PolitePEER: does peer review hurt? A dataset to gauge politeness intensity in the peer reviews

Article *in* Language Resources and Evaluation · May 2023

DOI: 10.1007/s10579-023-09662-3

Prabhat kr Bharti
Indian Institute of Technology Patna
11 PUBLICATIONS **73** CITATIONS
SEE PROFILE

Mayank Agarwal
Indian Institute of Technology Patna
73 PUBLICATIONS **557** CITATIONS
SEE PROFILE

Asif Ekbal
Indian Institute of Technology Patna
500 PUBLICATIONS **7,941** CITATIONS
SEE PROFILE

Language Resources and Evaluation
https://doi.org/10.1007/s10579-023-09662-3

ORIGINAL PAPER

PolitePEER: does peer review hurt? A dataset to gauge politeness intensity in the peer reviews

Prabhat Kumar Bharti[1] · Meith Navlakha[2] · Mayank Agarwal[1] · Asif Ekbal[1]

Accepted: 27 April 2023

Abstract

Even though peer review is a central aspect of scientific communication, research shows that the process reveals a power imbalance. The position of the reviewer allows them to be harsh and intentionally offensive without being held accountable. It casts doubt on the integrity of the peer-review process and transforms it into an unpleasant and traumatic experience for authors. Accordingly, more effort should be given to provide critical and constructive feedback. Hence, it is necessary to remedy the growing rudeness and lack of professionalism in the review system, by analyzing the tone of review comments and creating a classification on the level of politeness in a review comment. To this end, we develop the first annotated *PolitePEER* dataset encompassing five levels of politeness: (1) highly impolite, (2) impolite, (3) neutral, (4) polite, and (5) highly polite. The review sentences accrued from multiple venues, *viz.*, ICLR, NeurIPS, Publons and ShitMyReviewersSay. We have formulated our annotation guidelines and conducted a thorough analysis of the *PolitePEER* dataset, ensuring the dataset quality with an inter-annotation agreement of 93%. Additionally, we have benchmarked *PolitePEER* for multiclass classification and provided an extensive analysis of the proposed baseline. As a result, the proposed *PolitePEER* can aid in developing a politeness indicator to notify the reviewer and the editors to amend and formalize the review accordingly. Our dataset and codes are available at https://github.com/PrabhatkrBharti/PolitePEER.git for the community to explore further.

Keywords Review comments · Politeness scaling · Dataset applications · Digital libraries

1 Introduction

The most prestigious scientific journals/conferences have widely used peer review to accept various submissions. A peer review process[1] [2] allows the expert domains to assess the author's contributions, highlighting the paper's strengths and weaknesses

[1] https://blog.f1000.com/2020/01/31/a-brief-history-of-peer-review/.

[2] https://mitcommlab.mit.edu/broad/commkit/peer-review-a-historical-perspective/

Extended author information available on the last page of the article

Published online: 14 May 2023

and offering valuable insights and constructive comments (Shema, 2022), thereby encouraging authors to correct their mistakes. Still, on the other hand, the surging submissions (Bornmann & Mutz, 2015) impel the reviewing system to delegate the undertaking to the unseasoned reviewers. Unfortunately, extending the peer-reviewing process to non-expert reviewers has harmed the system's efficacy, attracting unprofessional comments, and offensive remarks on the authors, races, or genders rather than focusing on technical merits (Silbiger & Stubler, 2019; Wilcox, 2019). For instance, the following comments from ShitMyReviewersSay and https://www.discovermagazine.com/mind/years-best-peer-review-comments-papers-that-suck-the-will-to-live.

1. The author's last name sounds Spanish. I didn't read the manuscript because I'm sure it's full of bad English, Sorry guys, I'm throwing in the towel.
2. This paper is desperate. Please reject it completely and then block the author's email ID so they can not use the online system in the future.

Such unprofessional and offensive comments pose a harmful impact on the confidence and enthusiasm of the scientific community, causing psychological distress, especially on the budding authors from disparaged communities (Beaumont, 2019), as discussed by Spencer et al. (2016) and Hyland and Jiang (2020). Besides, in most cases, the reviewers try to address the problems in the papers; moreover, unseasoned reviewers express their suggestions harshly and need more professionalism. Here are some comments from Hyland and Jiang (2020).

