commenter behaviors.

There are some limitations of our dataset that warrant elaboration. The NYT API does not return any comments that were flagged by moderators and were thus unpublished. Therefore, our analysis only considers *approved* comments (i.e., those visible on the site), which is a potential bias since there are some comments in a user's history that are missing from our analysis. According to the only study we could find using a complete dataset, 14.5% of NYT comments were unpublished (i.e., rejected) from Oct 2007 to Aug 2013 [41]. Given that our data is probably missing about 1 in 7 comments, we solely focus our findings on the comment quality and user behavior around *approved* comments. In an effort to characterize the threat of missing data to the validity of our findings, we consider several hypotheticals about the distribution of missing data. If the majority of the missing data occurs before the Picks for the receivers, receiving Picks could have a stronger positive correlation with subsequent quality boost than our findings indicate and a weaker association with future commenting frequency than our findings indicate. Likewise, if more missing data occurs after the Picks for the receivers, it would mitigate the quality correlation in the findings and strengthen the frequency association in the findings. If missing data is evenly distributed across commenters' commenting history (which we believe is a reasonable assumption), the magnitude of the correlations in findings should remain stable. Furthermore, since we don't have access to the rejected comments, we cannot be entirely certain of how observing high-quality comments and Picks may be associated with rejected comment quality. This suggests the need for future work examining how observing high-quality comments may influence inappropriate commenting behaviors (i.e., those comments that are ultimately rejected). While we have taken care to scope all of our claims in this paper to *approved* comments, future research should strive to compare findings that also take rejected comments into account.

While the most recent comments (i.e., from the last 4 years) from the NYT are missing from this study, we do not believe this presents an issue for the validity of our results. We are concerned with durable social behavior, and the extended time period of data that we do analyze is a strength that should help to smooth over any variations in how the media environment has changed over the 8 years that our data spans.

Another limitation of the current study is that we only consider the effect of NYT Picks on commenting behaviors and ignore the complex social feedback loops from the community (e.g., feedback from recommendations or down-votes) combined with NYT Picks. Future work should address these questions, including whether community feedback reinforces the impact from NYT Picks.

Lastly, we are aware that our machine learning model might not be the best estimation of comment quality for other online communities since it reflects NYT editorial standards and may exhibit certain biases peculiar to how NYT Picks are defined and chosen. However, our approach could still be helpful for researchers to better design their machine learning models to estimate community-specific comment quality using human validation.

References

[1] N. J. Stroud, E. Van Duyn, and C. Peacock, "Survey of commenters and comment readers," 2016.

[2] N. Diakopoulos and M. Naaman, "Towards quality discourse in online news comments," in *Proc. CSCW*, pp. 133–142, acm, 2011.

[3] N. Springer, I. Engelmann, and C. Pfaffinger, "User comments: Motives and inhibitors to write and read," *Information, Communication & Society*, vol. 18, no. 7, pp. 798–815, 2015.

[4] T. Gillespie, *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press, 2018.

[5] D. Park, S. Sachar, N. Diakopoulos, and N. Elmqvist, "Supporting comment moderators in identifying high quality online news comments," in *Proc. CHI*, pp. 1114–1125, ACM, 2016.

[6] N. J. Stroud, E. Van Duyn, A. Alizor, A. Alibhai, and C. Lang, "Comment section survey across 20 news sites," 2017.

[7] A. Bandura and R. H. Walters, *Social learning theory*, vol. 1. Prentice-hall Englewood Cliffs, NJ, 1977.

[8] S. Kiesler, R. Kraut, P. Resnick, and A. Kittur, "Regulating behavior in online communities," *Building successful online communities: Evidence-based social design*, pp. 125–178, 2012.

[9] J. Hattie and H. Timperley, "The power of feedback," *Review of educational research*, vol. 77, no. 1, pp. 81–112, 2007.

[10] R. Waldersee and F. Luthans, "The impact of positive and corrective feedback on customer service performance," *Journal of Organizational Behavior*, vol. 15, no. 1, pp. 83–95, 1994.

[11] J. Fanfarelli, S. Vie, and R. McDaniel, "Understanding digital badges through feedback, reward, and narrative: a multidisciplinary approach to building better badges in social environments," *Communication Design Quarterly Review*, vol. 3, no. 3, pp. 56–60, 2015.

[12] B. R. Watson, Z. Peng, and S. C. Lewis, "Who will intervene to save news comments? deviance and social control in communities of news commenters," *New Media & Society*, vol. 21, no. 8, pp. 1840–1858, 2019.

[13] E. Fehr, U. Fischbacher, and S. Gächter, "Strong reciprocity, human cooperation, and the enforcement of social norms," *Human nature*, vol. 13, no. 1, pp. 1–25, 2002.

[14] J. N. Matias, "Preventing harassment and increasing group participation through social norms in 2,190 online science discussions," *Proceedings of the National Academy of Sciences*, vol. 116, no. 20, pp. 9785–9789, 2019.

