The lack of theory is painful: Modeling Harshness in Peer Review Comments

Rajeev Verma[1], **Rajarshi Roychowdhury**[2], **Tirthankar Ghosal**[3]
[1]Independent Researcher
[2]Jadavpur University, India
[3]Charles University, Faculty of Mathematics and Physics,
Institute of Formal and Applied Linguistics, Czech Republic
[1]rajeev.ee15@gmail.com, [2]rroychoudhury2@gmail.com, [3]ghosal@ufal.mff.cuni.cz

Abstract

The peer-review system has primarily remained the central process of all science communications. However, research has shown that the process manifests a power-imbalance scenario where the reviewer enjoys a position where their comments can be overly critical and wilfully obtuse without being held accountable. This brings into question the sanctity of the peer-review process, turning it into a fraught and traumatic experience for authors. A little more effort to still remain critical but be constructive in the feedback would help foster a progressive outcome from the peer-review process. In this paper, we argue to intervene at the step where this power imbalance actually begins in the system. To this end, we develop the first dataset of peer-review comments with their real-valued *harshness* scores. We build our dataset by using the popular *Best-Worst-Scaling* mechanism. We show the utility of our dataset for text moderation in peer reviews to make review reports less hurtful and more welcoming. We release our dataset and associated codes in https://github.com/Tirthankar-Ghosal/moderating-peer-review-harshness. Our research is one step towards helping create constructive peer-review reports.

1 Introduction

The peer-review system has largely remained the central and universal quality control system in all scientific fields. Hyland and Jiang (2020) argues that the peer-review system embodies *Universalism* and *Organized skepticism* where the former means "an adherence to objectivity rather than self-interest," and the latter calls to the spirit that "no theory is accepted merely on the authority of the proponent." Both these goals are crucial to the success of this science scrutiny system that has been the *de-facto* method for scientific validation for ages. Nonetheless, the past few years have put this system to stress test with ever-increasing research submissions (Ghosal et al., 2019a), a dearth of experienced reviewers, and criticisms like exclusionary, arbitrary, inconsistent, etc. being leveled at this fundamental process of science evaluation (Ghosal, 2022). These challenges have the potential to turn this central process into a *fraught*, and *traumatic* experience, especially for young authors when the reviewers are overly critical or wilfully obtuse (Wilcox, 2019). In an ever-increasing competition in the academic job market, where the career of researchers depends on the impact and prestige of where their work is published, this leads to a natural disdain among the authors for the peer-review process, which is laden with these critical issues. While the peer-review process is by definition a process to evaluate the research under submission — *a litmus test to separate the sweet from the sour*[1], sometimes what hurt the most to the enthusiastic prospective author is the way reviewers express themselves in the reviews. Hyland and Jiang (2020) notes that *"review comments can be blunt, perhaps because of reviewer anonymity, a hurried report, personal style, or even a lack of pragmatic experience."* They also express that the peer-review process exhibits a power imbalance:

"The very act of evaluating another's work is a thinly disguised instructional relationship of authority; an inherently unequal interaction because the power to criticise is non-reciprocal and lies exclusively with the reviewer. This is perhaps made more threatening by the fact that reviewers are "mysterious and intimidating figures" (Tardy, 2018), masked by anonymity, with the power to influence our professional lives. Clearly, reviewers' reports can be demoralizing, and while anonymity might help prevent personal bias, it can make reviewers less accountable."

Towards the overarching goal of improving the

[1]https://www.humanities.hk/news/this-paper-is-absolutely-ridiculous-ken-hyland

Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 925–935
November 20–23, 2022.

review quality standards and making the peer-reviewing process more inclusive, an interesting direction would be to *intervene* at the very step where this *power imbalance* actually begins. Present-day scientific progress is critically dependent on the peer-review process. Hence an inclusive and constructive environment is critical to foster a progressive scientific temperament. Here in this work, we intend to make the review reports more welcoming so that they do not seem *hurtful* and actually focus on their intended objective, i.e., to provide *helpful* feedback to the authors on their submitted manuscript. Given the scale of the peer-review process, an automatic system for this *intervention* would be of high value. Here, we model the various facets of how review comments can be perceived as *hurtful*, a quality we henceforth call as *harshness*. We build upon the reviewer guidelines in major Artificial Intelligence (AI) conferences to categorize how this *harshness* is expressed in the peer-review reports. We use a comparative annotation scheme, called *Best-Worst-Scaling*, to map review sentences into real-valued harshness scores and make this dataset publicly available. We envision that our research and accompanying dataset will be helpful in automatic peer-review text moderation.

Let us study a recent example from a meta-review in NeurIPS 2021, which was rather harsh and unnecessary[2]:

"I do have experience with social science research, and this paper lacks insightfulness or originality from that perspective, so I recommend rejection," and *"This paper will eventually be published somewhere, but it won't have great impact."*

On gaining visibility and criticism in social media on these open access reviews[3], these comments were later manually moderated. Thus previous research and evidence such as the above example show that *unkind* review comments are common. Due to the confidential nature of the reviewing process, reviewers do not disclose their identity and hence cannot be held accountable for their unprofessional and unnecessary hard comments. Hence this phenomenon has the potential to *silently* make the whole publishing process a traumatic experience for researchers.