1. The biggest problem with this manuscript, which has nearly sucked the will to live out of me, is the terrible writing style.
2. The authors report results from pages 16–26. This section reflects what I would brutally call " death by figures".

Here the reviewers attempted to point out the mistakes like " *poor style of writing*" and " *poor results*" but ended up commenting on it sarcastically.

One of the precursory aspects of an apt peer review is the ability to formally and optimistically underscore all the strengths and mistakes of the research paper. The reviews must be encouraging and organized rather than condescending and demeaning. Hyland and Jiang (2020) in his study,[3] analyzed the harsh peer reviews and stressed the need for an impartial and polite screening technique. Despite the widespread prevalence of unprofessional and impolite reviews in the review process, the scientific community has not considered the effects of politeness on authors.

Politeness and peer review are inherently subjective phenomena. Such politeness tone detection models would play a crucial role in the reviewing system, not only for the editors of the journals to quickly analyze the overall tone of the reviews but also

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

for the novice reviewers by raising an " *impolite flag*" when the reviewer addressed the authors in an impolite way. Thus, notifying the reviewer with such a politeness flag would apprise the reviewer to amend and formalize the review accordingly, thereby abating the tendency to unconsciously write a harsh review. Our current work is a step in that direction.

In this paper, we introduce a novel *PolitePEER* dataset encompassing levels of politeness intensity of peer reviews accrued from multi-disciplinary venues like the international conference on learning representations (ICLR),[4] neural information processing systems (NeurIPS)[5], ShitMyReviewersSay[6] and Publons.[7] Moreover, instead of limiting the politeness of a review to a binary form (i.e., polite/impolite), we extended it to 5 discrete labels based on the intensity *viz.* highly impolite, impolite, neutral, polite, and highly polite, to offer an expansive base for analysis. In this work, we have focused on two significant aspects: an extensive statistical and information-theoretic analysis of the corpus and, second, a set of benchmark baseline models for multiclass classification of politeness on our proposed *PolitePEER*.

Key contributions of our current work are as follows:

1. This study presents a novel annotated politeness dataset (*PolitePEER*) alongside comprehensive annotation guidelines for peer-reviewed sentences.
2. We create and offer a corpus of 2500 labeled review sentences (annotated on the proposed politeness guidelines). In addition, we have 70k unlabelled review sentences for future experiments. We additionally benchmark multiclass classification models on the proposed *PolitePEER*.

The following section discusses the related work. Section III describes data procurement and description. In Section IV, we provide a brief description of the experiments that we conducted. Section V is dedicated to presenting the results and discussing their implications. Additionally, we conduct an error analysis to assess the model's performance. Finally, in Section VI, we conclude the paper and highlight the possible areas for future contributions.

2 Related work

Peer review aims to improve the scientific record by pointing out weaknesses, providing feedback, and identifying misleading findings. Unfortunately, peer review has come under fire for efficiency, bias, and fairness (Bohannon, 2013; Schwartz & Zamboanga, 2009; Silbiger & Stubler, 2019). Even though many scientists advocate peer review (Beaumont, 2019; Jefferson et al., 2006; Mulligan et al., 2013), there is little empirical evidence to support its effectiveness. There is often a lack of proper

[4] https://iclr.cc/.

[5] https://neurips.cc.

[6] http://shitmyreviewerssay.tumblr.com.

[7] https://publons.com/wos-op/.

professional etiquette in peer-reviewer comments, demeaning authors or concentrating on the author's gender, race, ethnicity, and nation of origin rather than the technical merit of the submission (Beaumont, 2019; Hyland & Jiang, 2020; Silbiger & Stubler, 2019). Psychological distress has been associated with unprofessional comments among early-career scientists (Bonn, 2020; Hyland & Jiang, 2020).

