Making the Pick: Understanding Professional Editor Comment Curation in Online News

Yupeng He,[1] Yimeng Gu,[2] Ravi Shekhar,[3] Ignacio Castro,[2] Gareth Tyson,[1]

[1] Hong Kong University of Science and Technology, GZ
[2] Queen Mary University of London, London, UK
[3] University of Essex, UK

yhe382@connect.hkust-gz.edu.cn, yimeng.gu@qmul.ac.uk, r.shekhar@essex.ac.uk, i.castro@qmul.ac.uk, gtyson@ust.hk

Abstract

Online comments within news articles are a key way people share opinions. Discovering insightful comments can, however, be challenging for readers. A solution to this problem is using *comment curation*, whereby professional editors select the highest quality comments manually — referred to as "editor-picks". This paper studies the growing use of professional editor-curation for user-generated comments. We focus on the New York Times as a case study, using a dataset covering 80k articles. We study the characteristics of editor-pick comments, highlighting how editor criteria vary across news sections (*e.g.* sports, entertainment). We find that editor-pick comments tend to be longer, more relevant to the article, positive in sentiment, and contain low toxicity. Our analysis further reveals that editors within different news sections exhibit differing criteria when they perform comment selection. Thus, we finally propose a set of models that can automatically identify good candidate editor-picks. Our ultimate goal is to reduce editor and journalistic workload, increasing productivity and the quality of curated comments.

1 Introduction

The way we consume and disseminate news has undergone a dramatic shift. While traditional media outlets like newspapers still play a significant role in shaping public opinion, they are increasingly complemented by online sources of information. One of the most notable features of online news dissemination is the prevalence of user-generated content, particularly in the form of online comments. Online comments are a key way that people share news, opinions, and information with one another (Stroud, Van Duyn, and Peacock 2016) — whether it is sharing personal experiences, offering additional insights, or simply expressing support or criticism of the article.

Discovering insightful comments can, however, be challenging for readers. Popular articles can gain thousands of comments, making it hard to select the best to read. A growing solution to this problem is to use of *comment curation* (McInnis et al. 2021), whereby paid editors select the highest quality comments manually. Such comments can then be highlighted within news articles, improving the readers' experience.

This paper studies the growing use of professional editor-curation for news comments. As an exemplar, we focus on one of the world's most well-known news outlets: the New York Times (NYT). The comment section on the NYT is a carefully moderated and curated space that strives to uphold the publication's standards of quality and integrity (Diakopoulos 2015). Professional editors select comments based on their own criteria, which are then featured prominently within the comment section. Here, the editor's purpose is to enable readers to engage with diverse perspectives and deepen their understanding of the issues at hand. These selected comments are referred to as "editor-picks".

Unfortunately, running such a system is not without costs, constituting a significant manual workload. This makes such activities prohibitive for smaller news outlets. It further means that important comments can easily be missed, particularly for fast-moving news stories. Thus, we argue that developing tools supporting editors in their activities could create significant benefits (Park et al. 2016). With this in-mind, and exploiting a large-scale comment dataset from the NYT (Section 3), we explore three research questions:

RQ1 *What factors impact the likelihood of a comment being selected by an editor (aka. an "editor-pick")?* Through LIWC analysis (Tausczik and Pennebaker 2010), we identify psycho-linguistic factors in comments and then identify four other factors (comment length, relevance, sentiment, and toxicity) that seem to influence comment selection. We argue that the answer could provide key insights into the comment curation process and be further used to develop tools for supporting editors (Section 4).

RQ2 *Do the user engagement and editor selection preferences differ across different news sections (*e.g. *entertainment, sports)?* We subdivide articles into their news sections (*e.g.* entertainment, sports) and explore how their comment patterns differ. This sheds unique insight into how editors in different sections behave. We find that editors within different sections exhibit differing criteria, which must be taken into account when attempting to automate comment selection (Section 5).

RQ3 *Can the factors discovered in RQ1 and RQ2 be exploited to build tooling to automatically identify potential editor-picks, and could this then streamline comment*

selection? We aim to assist editors by designing tools to automatically shortlist suitable candidate comments for them to select from. We experiment with several classifiers and attain a 0.75 macro-F1 score on the global dataset but find that per-section models gain substantial improvements. Our top-performing news section is UPSHOT, which attains a 0.86 macro-F1 score by Gradient Boosting and Random Forest classifiers (Section 6).

2 Related Work

NYT Comment Analysis. Previous efforts have been undertaken in analyzing NYT comments from different angles. Diakopoulos (2015) found an association between editor-pick comments and article relevance or conversational relevance on NYT; Karpova, Best, and Bayat (2020) analyzed voluntary online comments posted on NYT to study consumer views towards sustainability in the fashion industry; Kolhatkar and Taboada (2017a,b) defined and identified constructive comments from news articles (NYT and Yahoo News), and further examined the relationship between constructiveness and toxicity. Juarez Miro (2022) examined the characteristics of comments both in the NYT Picks section and Reader Picks section using qualitative content analysis. Additionally, Muddiman and Stroud (2016) investigated how the technical redesign of the comment section affects user behaviors in posting comments. Moreover, Park et al. (2016) designed and evaluated a UI system to assist editors in interactively identifying high-quality comments. Wang and Diakopoulos (2022) examined the quality and frequency of commenting on the site in response to NYT Picks. Although there has been diverse analysis carried out on NYT comments, many of them are qualitative and little work has shed light on understanding the editor's selection preferences.

Comment Classification. Other works attempt to develop automatic methods for classification of online comments. Kolhatkar and Taboada (2017b) leveraged classifiers to identify constructive comments based on the positive samples from NYT and negative samples from Yahoo. Shekhar et al. (2020) presented a benchmark study for automated news comment moderation using limited resources in Croatian and Estonian. Based on feature extraction of BERT, Wei et al. (2021) proposed a news comment relevance classification algorithm. Jang, Kim, and Kim (2019) proposed a classification model based on convolutional neural networks and Word2vec embeddings for the classification of NYT articles and tweets. Naeem et al. (2022) presented a classification approach using term frequency-inverse document frequency and optimized machine learning algorithms, and achieved high accuracy rates on a movie comment dataset. Although these prior works explore the classification of comments, most of them focus on the classifier design itself. There has been little work exploring the preferences of editors. Further, little prior work focuses on the goal of helping editors with comment selection.

There are also journalism studies looking into comment curation. Wolfgang (2018) used gatekeeping theory to study the content management through content moderation. Ferrucci and David Wolfgang (2021) also studied how the moderation of reader comments may decrease journalistic autonomy. Our work differs in that we devise tooling through an empirical exploration of editor picks. The goal is to help with news comment moderation in the future.

Table 1: The attributes we used for subsequent analysis in the New York Times dataset.