Our dataset can be used to filter out review sentences based on different thresholds to detect *impolite* review comments. A system to predict a *harshness score* of review sentences would help (senior) area chairs or editors to not allow such comments to go out in public or to the authors. Similarly, a reviewer-assistant tool could use such a predictor to flag/alert reviewers when they write such *harsh* comments (or are repeated offenders). We understand that the peer-review process and *harshness* is inherently a subjective phenomenon. However, we should strive to make the peer-review process more welcoming so that the fundamental process of scrutinizing science remains *objective*. Our current work is a step in that direction.

2 Related Work

There is a growing body of literature on Natural Language Processing (NLP) for peer reviews and scientific literature in general. For example, datasets like PeerRead (Kang et al., 2018), CiteTracked (Plank and van Dalen, 2019), ASAP-Review (Yuan et al., 2021), Peer-Review-Analyze (Ghosal et al., 2022) are proposed in the literature to support NLP research on few downstream problems in peer-reviews. Recently, Bharti et al. (2022a) proposed a binary-class dataset to determine if a peer-review statement is constructive or not. Among the computational approaches, Ghosal et al. (2019b); Kumar et al. (2022) use sentiment information in peer-review comments to predict the reviewer recommendation score and the acceptance/rejection decision of a manuscript. Wang and Wan (2018); Kumar et al. (2021) proposed deep neural methods for sentiment analysis on peer reviews. Our work is different from their works as we model the *harshness* of a review comment, which is a much richer signal than sentiment label or intensity. In essence, our work is closer to hate speech, and offensive language detection research in NLP (Schmidt and Wiegand, 2017; Waseem and Hovy, 2016; Davidson et al., 2017; Founta et al., 2018a; Sap et al., 2020; Breitfeller et al., 2019). However, we assert that our investigation on hurtfulness or offensiveness in peer-review texts differs from the regular toxicity and abusiveness studied in these works. Here we are working on scientific peer-review texts where these notions of harshness are usually very subtle due to the formal academic style of writing. Secondly, much of the hate speech research in NLP is focused on some targeted groups depending on factors like race, gender, ethnicity, etc. Some aspect of our work resem-

[2] https://twitter.com/Abebab/status/1464230544619806720

[3] The NeurIPS conference uses the open review platform: https://openreview.net

bles Wulczyn et al. (2017). They study aggression, personal attacks, and toxicity in Wikipedia Talk pages, where aggression and personal attacks also manifest the *harshness* that we model in this paper. However, their work is not directly applicable to us due to the different domains (Wikipedia vs. peer-reviews). Our methodology to map review comments to a real-valued score is similar to Hada et al. (2021), who also uses Best-Worst-Scaling (BWS) to map Reddit comments to real-valued offensiveness scores. To our knowledge, this is the first work towards developing resources and computational approaches for text moderation in the peer-review domain.

3 Definition of Review Harshness

We define *review harshness* as a metric encompassing two orthogonal dimensions. The first dimension concerns the *evaluative focus* of the comment, and the second dimension deals with the comment's *critical stance*.

3.1 Critical Stance

Peer reviews evaluate the submitted research work across several criteria, such as novelty, correctness/soundness, impact, appropriateness, etc. As such, review texts can be (and are expected to be) critical in their expression. By *harshness* in review texts, we not only mean the presence of criticality or the negative sentiment in them but how these attributes are expressed. Hyland and Jiang (2020) studies *critical stance* in purported harsh peer-review comments as *"features which refer to the ways writers present themselves and convey their judgements, opinions, and commitments..."*, and identify *evidentiality*, *effect*, and *presence* as the key markers of such expression. Evidentiality deals with the use of hedges and boosters (Ghosal et al., 2022) to signal the certainty of a statement. Presence means using first-person pronouns and possessive determiners to express authority. Affect concerns the use of attitude markers to express the attitude of the reviewers emphatically. Furthermore, Boosters (Evidentiality) and Self-mention (Presence) make up the most frequently occurring markers signaling the reviewer's conviction in their judgment, eliminating all doubts about their opinions in an authoritative manner. Hyland and Jiang (2020) mention a clear downplay of power imbalance here where harsh review comments are served *without dressing or varnish*. Interestingly, our example peer review comment (in Section 1) from NeurIPS 2021 contains two of these markers: *evidentiality* - "it **won't** have great impact," and *presence* - "**I do** have experience."