Review articles are now far less of an obscure genre (Swales, 1996) than they were 20 years ago, and we can better predict what we might find in them. A majority of reviewers, for example, concentrate on the content and language of submissions (Coniam, 2012; Duenas, 2012; Mungra & Webber, 2010). In contrast, criticisms of style or language are not uncommon but are rarely decisive factors in rejection (Belcher, 2007). A revision recommendation is typically significantly longer than a rejection or acceptance recommendation (Coniam, 2012; Hewings, 2004). Editors value longer reviews because they provide additional commentary on the goals, analyses, and claims made in the review (Falkenberg & Soranno, 2018).

A significant part of a review is criticism, which may constitute half of all comments (Fortanet, 2008). Some of the comments can be rather blunt (Kourilova, 1996), perhaps because the reviewers were anonymous, the report was hurried, had a personal style, or had no practical experience. Reviewers from English as an Additional Language (EAL) communities are increasingly invited to participate in the review process (Hyland, 2016), and their reviews are more direct than those of native English speakers (Paltridge, 2017). Study (Hyland, 2005) of 150 peer reviews in leading applied linguistics journals found that direct criticism and suggestions were rare. However, the studies in Belcher (2007); Hyland (2018) found considerable politeness.

Rather than imposing their criticism directly, the reviewers subordinated criticism to praise by following negative statements with positive remarks and minimizing the overall criticism's severity by using hedges and asking questions rather than imposing them. In addition, analyses of review articles have indicated that reviewers frequently use attitude markers and self-mentions to enhance their sense of authority and confidence (Paltridge, 2017). Finally, reviewers generally strive to establish a positive, sympathetic relationship with authors.

2.1 A computational approach to politeness with application to peer review

Politeness is a universal aspect of communication (Brown et al., 1987; DanescuNiculescuMizil et al., 2013; Voigt et al., 2017) and central to modern pragmatic theory since it provides pragmatic enrichment, social meaning, and cultural variation (Brown & Levinson, 1978; Grice, 1975; Lakoff, 1973, 1977; Leech, 2016). Speakers can be more or less polite to their audiences in virtually any setting. Most research starts with Brown et al. (1987) theory. In recent years, politeness has been studied in online settings. According to researchers, politeness marking differs according to context, media type, and social groups (Brennan & Ohaeri, 1999; Burke & Kraut, 2008; Duthler, 2006; Herring, 1994). It's all about how language relates to power and status, which is why politeness marking is one part of the larger discussion (Gilbert, 2012; Peterson et al., 2011; Prabhakaran et al., 2012; Scholand et al., 2010). The current study, however, looks to leverage domain-independent politeness cues, whereas this work focuses on domain-specific textual cues and builds on the literature on how politeness affects workplace power structures and social dynamics (Andersson & Pearson, 1999; Holmes, 2005; Obeng, 1997; Rogers & Lee-Wong, 2003). According to DanescuNiculescuMizil et al. (2012), aspects of linguistic accommodation reveal power differences on Wikipedia. This paper complements this previous research by showing how politeness affects peer review.

This motivation led us to detect the politeness in the review text. Here, we proposed the PolitePEER dataset, which classifies each sentence in a peer review as highly impolite, impolite, neutral, polite, or highly polite based on various factors, including tone, attitude, formal writing etiquette, and overall articulation.

2.2 Currently available data on peer reviews

Over the past few years, there has been a significant increase in the availability of peer review datasets for various research purposes. Kang et al. (2018) introduced PeerRead, which is the first large-scale peer review dataset. It contains full-text papers from conferences like ACL, ICLR, and NIPS, along with their editorial accept/reject decisions and review comments. Gao et al. (2019) created the ACL-18 Numerical Peer Review Dataset, which includes review comments, author rebuttals, and scores given by reviewers before and after the rebuttal phase. Hua et al. (2019) developed AMPERE, a dataset that focuses on reviewers' argumentative propositions and their types. It contains 10,386 labeled propositions from 400 review comments. Plank and Dalen (2019) constructed CiteTracked, which is a dataset for citation count prediction based on papers from NeurIPS, their review comments, and the number of citations. Stappen et al. (2020) built the Interspeech 2019 Submission dataset, which contains the original text of all papers submitted to Interspeech, including their corresponding review comments, scores, and decisions, with a large number of rejected papers. Yuan et al. (2022) created ASAP-Review, which is the largest dataset of review comments in the field of computer science. It contains sentence-level aspect labels for 1,000 review comments. Singh et al. (2021) constructed COMPARE, a dataset of comparative discussion sentences collected from review comments, aimed at achieving more efficient peer review. Matsui et al. (2021) analyzed the peer review process based on data collected from PeerJ, studying the impact of peer review on papers in biology, environment, and computer science using traditional machine learning technologies. Choudhary et al. (2021) introduced ReAct, a review comment dataset with classified review comments used to understand the requirements of the reviewers to the authors. Ghosal et al. (2022) built the first multi-layered peer review dataset with 17k sentences from 1199 review comments, labeled across four layers to study the corresponding sections of the paper, the viewpoint of the review text, the role of the review text, and the importance of the review statement within the review. Finally, Shen et al. (2022) constructed MReD, which contains a total of 45,929 sentences from 7089 labeled meta-reviews. This dataset was used to propose structurally controllable extractive and abstractive text summarization models.