Comment-related	Article-related
articleID	abstract
articleURL	web_url
commentBody	
editorsSelection	
replyCount	
recommendations	

3 Dataset

To underpin our work, we gather a large-scale dataset from one exemplar news site: the New York Times (NYT).

3.1 Dataset Collection

We gather 80,524 articles from the NYT, covering 1,200,977 comments using the official NYT API.[1] We collected all article URLs in 2019, including all their comments. We employ a two-step process to download all comments from an article. The first step uses the NYT *url.json* endpoint to collect all comments on the article. This returns top-level comments and the first three replies. To download all reply comments, we then query the NYT *replies.json* endpoint to collect all remaining replies to comments that have more than three replies. Note, this step is missed by Kesarwani (2018), and in turn, they miss the complete discussion.

Table 1 summarizes the attributes of the comments and their corresponding articles. Not all articles have a comment, and we report only those articles which have at least one comment. Each article has a unique identifier ("articleID") and a URL ("articleURL"). In addition, each article belongs to a section of the NYT (*e.g.* BUSINESS, OPINION, MOVIES). Overall, there are 40 different sections in the dataset. As our focus is on understanding editor-picks, we discard all articles with no comments selected by the editors (since the lack of editor-pick comments prevents us from obtaining the editor's preferences for these articles). As a result, we discard 75,036 articles containing 585,890 non-editor-picks. Note, out of the 40 sections, 5 (MULTIMEDIA, SMARTER-LIVING, ADMIN, VIDEO, AUTOMOBILES) have no article with a comment selected by the editor. Our final dataset contains 615,087 comments from 5,488 articles within 35 sections.

Table 2 summarizes the number of articles and comments for the top 10 popular sections[2] in the original dataset (dataset before filtering), and the final dataset (dataset after filtering), respectively. From Table 2, we see that news

[1] https://developer.nytimes.com/apis

[2] The popularity is based on the number of articles after filtering.

Table 2: Number of articles and comments of top 10 popular sections before and after filtering the New York Times dataset. O-ART (**O**riginal number of **ART**icles) and F-ART (**F**inal number of **ART**icles) refer to the number of articles in the original dataset and the final dataset, respectively. O-COM (**O**riginal number of **COM**ments) and F-COM (**F**inal number of **COM**ments) refer to the number of comments in the original dataset and the final dataset, respectively.

Section	O-ART	F-ART	O-COM	F-COM
OPINION	2,979	2,146	333,269	249,371
US	2,416	1,345	232,063	155,431
WORLD	1,304	584	93,641	55,869
NYREGION	1,036	264	66,907	29,496
BUSINESS	806	260	64,050	30,549
HEALTH	261	88	20,003	10,159
MAGAZINE	466	87	34,890	9,973
ARTS	987	80	31,910	8,502
SPORTS	544	78	21,266	7,874
UPSHOT	275	71	30,401	8,919
Top 10 Sections	11,074	5,003	928,400	566,143

sections receives the highest number of user comments are OPINION, US, and WORLD.

3.2 Grouping Comments

Each comment body ("commentBody") is provided with the comment's metadata information. For each comment, we use the following attributes: "editorsSelection", indicating whether a comment is picked by the editors; "recommendation", indicating how many users recommend this comment; and "replyCount", indicating how many replies this comment gets. Of all the 615,087 comments in the filtered dataset, 19,727 comments are picked by editors. In order to carry out a more insightful analysis of the non-editor-picks, and to compare them with the editor-picks, we divide comments into four different groups:

- **Editor-picks.** This set covers comments selected by editors. There are 19,727 comments within this group.
- **User-picks.** This set covers any comment that is not picked by an editor but receives $\geq$10 recommendations or $\geq$3 replies. This is intended to cover comments that gain moderate attention from users (hence the term user-picks). These specific thresholds were selected as, intuitively, we believe that the top quarter of comments could be considered "popular" among users. Thus, we compare several natural thresholds, and select the one closest ($\geq$10 recommendations or $\geq$3 replies covers 26.91% of comments). We define this group in order to explore further the difference between editors' criteria and users' preferences (as measured by the attention received). There are 160,227 user-pick comments.
- **Zero-picks.** This set covers any comments receiving zero recommendations and zero replies. We select these to represent low-quality comments that gain little traction. There are 62,271 zero-pick comments, 10.46% of all non-editor-picks.

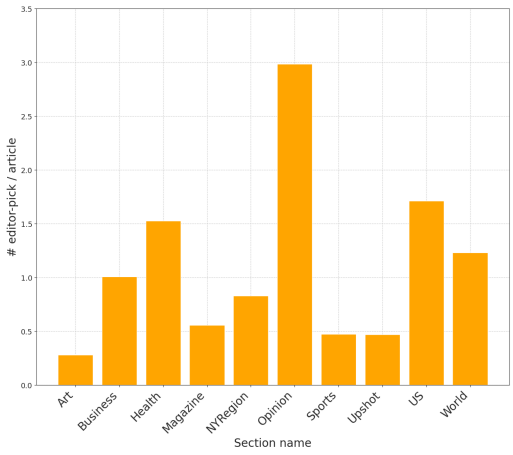


Figure 1: Number of editor-picks posted (normalized over the number of articles in the corresponding section) in 2019.

- **Other-picks.** Other-pick covers all the remaining comments. There are 372,862 comments within the other-picks set.

Our intention of introducing user-picks, zero-picks, and other-picks is to have a more fine-grained categorization of non-editor-picks. This also helps us to "filter out" low-quality comments, *i.e.* zero-picks and other-picks. Our subsequent analysis relies on this categorization to better understand the contrasts between comment types.

3.3 Dataset Overview

Before exploring our research questions, we provide a brief statistical description of our dataset to help contextualize our later results. Figure 1 presents the number of editor-picks (normalized over the number of articles within the corresponding section) posted in 2019. We plot the number of editor-picks per article of the top 10 popular sections. We see that among all the sections, articles in OPINION receive the highest number of editor-picks on average, reaching 2.985. Articles in US (covering US news) rank second, with the number of editor-picks being 1.712 on average. Articles in the ARTS section receive the least (0.279) editor-picks on average.

4 RQ1: Factors Affecting the Editor's Pick

In this section, we focus on **RQ1**, which aims to uncover the factors that affect editors' preferences for picking certain comments. Our analysis covers two perspectives: (*i*) *linguistic-based*: we inspect a set of linguistic features and identify the ones that are more significant in indicating whether a comment will become an editor-pick; and (*ii*) *attribute-based*: we then perform a look at four other attributes to better understand the editor's selection preferences. We aim to understand factors that can underpin our later classifier work in Section 6.

Table 3: 20 LIWC features assigned with top 10 coefficients (both positive and negative) by the LR classifier.