3.2 Evaluative Focus

This dimension deals with the actual content of the review comments. Building upon the reviewer guidelines for the IEEE Conference on Computer Vision and Pattern Recognition (IEEE CVPR), we identify several facets of review texts that are unwelcoming and demonstrate *bad* reviewing practices. Some of these practices are also mentioned in Rogers and Augenstein (2020). These include:

1. **Blank Assertions and Pure Opinions** These are ungrounded statements with no evidence to support the reasoning. Peer reviews are supposed to be the objective evaluation of the submitted work and should provide actionable comments to the authors. These ungrounded statements can sometimes take a very disparaging tone and blatantly attack authors, and the overall research (Hyland and Jiang, 2020).

2. **Intellectual Laziness and Novelty Fallacy** *Intellectual Laziness* refers to narrow-minded reviewing practices. Instead of focusing on a comprehensive evaluation of the submitted research, reviewers can sometimes choose to overemphasize certain factors. For example, if the paper surpassed the state-of-the-art (SOTA) results, (Rogers, 2020a), minor issues like writing and presentation style, minor issues that can be easily fixed, etc. Similarly, reviewers penalize simple methods, non-mainstream research (Rogers and Augenstein, 2020), etc. *Novelty Fallacy* refers to the rigid fixation to the novelty criteria, and not focusing on whether the concerned research advances scientific knowledge even if it is not significantly novel.

3. **Policy Entrepreneurism** stands for reviewers imposing their own policies in review comments which are against sound scientific reviewing practices. For example, sometimes reviewers ask the authors to compare with a recent arXiv preprint (not peer-reviewed or a contemporaneous article), reviewers in some venues show bias against resource papers (Rogers, 2020b; Rogers and Augenstein,

2020), some reviewers show bias against empirical research and demands theorems and theoretical results[4], etc.

We note that the boundaries across the above categories are ill-defined, making the categorical annotation challenging. We further assert that both the dimensions of our definition are orthogonal to each other, and the harshness score is a monotonically increasing function of both these two dimensions.

4 Dataset Source and Curation

Access to peer reviews is still restricted since much of the peer-review system operates behind closed doors. Fortunately, many venues in Artificial Intelligence research have adopted an open-access peer review platform called OpenReview[5] to manage the reviewing procedure. For our study, we make use of the Peer-Review-Analyze dataset (Ghosal et al., 2022). Peer-Review-Analyze contains 1199 reviews ($\sim 17K$ review sentences) from the 2018 edition of the International Conference on Learning Representations (ICLR). The ICLR reviewing process operates in the OpenReview platform. Each review sentence in this dataset is annotated for review-paper section correspondence, review-paper aspect category, review-statement purpose, and review-statement significance, along with their associated sentiment label (POS, NEG, NEU). Please refer to the original paper (Ghosal et al., 2022) for full details on the dataset. Our goal in this study is to model *harshness* in peer-review sentences. However, annotating each of the $17K$ sentences individually is expensive. As indicated in the paper, most of these review sentences are neutral in sentiment due to the inherent academic style in writing reviews. We, therefore, use an Active Learning technique to efficiently create a smaller collection of potentially *harsh* sentences. Active Learning assumes access to a small seed dataset for its operationalization. Active Learning aims to select the most informative samples for labeling according to some uncertainty or diversity measures. We refer the reader to Ren et al. (2021) for an exhaustive survey on active learning techniques in deep learning.

As a seed dataset, we crawl 1093 review sentences using the Twitter API[6] from the public Twitter handle *ShitMyReviewersSay*[7]. The Twitter handle *ShitMyReviewersSay* is a dedicated public platform where authors can anonymously post their review sentences that they find unwelcoming, disparaging, scathing, or discouraging. It tweets self-explanatory review sentences from diverse scientific backgrounds, which authors share to vent their frustrations. Since authors made the efforts to share these review comments on a public forum, we consider them to be a gold standard of the *harshness* we aim to model. However, these sentences are also extreme in their tone and are not representative of subtle/intrinsic *harshness* in most academic reviews. Therefore, we use both the samples from ICLR and *ShitMyReviewersSay* in our final annotations to model a more generic *harshness* scale.

4.1 Active Learning

In this work, we use the Cartography Active Learning (CAL) algorithm (Zhang and Plank, 2021) for sampling. CAL is a model-agnostic active learning sampling procedure based on *data-maps* (Swayamdipta et al., 2020). Specifically, it considers the training statistics of a model on a seed dataset to select informative samples. Swayamdipta et al. (2020) showed that the training dynamics of a downstream model on individual instances results in categorization of the input samples in the dataset into three categories, *ambiguous* examples, *easy-to-learn* examples, and *hard-to-learn* examples. CAL proposes to query *ambiguous* examples for labeling as these are the examples the model would learn from the most. Procedurally, it uses a limited labelled seed dataset $\mathcal{L}$ to train a classifier f_{θ^*} and record training statistics, namely *confidence*, *variability*, and *correctness* for each example in the seed data. It then uses information from the training statistics to train another binary classifier g_{ϕ^*} on the representations of f_{θ^*} to demarcate the decision boundary between *hard-to-learn* and *ambiguous* examples. It then uses g^*_ϕ to sample examples from the pooled unlabelled dataset $\mathcal{U}$ for labeling. It is an iterative procedure, where after each iteration, the newly labeled examples from $\mathcal{U}$ are added to $\mathcal{L}$, and the procedure is repeated. We refer the reader to the original paper (Zhang and Plank, 2021) for a complete description of the algorithm.