[15] A. Anderson, D. Huttenlocher, J. Kleinberg, and J. Leskovec, "Steering user behavior with badges," in *Proceedings of the 22nd international conference on World Wide Web*, pp. 95–106, ACM, 2013.

[16] J. Cheng, C. Danescu-Niculescu-Mizil, and J. Leskovec, "How community feedback shapes user behavior," in *Proc. ICWSM*, 2014.

[17] J. Seering, R. Kraut, and L. Dabbish, "Shaping pro and anti-social behavior on twitch through moderation and example-setting," in *Proc. CSCW*, pp. 111–125, ACM, 2017.

[18] N. A. Diakopoulos, "The editor's eye: Curation and comment relevance on the new york times," in *Proc. CSCW*, pp. 1153–1157, ACM, 2015.

[19] H. Zhu, A. Zhang, J. He, R. E. Kraut, and A. Kittur, "Effects of peer feedback on contribution: a field experiment in wikipedia," in *Proc. CHI*, pp. 2253–2262, ACM, 2013.

[20] J. Schneider, A. Passant, and J. G. Breslin, "Understanding and improving wikipedia article discussion spaces," in *Proc. SAC*, pp. 808–813, ACM, 2011.

[21] C. Lampe and E. Johnston, "Follow the (slash) dot: effects of feedback on new members in an online community," in *Proc. SIGGROUP*, pp. 11–20, ACM, 2005.

[22] E. Joyce and R. E. Kraut, "Predicting continued participation in newsgroups," *Journal of Computer-Mediated Communication*, vol. 11, no. 3, pp. 723–747, 2006.

[23] C. Cheshire, "Selective incentives and generalized information exchange," *Social Psychology Quarterly*, vol. 70, no. 1, pp. 82–100, 2007.

[24] M. Burke, C. Marlow, and T. Lento, "Feed me: motivating newcomer contribution in social network sites," in *Proc. CHI*, pp. 945–954, ACM, 2009.

[25] A. Sukumaran, S. Vezich, M. McHugh, and C. Nass, "Normative influences on thoughtful online participation," in *Proc. CHI*, pp. 3401–3410, ACM, 2011.

[26] S.-H. Han and L. M. Brazeal, "Playing nice: Modeling civility in online political discussions," *Communication Research Reports*, vol. 32, no. 1, pp. 20–28, 2015.

[27] J. Cheng, M. Bernstein, C. Danescu-Niculescu-Mizil, and J. Leskovec, "Anyone can become a troll: Causes of trolling behavior in online discussions," in *Proc. CSCW*, pp. 1217–1230, ACM, 2017.

[28] R. Fredheim, A. Moore, and J. Naughton, "Anonymity and online commenting: The broken windows effect and the end of drive-by commenting," in *Proc. WebSci*, p. 11, ACM, 2015.

[29] V. Johansson, "Lexical diversity and lexical density in speech and writing: A developmental perspective," *Lund Working Papers in Linguistics*, vol. 53, pp. 61–79, 2009.

[30] G. H. Mc Laughlin, "Smog grading-a new readability formula," *Journal of reading*, vol. 12, no. 8, pp. 639–646, 1969.

[31] C. J. Hutto and E. Gilbert, "Vader: A parsimonious rule-based model for sentiment analysis of social media text," in *Eighth international AAAI conference on weblogs and social media*, 2014.

[32] E. Wulczyn, N. Thain, and L. Dixon, "Ex machina: Personal attacks seen at scale," in *Proc.WWW*, pp. 1391–1399, International World Wide Web Conferences Steering Committee, 2017.

[33] Z. Papacharissi, "Democracy online: Civility, politeness, and the democratic potential of online political discussion groups," *New media & society*, vol. 6, no. 2, pp. 259–283, 2004.

[34] L. Breiman, "Bagging predictors," *Machine learning*, vol. 24, no. 2, pp. 123–140, 1996.

[35] K. Benoit, D. Conway, B. E. Lauderdale, M. Laver, and S. Mikhaylov, "Crowd-sourced text analysis: Reproducible and agile production of political data," *American Political Science Review*, vol. 110, no. 2, pp. 278–295, 2016.

[36] A. Kittur, E. H. Chi, and B. Suh, "Crowdsourcing user studies with mechanical turk," in *Proc.CHI*, pp. 453–456, ACM, 2008.

[37] R. Artstein and M. Poesio, "Inter-coder agreement for computational linguistics," *Computational Linguistics*, vol. 34, no. 4, pp. 555–596, 2008.

[38] D. Ho, K. Imai, G. King, and E. Stuart, "Matchit: Nonparametric preprocessing for parametric casual inference. r package version 2-2," 2006.

[39] D. Y. Wohn, "The effects of feedback and habit on content posting in an online community," *iConference 2015 Proceedings*, 2015.

[40] S. S. Sachar and N. Diakopoulos, "Changing names in online news comments at the new york times," in *Proc. ICWSM*, 2016.

[41] A. Muddiman and N. J. Stroud, "10 things we learned by analyzing 9 million comments from the new york times," 2016.