Feature	Coef.(+)	Feature	Coef.(-)
conversation	0.1402	*emoji*	-0.5223
culture	0.0768	*nonflu*	-0.2110
emo_neg	0.0721	*swear*	-0.1866
emo_pos	0.0565	*filler*	-0.1843
we	0.0472	*assent*	-0.1791
achieve	0.0414	*netspeak*	-0.1588
affect	0.0407	*ethnicity*	-0.0866
power	0.0363	*politic*	-0.0837
leisure	0.0356	*sexual*	-0.0801
affiliation	0.0334	*tone_pos*	-0.0531

4.1 Linguistic-based Analysis

To gain an understanding of factors affecting editor choices, we start by computing linguistic features' (118 in total) scores for the comments using LIWC-22 (Linguistic Inquiry and Word Count) (Boyd et al. 2022). LIWC is a dictionary that maps important psycho-social constructs to words, phrases, and other linguistic constructions. After acquiring the linguistic feature vectors, we train a group of logistic regression (LR) classifiers on them to try and classify editor-picks *vs.* user-picks. Recall, in the filtered dataset, there are 19,727 editor-picks and 160,227 user-picks. To balance the two labels, we randomly permute the user-picks and divide them by 19,727 to form 8 non-overlapping groups. Then for each group of 19,727 user-picks plus 19,727 editor-picks, we split them into training and testing sets with a ratio of 80:20. On the 8 test sets, the LR classifier achieves an average accuracy of 0.632, giving us confidence in our following analysis based on the LR classifier's coefficients.

We next inspect the coefficients (averaged over coefficients of 8 LR classifiers) assigned to each feature to understand the types of LIWC features that correlate with the editors' preference. Table 3 reports the features that receive the 20 largest coefficients (10 positive and 10 negative). We see that *conversation*, *culture (*e.g. *car, united states, govern, phone)*, and *emo_neg* are the three LIWC features that affect the editor's pick most positively. On the other hand, *emoji*, *nonflu (*e.g. *oh, um, uh)* and *swear* are the three LIWC features that affect editor's pick most negatively. This indicates, for example, that comments containing emojis are less likely to be selected by editors. Nevertheless, positive coefficients are less prominent than their negative counterparts, suggesting that features receiving high positive coefficients are less dominant in the class prediction. Furthermore, among the 10 features receiving negative coefficients, two of them are related to sentiments, *i.e. emo_neg and emo_pos*.

We further test the statistical significance of the coefficients. We use $p<0.05$ (z-test) as the significance criterion. For the positive coefficients, *we* (0.024) is statistically significant. For the negative coefficients, *swear* (0.042), *ethnicity* (0.028) *sexual* (0.001) are statistically significant.

Table 4: Group-wise comparison on the percentage of comments at different length levels. For convenience, we embolden any values that are mentioned as examples in the text.

Length	Editor -picks(%)	User -picks(%)	Other -picks(%)	Zero -picks(%)
≥ 20	95.110	86.292	81.796	71.934
≥ 30	89.947	75.372	70.040	58.349
≥ 40	83.825	65.664	59.914	48.126
≥ 50	77.199	56.964	51.219	40.003
≥ 60	69.887	49.138	43.799	33.369
≥ 70	61.972	42.403	37.445	28.017
≥ 80	54.491	36.596	32.044	23.449
≥ 90	**47.746**	31.420	27.384	**19.873**

4.2 Attribute-based Analysis

After analyzing the linguistic features of text, we now turn our attention to other attributes. To explore what kind of factors contained in the comment have greater influence on the likelihood of editor selection, we next select four factors: the (*i*) comment length, (*ii*) relevance between the comment and its corresponding article, (*iii*) sentiment, and the (*iv*) toxicity of comments. As defined in Section 3.2, there are four different groups of comments in the dataset, *i.e.* editor-picks, user-picks, zero-picks, and other-picks. In order to obtain the preferences of editors, we compare the differences between editor-picks and the other 3 groups.

Comment Length. Table 4 shows the percentage of comments that have each comment length across the four different comment groups. The editor-picks group contains the highest percentage (95.11%) of comments with over 20 words. For comments with a length longer or equal to 90 words, the percentage of the editor-picks group (47.74%) is 16% higher than the percentage of the user-picks group (19.87%). This suggests that editors prefer longer comments. To complement this, Figure 2 (a) plots the distribution of the comments lengths. We find that the distribution of lengths for different comment groups is all different. The distance between editor-picks and zero-picks curves is the largest. The Wasserstein distance between these 2 curves is 0.0032, compared with 0.0022 between editor-picks and user-picks; 0.0024 between editor-picks and other-picks. There is also a large distance between the editor-picks curve and the user-picks curve (with Wasserstein distance 0.0022). We further conduct Kruskal-Wallis test (K-W test) and conduct post-hoc tests to measure the significance of the difference among these groups. The results confirm that editor-pick comments are statistically different to all other groups regarding the length of comments (all post hoc p-values are <0.001). This shows that editors are more likely to select comments with longer lengths compared with comments from the other 3 groups.

Relevance. An intuitive assumption is that editors might select comments that are more relevant to the content of the article. This, for example, can help filter irrelevant or spam postings. To verify the validity of this assumption, we an-

Table 5: Group-wise comparison on the percentage of comments at different similarity score levels.

Similarity	Editor -picks(%)	User -picks(%)	Other -picks(%)	Zero -picks(%)
≥ 0.67	50.678	40.106	28.382	35.603
≥ 0.70	35.688	26.446	17.575	22.783
≥ 0.73	22.255	15.609	9.723	13.063
≥ 0.76	11.877	7.936	4.726	6.482
≥ 0.79	5.287	3.410	1.888	2.704
≥ 0.82	1.912	1.233	0.616	0.928
≥ 0.85	0.546	0.344	0.167	0.253
≥ 0.88	0.129	0.070	0.045	0.059

Table 6: Group-wise comparison on the percentage of comments at different sentiment score levels.

Sentiment	Editor -picks(%)	User -picks(%)	Other -picks(%)	Zero -picks(%)
$\leq$ -0.6	26.686	24.006	21.445	17.441
$\leq$ -0.7	22.997	19.923	17.453	13.925
$\leq$ -0.8	17.988	14.826	12.713	9.794
$\leq$ -0.9	10.960	8.393	7.017	5.379
≥ 0.9	15.474	10.175	8.373	6.262
≥ 0.8	24.991	18.230	15.775	12.480
≥ 0.7	31.432	24.446	21.902	18.283
≥ 0.6	35.966	29.502	27.126	23.705

alyze the relevant similarity between the comment text and the article text.