Our goal in this paper is to sample the subtle/implicit cases of *harsh* comments from the aca-

[4] https://twitter.com/tomgoldsteincs/status/1484609309778587653

[5] https://openreview.net/

[6] https://developer.twitter.com/en/docs

[7] https://twitter.com/yourpapersucks?lang=en

demic peer-review texts. We reason such comments lie in between the two extremes of rather explicitly *harsh* comments from *ShitMyReviewersSay* (`class 1`) and the more academically factual comments in ICLR (`class 2`). Furthermore, we hypothesize that such comments would be *ambiguous* for a classifier trained to predict whether a sentence belongs to `class 1` or `class 2`. Thus, we can create $\mathcal{L}$ by picking examples from both the classes and running CAL to sample *ambiguous* samples. However, it marks a majority of valid negative sentiment sentences (and not *harsh*) as *ambiguous*. Here, we would like to note that the Peer-Review-Analyze dataset contains the sentiment (POS-NEG-NEU) associated with the review comment as well. We found boosting `class 2` with positive and valid negative sentences works better. We, therefore, create our seed dataset $\mathcal{L}$ by randomly picking 250 examples from the *ShitMyReviewersSay* set and 750 examples from Peer Review Analyze dataset split equally across all the three sentiments (POS, NEG, NEU) classes. For our pooled unlabelled dataset $\mathcal{U}$, we consider all the remaining NEG sentiment sentences from the Peer Review Analyze dataset. We run CAL based on the defined $\mathcal{L}$ and $\mathcal{U}$, and create a smaller set of 391 potentially *harsh* review comments. To maximize the diversity of the dataset for final annotation, we inflate this set to 500 samples by randomly including NEG sentiment sentences from the Peer Review Analyze dataset.

5 Annotation Process

As stated before, we aim to model review comment *harshness* on a real-valued scale. Our choice is motivated by the fact that a review text can be hurtful/harsh to a varying degree and by the downstream application of more fine-grained review text moderation. Contrary to the categorical annotation of marking whether a review comment is hurtful or not (Bharti et al., 2022a), we employ the comparative annotation mechanism. We argue that eliciting categories for review comment harshness is challenging due to the inherent subjective perceptual nature of the task. Additionally, such an annotation procedure is not reliable and could lead to ambiguities and inconsistencies (Founta et al., 2018b). We argue that these issues can manifest to a greater degree due to the academic nature of our data. All these problems can be mitigated using a comparative annotation setup (Asaadi et al., 2019; Kiritchenko and Mohammad, 2017). The comparative annotation works by asking the annotator which one among the two samples demonstrates the desired quality to a greater extent. This is more suited for our academic data, as comparing two review comments to see which one is more hurtful is an easier task. We use the *Best-Worst-Scaling* (BWS) (Louviere, 1991; Louviere et al., 2015; Kiritchenko and Mohammad, 2016, 2017) setup for our annotation.

5.1 Best Worst Scaling (BWS)

For N samples, a naive comparative annotation mechanism would need to compare N^2 pairs. This is obviously expensive in practice. BWS is an efficient comparative annotation mechanism where we need only $2N$ comparisons. However, instead of comparing in a pair, we ask our annotators to mark a `Best Item` and a `Worst Item` according to some quality of interest in a set of four comments (4-tuple). We follow Kiritchenko and Mohammad (2016) to obtain 4-tuples according to a generation procedure called *random-maximum-diversity-selection (RMDS)*. RMDS aims to maximize the diversity (according to the quality of interest) in a tuple by maximizing the number of items that each item co-occurs with. This way, $2N$ distinct 4-tuples are generated, such that each comment is seen in 8 different 4-tuples, and no 2 4-tuples have more than 2 items in common. This process aims to cover the entire range of the quality of interest in each tuple. We then convert the `Best Item`, and `Worst Item` annotations from BWS to the real-valued scores using a simple counting procedure (Orme., 2009; Flynn and Marley, 2014), associating with each sample a real-valued score according to the quality of interest. For each example, this score is the proportion of times the given example is chosen as the `Best Item` minus the times the concerned example is chosen as the `Worst Item`.