To compute the similarity, we first use SentenceTransformer (Reimers and Gurevych 2019) to obtain the vector representation of the comment text and abstract text of the article.[3] Note, SentenceTransformer is a pre-trained model in a popular open-source library Hugging Face (Wolf et al. 2019), which can convert sentences into dense vector representations that capture the meaning of the text. After being embedded by SentenceTransformer, we calculate the cosine similarity between the vector representations of the comment text and abstract text. In order to validate whether the similarity between vectors from SentenceTransformer can capture the relevance between comment text and the article abstract, we manually check 100 samples. Specifically, we first list the 200 comments with the lowest similarity in all samples (similarity ≤ 0.46), and randomly pick 50 examples. We then extract the 200 comments with the highest similarity in all samples (similarity ≥ 0.85), and randomly pick another 50 examples. We then manually estimate the relevance between the comment text and the abstract. We find that the comment text of examples with high similarity are indeed far more relevant to the abstract compared with examples with low similarity. Hence, we are confident that the vector similarities serve as an effective proxy metric for relevance.

Figure 2 (b) presents the similarity ECDF curves of comments in the dataset. We observe that there are different distributions across different comment groups. This indicates that editors have different selection preferences compared with the other three groups. We see that editors tend to prefer comments closely related to their corresponding articles, *i.e.* with high similarity. As shown in Table 5, for each similarity level, the percentage of the editor-picks group is the highest compared with the other three groups. For similarity ≥ 0.67, the percentage of editor-picks (50.678%) is 10% higher than user-picks (40.106%) and 15% higher than zero-picks (35.603%). For this feature, we also conduct a K-W test and post-hoc tests to measure the significance of the difference among these groups. We confirm that editor-pick comments are statistically different to all other groups regarding the similarity of comments (all post hoc p-values are <0.001). We thus conclude that editors prefer comments with high similarity to the article.

[3] An abstract is a summary of the article provided by the NYT.

Table 7: Group-wise comparison on the percentage of comments at different toxicity score levels. For convenience, we embolden any values that are mentioned as examples in the text.

Toxicity	Editor -picks(%)	User -picks(%)	Other -picks(%)	Zero -picks(%)
≥ 0.1	54.202	55.182	45.563	52.686
≥ 0.2	27.923	32.409	24.628	30.011
≥ 0.3	14.227	18.471	12.969	16.721
≥ 0.4	4.684	7.783	4.911	6.766
≥ 0.5	1.309	3.072	1.767	2.620
≥ 0.6	0.417	1.198	0.627	0.953
≥ 0.7	0.062	0.292	0.120	0.240
≥ 0.8	**0.005**	**0.038**	0.015	0.031

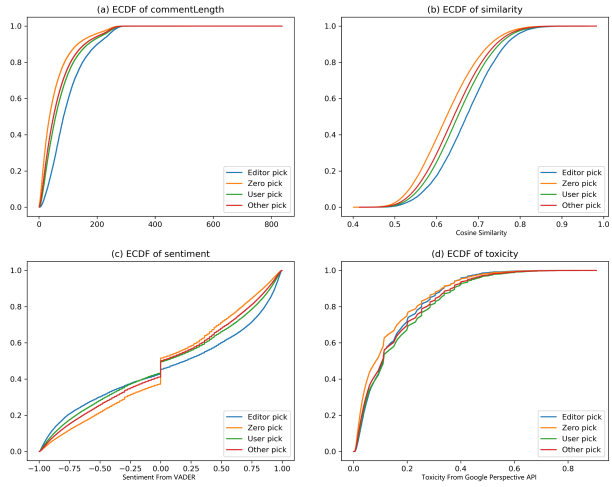


Figure 2: ECDF of comment length, similarity score, sentiment score and toxicity score in the article-filtered dataset.

Sentiment. We next explore the influence that comment sentiment has on the decision of editors. We use VADER (Valence Aware Dictionary and sEntiment Reasoner) (Hutto and Gilbert 2014), a pre-trained tool for sentiment analysis.

The sentiment value of a comment varies from -1 to 1. If the sentiment of a comment is smaller than 0, the emotional tendency of this comment is negative, and vice versa.

Figure 2 (c) presents the sentiment per comment as an ECDF. We again see that the sentiment correlates with the preferences of editors and users since the curves of 4 comment groups are distinct. The curves suggest that editors are more likely to select comments with high sentiment (highly positive or negative) compared with other categories. To confirm this difference, we perform a Kruskal-Wallis test (K-W test) and post-hoc tests to measure the significance of the difference among these groups. The results confirm that editor-pick comments are statistically different to all other groups regarding the sentiment of comments (all post hoc p-values are <0.001). Note, only few editor-picks are neutral, suggesting that editors do not like comments that are too emotionally neutral.

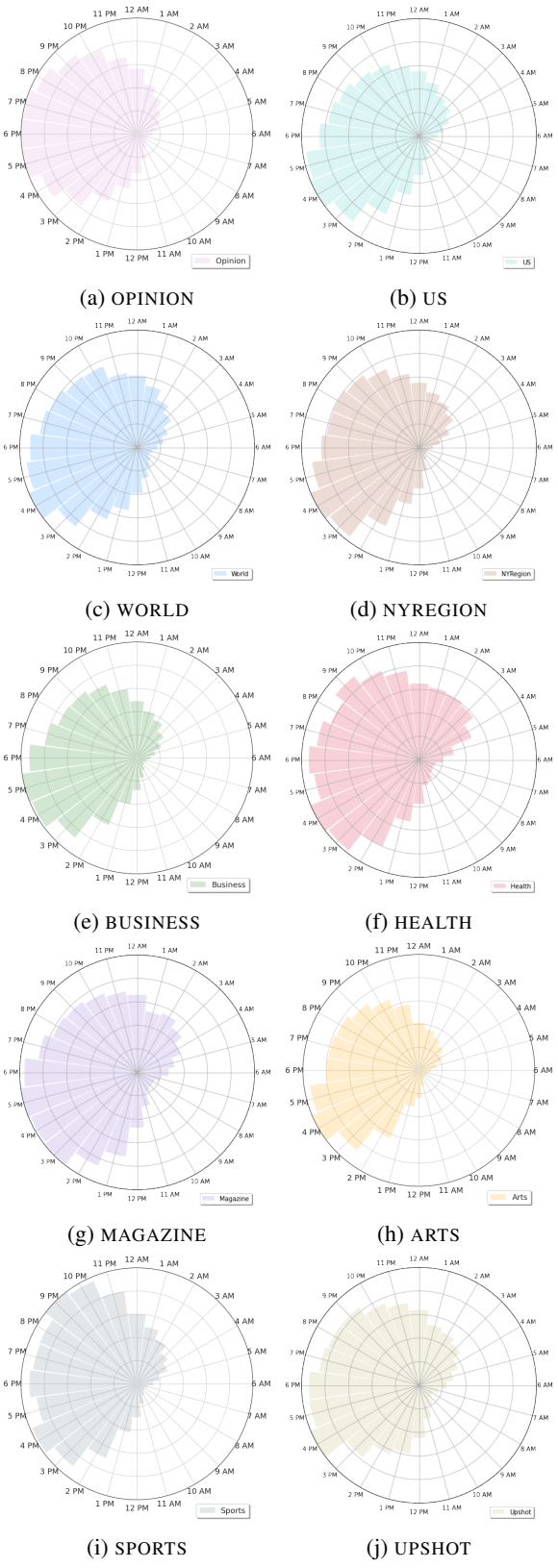
Toxicity. We finally use the Google Perspective API (Jigsaw 2022) to label the toxicity of comments. We are curious to understand how toxic comments are responded to by editors. The toxicity score provided by the API ranges from 0 to 1, with 0 indicating that the text is not toxic and 1 indicating that the text is highly toxic. Figure 2 (d) presents the distribution of toxicity across the four category groups. The difference in distributions across all comment groups is less than compared to the prior three metrics. This is likely because NYT comments are moderated, and highly toxic comments are unlikely to appear. Despite this, on average, the editor-picks comments are less toxic than user-picks. We further conduct a K-W test and post-hoc tests to measure the significance of the difference for toxicity among these groups. This confirms that editor-picks are statistically different to all other groups regarding the toxicity of comments (all post hoc p-values are <0.001). To expand, Table 7 presents the percentage of comments classified as toxic based on several thresholds. Again, this confirms the percentage of editor-picks is the lowest compared with the other 3 groups. As the highlighted values show in the table, the percentage of user-picks with toxicity ≥ 0.8 (0.038%) is 8x the editor-picks (0.005%). This confirms that editors prefer to select comments with less toxic material.