5.2 Annotation Tool and Annotators

For our task, `Best Item` stands for the most *harsh* review comment, and `Worst Item` means the least *harsh* comment. In simple terms, our annotation task refers to showing each annotator a 4-tuple of review comments and asking them to select which is the most *harsh* comment and which is the least *harsh* comment. Since *harshness* is a subjective perceptual quality, crowdsourcing annotations would have been ideal. However, we are working with specific scientific data which requires

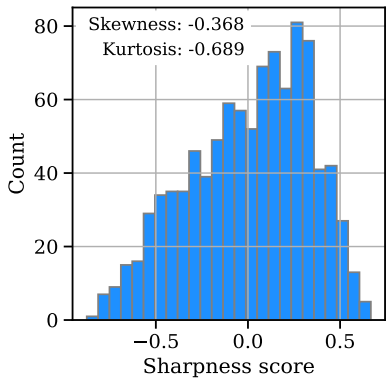

Figure 1: *Histogram of the Harshness (harshness) score.* As can be seen, the distribution of the sample scores is moderately left-skewed and has "thinner" tails.

some training to get acquainted with. Therefore, we deliberately hire annotators from diverse academic backgrounds. We hire six annotators; four hold graduate degrees in Linguistic and English Literature, one holds a bachelor's degree in Computer Science and Engineering (CSE), and another is an undergraduate student in CSE. The annotators are duly paid according to the annotation payment standards in India. Each annotator underwent an exposition and training session about the *Evaluative Focus* dimension in our definition of *harshness*. We asked each annotator to read Hyland and Jiang (2020) paper to understand the *Critical stance* dimension. Additionally, we had each annotator take a challenge annotator test to check their readiness for the task. During the annotation period, we held weekly meetings to discuss their doubts and resolve their concerns. However, we strictly asked annotators to not discuss specific comments with each other, and with the authors. We developed a simple easy-to-use annotation tool as an in-house web application hosted on Amazon Web Services (AWS) for the purpose. We carried out the data annotation for a month.

5.3 Data Annotation

In order to cover the entire range of *harshness* scale, we use 500 samples randomly selected from the *ShitMyReviewersSay* set, and 500 samples as procured from the process described in section 4.1. Thus, we have $N = 1000$, resulting in 2000 tuples for BWS. We have six annotators, and since each review comment is seen in eight different 4-tuples, we get 48 judgments per review comment.

5.4 Reliability of Annotations

To calculate the reliability of our annotations obtained through BWS, we use *split-half-reliability* (SHR) values over 10 trials. SHR is a commonly used metric to calculate internal consistency, a desirable quantity for the annotations to be reliable. We follow the methodology in Hada et al. (2021) and compute the SHR values by splitting the annotations for 4-tuples in our dataset in two halves to determine the two sets of rankings. We then measure the correlation between these two rankings; a higher correlation means higher consistency. We repeat this procedure for 10 trials and calculate the final average correlation across these trials to be 0.73, indicating good annotation reliability. We found that 10 trials were sufficient to converge to the final correlation value, and further increasing the number of trials does not significantly affect the average correlation value.

6 Data Analysis

Our final dataset contains 1000 review sentences annotated for their *harshness* value on a scale of -1.0 (most *harsh*) to 1.0 (least *harsh*). In this section, we study the distribution of the *harshness* score and qualitatively examine the samples on varying positions in the *harshness* scale.

Distribution of *Harshness* Scores We visualize the histogram of the *harshness* scores in our sample dataset in Figure 1. We can see that the distribution of the scores in our sample is moderately left-skewed (skewness metric = -0.368). We further infer the population *harshness* scores using the widely known statistical test for skewness (Duncan, 1997). We calculate the test statistic $t = skewness/SES$, where SES means the standard error of skewness defined as $SES = \sqrt{\frac{6N(N-1)}{(N-2)(N+1)(N+3)}}$. The calculated test statistic for our test is $t = -4.699$, which suggests that the population *harshness* scores are skewed negatively with high confidence. This observation is not surprising, as most of the academic writing is formal, and very *harsh* (overly sentimental/caustic, etc.) texts are a rare class in an academic context. However, this observation also asserts the challenges in modeling the *harshness* of peer-review comments. Our methodology of using Active Learning and comparative annotations through BWS efficiently circumvents these issues and closely models a statistic of *harshness* scores in peer-review

Bin	Review comment	Score
1	**a).** An article like this is just a waste of peer-reviewing resources.	-0.708
	b). This paper reads like a woman's diary, not like a scientific piece of work.	-0.625
	c). The manuscript is a collection of fragmented and disconnected descriptive observations.	-0.667
	d). What were you thinking?	-0.625
2	**a).** The lack of theory is painful at times.	-0.521
	b). The author should abandon the premise that his work can be considered research.	-0.583
	c). A failing course paper written by an undergrad.	-0.438
	d). Overall, I think this manuscript is a waste of time.	-0.562
3	**a).** I don't see much science in this manuscript.	-0.333
	b). Many questions on the text, for example, cause embarrassment in understanding the text.	-0.250
	c). Most part of methodology is useless, most of the paragraphs are irrelevant to the main topics.	-0.333
	d). The authors use a log transformation, which is statistical machination, intended to deceive.	-0.396
4	**a).** None of these results beat state-of-the-art deep NNs.	-0.188
	b). Your proposed method should be compared with another method that introduced in a prestigious paper.	-0.001
	c). That can hardly be true (if it is, then it puts the entire paper into question! If trivial uncertainty is almost as good as this method, isn't the method trivial, too?).	-0.021
	d). I don't believe in simulations.	-0.188
5	**a).** They do not really provide any substantial theoretical justification why these heuristics work in practice even though they observe it empirically.	0.083
	b). The results look like a smorgasbord of data	0.021
	c). Unfortunately, in your Figure 2, this is not as obvious and not real since it is using simulated delays.	0.042
	d). Furthermore, the paper lacks in novelty aspect, as it is uses mostly well-known techniques.	0.083
6	**a).** Since the adaptions to DTP are rather small, the work does not contain much novelty.	0.208
	b). RBMs are not state-of-the-art in topic modeling, therefore it's difficult to assess whether this is helpful.	0.375
	c). there is not much innovation in the model architecture.	0.208
	d). From a novelty standpoint though, the paper is not especially strong given that it represents a fairly straightforward application of (Andrychowicz et al., 2016).	0.312
7	**a).** the paper suffers from many problems in clarity, motivation, and technical presentation.	0.458
	b). The authors need to provide more justification for this motivation.	0.417
	c). The legends in the figures are tiny, and really hard to read.	0.438
	d). The text is also difficult to follow. The three contributions seem disconnected and could have been presented in separate works with a more deeper discussion.	0.479
8	**a).** It is not clear what is the stopping criterion for each of the methods used in the experiments.	0.604
	b). Some of the figures are hard to read (in particular Fig 1 & 2 left) and would benefit from a better layout.	0.604
	c). It would, however, seem that the truncated iterations do not result in the approximation being very accurate during optimization as the truncation does not result in the approximation being created at a mode.	0.521
	d). The paper misses some more recent reference, e.g. [a,b].	0.521

Table 1: *Representative sample comments and their scores across 8 bins on the harshness scale.*

comments.

Qualitative Analysis We further analyze our dataset to gauge the patterns along the continuous *harshness* scale. For this, we split the scale into 8 bins, Bin 1: $score \leq -0.6$, Bin 2: $-0.6 \leq score \leq -0.4$, Bin 3: $-0.4 \leq score \leq -0.2$, Bin 4: $-0.2 \leq score \leq 0.0$, Bin 5: $0.0 \leq score \leq 0.2$, Bin 6: $0.2 \leq score \leq 0.4$, Bin 7: $0.4 \leq score \leq 0.5$, and Bin 8: $score \geq 0.5$. We list representative samples from each bin along with the associated score in Table 1. We can see that as the *harshness* score increases from one end to another, the review comments go from extremely disparaging (Bin 1) to standard review comments (Bin 8). Furthermore, review comments across bins also manifest specific qualities according to our definition of *harshness*, denoting that the modeled continuous *harshness* scale capture these properties. For example, comments from Bin 4 exhibit "intellectual laziness" (4a. fixation on SOTA), "policy entrepreneurism" (4b. comparison to a prestigious paper), "personal opinions" (4c. not believing in simulations). Similarly, some comments from Bin 5 and Bin 6 show "novelty fallacy". However, comments in Bin 7 and Bin 8 are standard review comments. These observations also show that one can easily employ a threshold on the scale to filter out *harsh* review comments based on some criteria.

7 Baseline Prediction Models

In this section, we use common computational models to predict the *harshness* scores for review comments. Our problem is a regression task; for each review sentence $\mathbf{s}$, predict the real-valued score. Since we have a relatively smaller size dataset, we use 5-fold cross-validation to evaluate the predictive models. Furthermore, to account for outliers in the dataset, we use smooth L1-loss instead of the regular mean squared error (MSE) loss for the regression task. Besides the regression task, we

Models → Metric ↓	ASE	BiLSTM	BERT	HateBERT
L1-Loss	1.870 ± 0.050	1.629 ± 0.071	1.536 ± 0.112	1.521 ± 0.092
Accuracy	61.12 ± 0.012	67.35 ± 0.009	71.23 ± 0.005	72.08 ± 0.047

Table 2: *Benchmark Results for Common Predictive Models both in Regression and Classification Setting.* We report average L1-loss metric (Regression) and Accuracy (Classification) across all the five folds of cross-validation.

also use the predictive models in the classification setting. As we have seen earlier, different regions on the *harshness* scale show different properties. Therefore, we categorize our dataset into 3 different classes based on the score; class 1 means the score is less than -0.2, class 2 for a score between -0.2 and 0.3, and class 3 for a score greater than 0.3. In this way, class 1 has disparagingly *harsh* comments, class 2 contains review comments exhibiting bad reviewing practices, and class 3 contains regular review comments. In the next subsection, we describe our baseline models for prediction.

7.1 Models

7.1.1 Average Sentence Embeddings (ASE)

We construct the review comment representation using the average of the word embeddings. We use 300 dimensional GoogleNews word2vec vectors for this and pass the sentence representation to the feedforward linear layers for prediction.