5 RQ2: Differences Across News Sections

Recall, comments come from different sections (*e.g.* SPORTS, ENTERTAINMENT). Thus, we next answer **RQ2** and explore if user engagement and editor selection preferences differ across different news sections. We focus on the top 10 sections, as these cover 92% of all comments.

5.1 User Engagement Per-Section

To explore the user engagement per section highlighted in RQ2, we first inspect the timestamp of when comments are posted. This is important because we conjecture that editors may work particular hours, and the comment time could bias their selection on a per-section basis. Note, the comment timestamp is in EST (New York time) and the majority of users are from North America. After manual check on 100



(a) OPINION (b) US (c) WORLD (d) NYREGION (e) BUSINESS (f) HEALTH (g) MAGAZINE (h) ARTS (i) SPORTS (j) UPSHOT

Figure 3: Statistics of comment posting time (in hours) in 10 sections

random comments, we find that 95 are from North America (mainly US) according to the 'userLocation' attribute. Note, the 'userLocation' has non-regular forms, *e.g.* 'miami beach', 'NYC', 'Okalahoma City', 'some city', hence resorting to manual annotation. Figure 3 plots the number of comments posted in each hour range across the top 10 sections. Each bar is normalized over the maximum value in that subplot.

We find that the different news sections tend to have different user engagement levels over time, *i.e.* comment posting peaks. For instance, active users in HEALTH, NYREGION and MAGAZINE tend to comment between 3PM and 6PM, while active users in ARTS, BUSINESS, US, WORLD and UPSHOT are most likely to post comments between 4PM and 5PM. However, in OPINION, users tend to comment slightly later — between 5PM and 7PM. Furthermore, SPORTS sees more "night owls" posting comments between 9PM and 10PM. Moreover, the duration of comment posting peaks varies across different sections. This discrepancy in user engagement might be caused by the different tastes and behaviors of user groups, and also by when the news events take place in the corresponding sections. Our preliminary analysis suggests that there is a notable difference in comment timestamps across different sections. This motivates us to build section-specific classifiers in the subsequent analysis (Section 6) to better accommodate the classification in the targeted section.

5.2 Editor Preferences Per-Section

Next, to explore the difference in editor preferences across sections in RQ2, we revisit the four features discussed in Section 4 and check how they differ across the NYT sections.

Comment Length. To measure the statistical significance of differences in editor preferences across the top 10 sections, we first conduct K-W test for comment length among these sections. The result confirms that these sections have a significant difference regarding comment length (p-value is $<$0.001). Table 8 presents the percentage of editor-pick comments broken down into different comment lengths. We find that editors in different sections tend to have varied preferences for the length of comments. The UPSHOT section has the longest comments picked by editors (a section dedicated to using data to understand politics). To be specific, UPSHOT has the highest percentage (43.284%) of comments with length $\geq$ 100 words, and 26.119% (also the highest) for comments with length $\geq$ 160 words. In contrast, editors in the NYREGION and WORLD sections tend to select shorter comments when compared to other sections. NYREGION has the lowest percentage (25.112%) of comments with length $\geq$ 100, and WORLD has the lowest percentage (9.648%) of comments with length $\geq$ 160.

Relevance. Next, Table 9 shows the similarity (between comment text and the abstract of the corresponding article, as mentioned in Section 4) for editor-picks across the top 10 sections. We also conduct a K-W test for similarity among these sections. The result confirms that the preferences of editor in these sections are significantly different regarding with similarity (p-value is $<$0.001). The percentage of editor-picks is lowest in section OPINION (43.992%), where the value of similarity is greater or equal to 0.67. This indicates that editors in OPINION do not consider similarity as important as in other sections. This seems largely because of the more subjective and diverse nature of the OPINION section. In contrast, editors in HEALTH and SPORTS tend to select comments that connect closely with corresponding articles, since these two sections have the highest similarity value in the shown levels. For example, HEALTH has the highest (1.478%) for comments with similarity $\geq$ 0.85 , SPORTS has the highest (1.128%) for comments with similarity $\geq$ 0.88. This is likely because of the nature of such sections, where facts (rather than opinions) are focused on.

Sentiment. To measure the statistical difference in sentiment across the top 10 sections, we again conduct K-W tests among these sections. The result confirms that these sections has significant difference regarding with editor preferences in terms of sentiment (p-value is $<$0.001). Table 10 presents the sentiment scores for editor-picks in each section. We see that section WORLD has the highest percentage (12.864%) of comments with sentiment scores ≤ -0.9 (representing extremely negative sentiment). Additionally, WORLD has the lowest percentage (9.405%) of comments with sentiment scores ≥ -0.9 (representing extremely positive sentiment). Therefore, we conclude that in WORLD, editors have the highest tolerance of comments with highly negative sentiment. The ARTS section has the highest percentage of comments with a positive sentiment (*e.g.* 24.561% for comments with sentiment $\geq$ 0.9), and the lowest percentage (10.877%) of comments with sentiment score ≤ -0.8. This indicates that editors in ARTS prefer to select positive comments compared with the other sections. Note, this is partly impacted by the more emotive language used in discussing art news. For comments with a sentiment $le - 0.9$, the percentage in UPSHOT is the lowest (7.463%). Although it is unclear exactly why, this indicates that editors responsible for UPSHOT dislike comments with extremely negative sentiment.