7.1.2 Bidirectional LSTM

We use the LSTM (Hochreiter and Schmidhuber, 1997) networks using word2vec word vectors (Mikolov et al., 2013). Specifically, we use 300 dimensional GoogleNews word vectors and use the representations from a 2-layered BiLSTM model to predict the *harshness* score.

7.1.3 BERT

We finetune the pre-trained BERT model (Devlin et al., 2019), specifically bert-base-large using Huggingface (Wolf et al., 2020). The model takes a review text as the input, and the review representation is taken from the [CLS] token, which is then passed to the feedforward linear layers for prediction.

7.1.4 HateBERT

Our task of predicting *harshness* score for review comments somewhat resembles the task of abusive language and toxicity prediction in NLP. Therefore, we also use a standard benchmark for our dataset. We finetune the HateBERT model (Calvetti and Reichel, 2003) on our dataset. HateBERT is a pre-trained BERT model for abusive language detection and outperforms the regular BERT model for abusive language detection.

7.2 Training Setting

For all our models, we use a learning rate of $1e-3$ and a batch size of 32. For ASE and BiLSTM models, we use the Adam optimizer with a weight decay of $1e-3$. For the BERT model, we use the AdamW optimizer. Since the *harshness* score lies between -1 to 1, we use *tanh* non-linearity function at the final prediction layer in all our regression task models. We use Pytorch to implement the models.

7.3 Results

The results for our benchmark models are shown in Table 2. We can see that BERT models perform better in both task settings. However, what is interesting to see is that HateBERT does not provide greater performance gains compared to the regular BERT model. This signifies that the nature of *harshness* in peer-review comments is different that toxicity and abusiveness as it is studied widely in the NLP literature. Thus, there is a great scope for improvement for better predictive models to detect the *harshness* of the review scores.

8 Conclusions

The peer-review process is central to all science research dissemination. However, it also exhibits a power-imbalance situation where the review comments can be overly critical and sometimes cross the boundaries to disparage while also demonstrating bad reviewing practices. This makes this process traumatic, especially for young researchers. The responsibility to moderate these review comments lies in the hands of (senior) area chairs and editors. However, it is not easy to manually moderate review comments with ever-increasing submissions in major AI conferences. In this work, we present a *first-of-its-kind* dataset of 1000 peer-review comments annotated for their *harshness* value. We define *harshness* in this paper based on two dimensions, critical stance and the evaluative

focus of the review comment. We then use a comparative annotation technique, Best-Worst-Scaling (BWS), to elicit a continuous real-valued *harshness* scale. Our analysis shows that the different regions of this scale represent different facets of *harshness* with comments going from disparaging at one end to standard evaluative comments at another. We then benchmark common predictive models on our dataset. We show scope for improvement in building computational predictive models. We believe our dataset will be useful in automatic review comments moderation. In the future, we would like to extend the dataset and investigate the impact of reviewer confidence (Bharti et al., 2022b) on peer-review text moderation.

Acknowledgement

This work has received funding from Cactus Communications, India, under award # CAC-2021-01, awarded to Tirthankar Ghosal. We acknowledge the contributions of our annotators to building the dataset. We thank Rishav Hada for helping with the project. We also thank our anonymous reviewers for their helpful comments on improving this paper.

References

Shima Asaadi, Saif Mohammad, and Svetlana Kiritchenko. 2019. Big BiRD: A large, fine-grained, bigram relatedness dataset for examining semantic composition. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 505–516, Minneapolis, Minnesota. Association for Computational Linguistics.

Prabhat Kumar Bharti, Tirthankar Ghosal, Mayank Agarwal, and Asif Ekbal. 2022a. Betterpr: A dataset for estimating the constructiveness of peer review comments. In *Linking Theory and Practice of Digital Libraries*, pages 500–505, Cham. Springer International Publishing.

Prabhat Kumar Bharti, Tirthankar Ghosal, Mayank Agrawal, and Asif Ekbal. 2022b. How confident was your reviewer? estimating reviewer confidence from peer review texts. In *Document Analysis Systems*, pages 126–139, Cham. Springer International Publishing.

Luke Breitfeller, Emily Ahn, David Jurgens, and Yulia Tsvetkov. 2019. Finding microaggressions in the wild: A case for locating elusive phenomena in social media posts. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1664–1674, Hong Kong, China. Association for Computational Linguistics.

Daniela Calvetti and Lothar Reichel. 2003. Tikhonov regularization of large linear problems. *BIT Numerical Mathematics*, 43(2):263–283.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Cramer Duncan. 1997. *Basic Statistics for Social Research*.

T.N. Flynn and A.A.J. Marley. 2014. Best-worst scaling: theory and methods. In Stephane Hess and Andrew Daly, editors, *Handbook of Choice Modelling*, Chapters, chapter 8, pages 178–201. Edward Elgar Publishing.

Antigoni Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018a. Large scale crowdsourcing and characterization of twitter abusive behavior. In *Twelfth International AAAI Conference on Web and Social Media*.

Antigoni-Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018b. Large scale crowdsourcing and characterization of twitter abusive behavior.