Toxicity. We finally conduct K-W tests among the top 10 sections to measure the significance of differences in toxicity. This confirms that these sections are significantly different regarding with preferences of editor in terms of toxicity (p-value is $<$0.001). Table 11 shows the percentage of editor-picked comments that have various toxicity levels. Unsurprisingly, we see that section OPINION has the highest percentage (0.011%) of comments with toxicity $\geq$ 0.8, while section BUSINESS has the highest percentage (0.238%) of comments with toxicity $\geq$ 0.7. The latter is 2x the OPINION (0.120%) and 10x the US (0.023%). Hence, we conclude that editors in these two sections have a stronger tolerance for highly toxic comments compared with the other 8 sections. This is likely necessary, considering the nature of the sections in question.

Table 8: The percentage of comments at different length levels in the Editor-picks group. For convenience, we embolden any values that are mentioned as examples in the text.

Length	OPINION	US	WORLD	NYREGION	BUSINESS	HEALTH	MAGAZINE	ARTS	SPORTS	UPSHOT
≥ 20	82.899	84.982	82.585	80.717	80.930	90.394	89.850	82.105	83.459	90.299
≥ 40	70.468	69.942	64.260	64.798	68.296	78.571	75.940	64.561	65.414	81.343
≥ 60	56.466	53.544	46.966	47.982	48.153	66.502	56.767	48.421	50.752	67.910
≥ 80	44.058	39.883	34.466	35.202	35.757	52.463	44.737	36.842	37.218	57.463
≥ 100	33.450	29.357	25.425	**25.112**	25.507	41.133	33.083	28.070	28.571	**43.284**
≥ 120	26.040	21.193	17.476	19.283	18.236	33.005	25.940	22.456	23.308	37.313
≥ 140	19.906	15.275	13.046	14.798	14.899	24.384	21.053	16.140	18.421	30.597
≥ 160	15.115	11.298	**9.648**	10.650	10.846	20.197	18.045	12.281	15.414	**26.119**

Table 9: The percentage of comments at different similarity score levels in the Editor-picks group. For convenience, we embolden any values that are mentioned as examples in the text.

Similarity	OPINION	US	WORLD	NYREGION	BUSINESS	HEALTH	MAGAZINE	ARTS	SPORTS	UPSHOT
≥ 0.67	**43.992**	55.673	56.250	55.381	56.615	66.256	60.526	55.439	**72.556**	45.522
≥ 0.70	30.427	40.094	41.687	39.013	40.048	54.433	45.865	40.351	51.128	31.343
≥ 0.73	18.487	26.363	27.367	24.439	25.268	37.438	28.571	25.965	36.466	14.179
≥ 0.76	9.844	14.643	14.563	14.013	14.422	24.138	11.278	13.684	22.556	3.731
≥ 0.79	4.289	6.854	6.371	7.175	6.198	11.330	6.015	5.263	9.398	1.493
≥ 0.82	1.637	2.596	1.760	1.906	1.788	4.433	3.008	1.404	4.135	0.746
≥ 0.85	0.589	0.725	0.243	0.785	**0.119**	**1.478**	0.752	0.702	1.128	0.746
≥ 0.88	0.131	0.164	-	0.224	-	-	-	0.702	**1.128**	-

6 RQ3: Automating Editor Picks

The task of producing editor-picks (selecting comments by editors) is labour intensive. **RQ3** seeks to build tooling to automate the identification of editor-picks. Therefore to answer this RQ, we next develop classifiers that can help semi-automate this process by generating a shortlist of comments suitable as Editor-picks.

6.1 Editor-Picks Classifier Design

To underpin our tooling, we experiment with a number of classification engines to automatically discriminate suitable editor-picks from zero-picks comments. Recall that our analysis in Section 5 revealed significant differences in editor criteria across different news sections. Thus, we train individual classifiers for each section. As the top 10 largest sections cover 92% of all comments, we focus on these. As a baseline, we further build a single "global" model trained using data from all sections.

We experiment with seven machine learning algorithms: Random Forest Classifier (RF), Logistic Regression (LR), K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Gradient Boosting Classifier (GB), Decision Tree Classifier (DT) and Gaussian Naive Bayes (GNB). The number of zero-picks is undersampled in order to make the two classes balanced. During the training process, 5-fold Cross-Validation is utilized, while grid search is exploited to optimize the parameters of each classifier. The parameters of each classifier are listed below:

- **Random Forest**: 'n_estimators': [5, 50, 100, 250]; 'max_depth': [2, 4, 8, 16, 32, None].
- **Logistic Regression**: 'penalty': ['l1','l2']; 'C': [0.001, 0.01, 0.1, 1, 10, 100, 1000].
- **K-Nearest Neighbors**: 'n_neighbors': [1, 3, 5, 7, 9]; 'algorithm': ['auto', 'ball_tree', 'kd_tree', 'brute'].
- **Support Vector Machine**: 'C': [0.001, 0.01, 0.1, 1, 10, 100, 1000]; 'kernel': ['rbf', 'linear', 'sigmoid'].
- **Gradient Boosting**: 'n_estimators': [100, 300, 500, 800]; 'learning_rate': [0.01, 0.1, 1, 10].
- **Decision Tree**: 'max_depth': [2, 4, 8, 16, 32, None].
- **Gaussian Naive Bayes**: 'var_smoothing': [1e-8, 1e-9, 1e-10].

To better understand the features that impact performance, we experiment with using multiple feature sets. The first feature set only uses the text embedding representation of the comments, while the other three experimental feature sets complement the text embedding representation by using comment length, similarity and sentiment, respectively.

6.2 Evaluation

Global model. Table 12 presents the performance (evaluated by F1 score) of the classifiers on the global dataset, trained across all 10 sections. In this table, the **Text only** column refers to the results of the first experimental group (which only leverage comment text embedding for classification); **Text⊕Length** refers to the results of exploiting both text embedding and comment length as features; **Text⊕Sim.** refers to exploiting both text embedding and similarity as features; **Text⊕Senti.** refers to both text embedding and sentiment as features; while **Text⊕All.** refers to exploiting

Table 10: The percentage of comments at different sentiment score levels in the Editor-picks group. For convenience, we embolden any values that are mentioned as examples in the text.

Sentiment	OPINION	US	WORLD	NYREGION	BUSINESS	HEALTH	MAGAZINE	ARTS	SPORTS	UPSHOT
$\leq$ -0.6	25.974	29.544	30.340	30.269	24.553	26.601	23.308	**18.246**	20.677	22.388
$\leq$ -0.7	22.449	25.637	25.850	25.673	20.977	22.414	20.301	**14.386**	17.669	19.403
$\leq$ -0.8	17.494	20.000	20.206	20.291	15.852	16.995	16.165	**10.877**	14.286	12.687
$\leq$ -0.9	10.892	11.953	**12.864**	12.332	8.701	10.099	9.774	8.772	8.271	**7.463**
$\geq$ 0.9	16.097	12.702	**9.405**	11.996	10.965	18.227	19.173	**24.561**	22.556	17.910
$\geq$ 0.8	25.527	21.754	18.507	20.628	21.216	28.571	32.331	37.193	32.331	36.567
$\geq$ 0.7	32.107	27.556	24.575	26.570	27.175	35.961	40.602	45.263	39.098	39.552
$\geq$ 0.6	36.593	32.374	29.369	31.726	33.373	39.901	45.489	48.772	42.105	44.030

Table 11: The percentage of comments at different toxicity score levels in the Editor-picks group. For convenience, we embolden any values that are mentioned as examples in the text.

Toxicity	OPINION	US	WORLD	NYREGION	BUSINESS	HEALTH	MAGAZINE	ARTS	SPORTS	UPSHOT
$\geq$ 0.1	55.888	57.731	59.830	51.570	44.815	29.557	48.496	47.719	43.233	50.746
$\geq$ 0.2	29.826	28.819	28.883	28.139	22.884	11.084	25.564	26.667	18.797	20.149
$\geq$ 0.3	15.421	15.018	13.167	13.117	10.608	4.680	11.654	15.439	8.271	11.194
$\geq$ 0.4	5.009	5.357	3.823	4.148	4.291	1.970	1.880	2.807	2.256	4.478
$\geq$ 0.5	1.572	1.380	0.728	1.233	0.954	0.985	0.752	1.053	0.752	-
$\geq$ 0.6	0.458	0.421	0.243	0.224	0.477	0.493	0.376	1.053	-	-
$\geq$ 0.7	0.120	0.023	-	-	**0.238**	-	-	-	-	-
$\geq$ 0.8	**0.011**	-	-	-	-	-	-	-	-	-

Table 12: Classifier performance for the global model, measured as macro F1-Score.

Model	Text only	Text ⊕Length	Text ⊕Sim.	Text ⊕Senti.	Text ⊕All
RF	0.678	↑ 0.713	↑ 0.688	0.678	↑ 0.721
LR	0.728	0.723	↑ 0.734	0.727	0.722
SVM	0.753	0.733	↑ 0.754	0.742	0.736
KNN	0.669	0.666	↑ 0.678	↑ 0.670	0.664
GB	0.718	↑ 0.729	↑ 0.726	0.712	↑ 0.737
DT	0.599	↑ 0.682	↑ 0.611	↑ 0.630	↑ 0.688
GNB	0.694	↑ 0.697	↑ 0.703	↑ 0.697	↑ 0.708

text embeddings together with all other features (comment length, similarity and sentiment).

The results show that the addition of the meta-features (length, similarity, sentiment) has differing impacts based on the classification model. Whereas some models improve (*e.g.* RF), others actually degrade (KNN). The top performing model is the SVM Text⊕Sim model (0.75 F1). This suggests that the global model *does* offer the potential for assisting editors with identifying suitable picks.

Per-Section Models. We next investigate if our per-section models can improve beyond the global baseline. Figure 4 shows the classification results and plots the F1 score using the per-section models. Due to space limications, Table 13 further highlights the results for the four sections with the highest improvement (ARTS, MAGAZINE, NYREGION and UPSHOT). We observe substantial performance improvements compared to the global baseline. All sections attain an F1 above 0.8. This is driven by the distinct editor behaviors across the different sections (Section 5.2). After adding comment length, similarity, and sentiment features, the performance of the classifiers improves noticeably. For example, there is a near 5% improvement of F1-Score after adding similarity for classification in the MAGAZINE section. The UPSHOT section gains the best F1 (0.86) using Text⊕Senti. These results confirm that the per-section features can help editors select comments. This can reduce the manual review workload of editors by presenting editors with the top-n comments predicted to be suitable picks.

Table 13: Classifier performance for per-section models in the best 4 sections, measured as macro F1-Score.

Section	Text only	Text ⊕Length	Text ⊕Sim.	Text ⊕Senti.	Text ⊕All
ARTS	0.838 (RF)	↑ 0.846 (RF)	↑ 0.845 (RF)	↑ 0.846 (RF)	0.829 (RF)
MAGAZINE	0.755 (RF)	↑ 0.761 (GB)	↑ 0.801 (RF)	↑ 0.783 (GB)	↑ 0.774 (RF)
NYREGION	0.784 (LR)	↑ 0.802 (RF)	↑ 0.787 (SVM)	↑ 0.794 (SVM)	↑ 0.788 (RF)
UPSHOT	0.825 (RF)	↑ 0.842 (RF)	0.825 (RF)	↑ 0.859 (GB)	↑ 0.859 (RF)

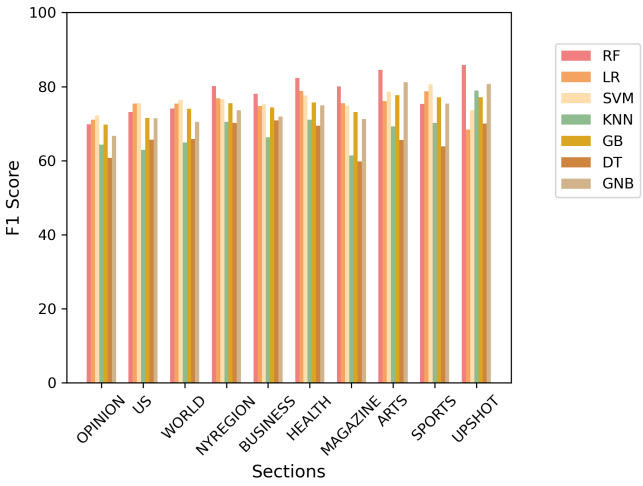


Figure 4: The per-section classification results between Editor-picks and Zero-picks comment categories in the top-10 sections. (The best results selected from experiments trained on different feature sets).

7 Discussion

By exploring comment curation based on the NYT, we have identified factors that impact the selection of editors, found variations of criteria across news sections, and built automatic tooling to identify potential editor-picks. We believe our findings and tooling can play a role in helping journalists and editors to identify relevant comments.

We wish to highlight and discuss certain limitations of our analysis. In part, these stem from the need to treat the editor selection process as a black box. We do not have details on the internal procedures of each editor. This inspires our study, and our findings shed insight into this. However, this means we have a limited vantage of certain complexities, *e.g.* delays caused by moderation queuing and article publication time. We believe studying these logistical aspects would be another interesting line of future work.

As part of this, there might also be a tacit causal relationship between editors' and users' preferences. Indeed, we find that comments picked by an editor have, on average, more recommendations and replies by users. The exact causality for this is unclear. Two possible explanations exist: either the editor is influenced by the users, or the users are influenced by the editor. For instance, editor picks gain higher visibility, therefore expanding the pool of potential people who could reply or recommend. Equally, comments that receive more recommendations and replies are more likely to be seen by editors. Unfortunately, we do not have timestamps for when recommendations were received; nor do we have timestamps for when editors made their decision. This prevents us from studying such causality, yet we argue this would be an interesting line of future work.