Tirthankar Ghosal. 2022. Studies in aspects of peer review: Novelty, scope, research lineage, review significance, and peer review outcome. *SIGIR Forum*, 55(2).

Tirthankar Ghosal, Sandeep Kumar, Prabhat Kumar Bharti, and Asif Ekbal. 2022. Peer review analyze: A novel benchmark resource for computational analysis of peer reviews. *PLoS ONE*, 17(1):e0259238.

Tirthankar Ghosal, Ashish Raj, Asif Ekbal, Sriparna Saha, and Pushpak Bhattacharyya. 2019a. A deep multimodal investigation to determine the appropriateness of scholarly submissions. In *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pages 227–236.

Tirthankar Ghosal, Kamal Kaushik Varanasi, and Valia Kordoni. 2022. Hedgepeer: A dataset for uncertainty detection in peer reviews. In *Proceedings of the 22nd ACM/IEEE Joint Conference on Digital Libraries*, JCDL '22, New York, NY, USA. Association for Computing Machinery.

Tirthankar Ghosal, Rajeev Verma, Asif Ekbal, and Pushpak Bhattacharyya. 2019b. Deepsentipeer: Harnessing sentiment in review texts to recommend peer review decisions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1120–1130.

Rishav Hada, Sohi Sudhir, Pushkar Mishra, Helen Yannakoudakis, Saif M. Mohammad, and Ekaterina Shutova. 2021. Ruddit: Norms of offensiveness for English Reddit comments. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Online. Association for Computational Linguistics.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.*, 9(8):1735–1780.

Ken Hyland and Feng (Kevin) Jiang. 2020. "this work is antithetical to the spirit of research": An anatomy of harsh peer reviews. *Journal of English for Academic Purposes*, 46:100867.

Dongyeop Kang, Waleed Ammar, Bhavana Dalvi, Madeleine van Zuylen, Sebastian Kohlmeier, Eduard Hovy, and Roy Schwartz. 2018. A dataset of peer reviews (peerread): Collection, insights and nlp applications. *arXiv preprint arXiv:1804.09635*.

Svetlana Kiritchenko and Saif Mohammad. 2017. Best-worst scaling more reliable than rating scales: A case study on sentiment intensity annotation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 465–470, Vancouver, Canada. Association for Computational Linguistics.

Svetlana Kiritchenko and Saif M. Mohammad. 2016. Capturing reliable fine-grained sentiment associations by crowdsourcing and best–worst scaling. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 811–817, San Diego, California. Association for Computational Linguistics.

Sandeep Kumar, Hardik Arora, Tirthankar Ghosal, and Asif Ekbal. 2022. Deepaspeer: Towards an aspect-level sentiment controllable framework for decision prediction from academic peer reviews. In *2022 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pages 1–11.

Sandeep Kumar, Tirthankar Ghosal, Prabhat Kumar Bharti, and Asif Ekbal. 2021. Sharing is caring! joint multitask learning helps aspect-category extraction and sentiment detection in scientific peer reviews. In *2021 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pages 270–273.

Jordan Louviere. 1991. Best-worst scaling: A model for the largest difference judgments. In *working paper*.

Jordan Louviere, T.N. Flynn, and A. A. J. Marley. 2015. *Best-worst scaling: Theory, methods and applications*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

B. Orme. 2009. Maxdiff analysis: Simple counting,individual-level logit, and hb. In *sawtooth software, inc.*

Barbara Plank and Reinard van Dalen. 2019. Citetracked: A longitudinal dataset of peer reviews and citations. In *BIRNDL@SIGIR*.

Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Brij B. Gupta, Xiaojiang Chen, and Xin Wang. 2021. A survey of deep active learning. *ACM Comput. Surv.*, 54(9).

Anna Rogers. 2020a. Peer review in nlp: reject-if-not-sota.

Anna Rogers. 2020b. Peer review in nlp: resource papers.

Anna Rogers and Isabelle Augenstein. 2020. What can we do to improve peer review in NLP? In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1256–1262, Online. Association for Computational Linguistics.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *ACL*.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain. Association for Computational Linguistics.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. 2020. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9275–9293, Online. Association for Computational Linguistics.

Christine - Tardy. 2018. *We are all Reviewer 2: A Window into the secret world of peer review*, pages 271–289. Springer International Publishing.

Ke Wang and Xiaojun Wan. 2018. Sentiment analysis of peer review texts for scholarly papers. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 175–184.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California. Association for Computational Linguistics.

Christie Wilcox. 2019. Rude reviews are pervasive and sometimes harmful, study finds. *Science*, 366(6472):1433–1433.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex machina: Personal attacks seen at scale.

Weizhe Yuan, Pengfei Liu, and Graham Neubig. 2021. Can we automate scientific reviewing?

Mike Zhang and Barbara Plank. 2021. Cartography active learning. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 395–406, Punta Cana, Dominican Republic. Association for Computational Linguistics.