As mentioned in Section 3.2, we also considered any comment that receives $\geq$10 recommendations or $\geq$3 replies to be a "user-pick". These thresholds were selected as an intuitive mechanism to identify more popular comments with users. Although we believe these thresholds are a good proxy to measure users' engagement, we hope to experiment with alternative thresholds in the future.

Finally, we emphasize that the New York Times is only a single USA-based news organization, which is considered to have a left-leaning bias. Our findings should therefore be interpreted with this in mind. To better understand generalizability, we wish to expand our analysis to other news outlets with diverse editorial policies and news types. We are also keen to pilot our tooling within news outlets and gain editor feedback. This is because one concern is that our tooling could narrow editor pick selections by over-fitting to prior behavior. Thus, we also wish to perform user experiments.

8 Conclusion

This paper has explored the growing use of comment curation in news media through the lens of the New York Times (NYT). To answer RQ1, we have identified a clear set of factors differentiating editor-picks from other comments in Section 4. The factors that impact the likelihood of comment selection include several psycho-linguistic factors and four other factors (comment length, relevance, sentiment, and toxicity). We found that editor-picks tend to be longer, more relevant to the article, positive in sentiment, and contain lower toxicity. We also found correlations with certain linguistic features, *e.g.* words related to conversation and power are more likely to feature in editor-picks, whereas the use of emojis is less likely. For RQ2, we then explored these factors across individual news sections in Section 5. We confirmed that editors within different news sections exhibit different criteria. For example, editor-picks in the OPINION section are less related to their articles than other sections. Inspired by the first two research questions, RQ3 then explored tooling to support editors in identifying eligible candidate comments to pick in Section 6. Our models accurately predict picks, with F1 scores as high as 0.86. We hope that this could help editors in reducing their manual workload.

9 Ethics and Broader Impacts

We rely on public data, which has been voluntarily published by authors in the knowledge that it will be viewed by readers. We obtain our dataset using the official New York Times API. We cannot release the dataset considering the copyright of NYT. That said, it is possible for other researchers to collect an equivalent dataset, and we make our code available to facilitate this.

We also note there are important positive outcomes of our work. Our goal is to build tooling to help editors easily identify high quality comments. This can improve the quality of comment discourse and reduce the manual tasks of editors, freeing their time to perform deeper manual reviews. Since the tooling selects comments based on prior editors, we do not expect that the approach will suppress voices of people who were previously heard. That said, our models could certainly pick up potential editor biases, *e.g.* avoiding comments from under-served communities. Thus, we adopt a human-in-the-loop model, whereby we only shortlist suitable comments — editors must make ultimate decisions.

References

Boyd, R. L.; Ashokkumar, A.; Seraj, S.; and Pennebaker, J. W. 2022. The development and psychometric properties of LIWC-22. *Austin, TX: University of Texas at Austin*, 1–47.

Diakopoulos, N. A. 2015. The editor's eye: Curation and comment relevance on the New York times. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, 1153–1157.

Ferrucci, P.; and David Wolfgang, J. 2021. Inside or out? Perceptions of how differing types of comment moderation impact practice. *Journalism Studies*, 22(8): 1010–1027.

Hutto, C.; and Gilbert, E. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, volume 8, 216–225.

Jang, B.; Kim, I.; and Kim, J. W. 2019. Word2vec convolutional neural networks for classification of news articles and tweets. *PloS one*, 14(8): e0220976.

Jigsaw. 2022. Perspective API.

Juarez Miro, C. 2022. The comment gap: Affective publics and gatekeeping in The New York Times' comment sections. *Journalism*, 23(4): 858–874.

Karpova, E.; Best, K. L. R.; and Bayat, F. 2020. Who Is Part of the Problem?: Critical Analysis of New York Times Readers' Comments on the Environmental Cost of Fashion Consumption. In *International Textile and Apparel Association Annual Conference Proceedings*, volume 77. Iowa State University Digital Press.

Kesarwani, A. 2018. New York Times Comments: Comments on articles published in the New York Times. https://www.kaggle.com/aashita/nyt-comments.

Kolhatkar, V.; and Taboada, M. 2017a. Constructive language in news comments. In *Proceedings of the first workshop on abusive language online*, 11–17.

Kolhatkar, V.; and Taboada, M. 2017b. Using New York Times Picks to Identify Constructive Comments. In *Proceedings of the 2017 EMNLP Workshop: Natural Language Processing meets Journalism*, 100–105. Copenhagen, Denmark: Association for Computational Linguistics.

McInnis, B.; Ajmani, L.; Sun, L.; Hou, Y.; Zeng, Z.; and Dow, S. P. 2021. Reporting the Community Beat: Practices for Moderating Online Discussion at a News Website. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2): 1–25.

Muddiman, A.; and Stroud, N. J. 2016. 10 Things We Learned by Analyzing 9 Million Comments from The New York Times. *Mediaengagement. Org*.

Naeem, M. Z.; Rustam, F.; Mehmood, A.; Ashraf, I.; Choi, G. S.; et al. 2022. Classification of movie reviews using term frequency-inverse document frequency and optimized machine learning algorithms. *PeerJ Computer Science*, 8: e914.

Park, D.; Sachar, S.; Diakopoulos, N.; and Elmqvist, N. 2016. Supporting comment moderators in identifying high quality online news comments. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 1114–1125.

Reimers, N.; and Gurevych, I. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Shekhar, R.; Pranjić, M.; Pollak, S.; Pelicon, A.; and Purver, M. 2020. Automating News Comment Moderation with Limited Resources: Benchmarking in Croatian and Estonian. *Journal for Language Technology and Computational Linguistics (JLCL)*, 34(1).

Stroud, N. J.; Van Duyn, E.; and Peacock, C. 2016. News commenters and news comment readers. *Engaging News Project*, 1–21.

Tausczik, Y. R.; and Pennebaker, J. W. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of language and social psychology*, 29(1): 24–54.

Wang, Y.; and Diakopoulos, N. 2022. Highlighting High-quality Content as a Moderation Strategy: The Role of New York Times Picks in Comment Quality and Engagement. *ACM Transactions on Social Computing (TSC)*, 4(4): 1–24.

Wei, H.; Zheng, W.; Xiao, Y.; and Dong, C. 2021. News-Comment Relevance Classification Algorithm Based on Feature Extraction. In *2021 International Conference on Big Data Analysis and Computer Science (BDACS)*, 149–152. IEEE.

Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Wolfgang, J. D. 2018. Cleaning up the "Fetid Swamp" examining how journalists construct policies and practices for moderating comments. *Digital journalism*, 6(1): 21–40.