Table 1 Currently available data on peer reviews

Name	Domain	Description
PeerRead (Kang et al., 2018)	AI	Acceptance prediction, score prediction
ACL-18 numerical (Gao et al., 2019)	NLP	Prediction of score changing after rebuttal
AMPERE (Hua et al., 2019)	AI	Analysis of review arguments
Interspeech 2019 submission (Stappen et al., 2020)	Speech	Acceptance prediction, score prediction
CiteTracked (Plank & Dalen, 2019)	AI	CiteTracked dataset for citation count prediction based on papers from NeurIPS, their review comments, and the number of citations
COMPARE (Singh et al., 2021)	AI	Discussion sentence comparison
Dataset of (Matsui et al., 2021)	BIO, CS, ENV	Analysis of peer review process
ReAct (Choudhary et al., 2021)	AI	Requirement analysis in review comments
MOPRD (Lin et al., 2022)	AI	A multidisciplinary open peer review dataset
Peer review analyze (Ghosal et al., 2022)	AI	Analysis of review comments
MRed (Shen et al., 2022)	AI	Meta-review generation
ASAP-review (Yuan et al., 2022)	AI	Review comment generation
BetterPR (Bharti et al., 2022a)	AI	Introduced a dataset focused on Binary classification of review comments as either constructive or not

Table 1 illustrates all of the peer review datasets mentioned above. However, the above research focuses on something other than a dataset of this kind. In this paper, we define for the first time the problem of the reviewer's tone. We propose a novel *PolitePEER* dataset encompassing five fine-grained politeness degrees, and provide actionable NLP baselines.

3 Data procurement and description

The sentences used for the *PolitePEER* dataset were collected from various sources such as ICLR, NeurIPS, Publons and ShitMyReviewersSay. These sources cover a broad range of subject areas including machine learning, natural language processing, artificial intelligence, environment and ecology, clinical medicine, Psychiatry and Psychology, and engineering. Our objective was to provide a diverse set of review styles and subject areas.

Table 2 displays the distribution of review sentences across different venues, revealing a higher ratio of neutral class sentences. While this may be seen as a limitation, we acknowledge that this is a reflection of the nature of the domain we are working in. Most reviewers tend to follow a neutral trend while expressing their views, and finding samples for highly impolite and highly polite sentences is an arduous task. Despite this limitation, we believe that our *PolitePEER* dataset is

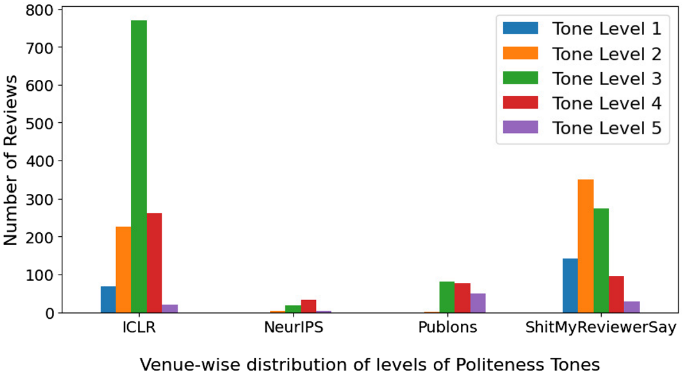
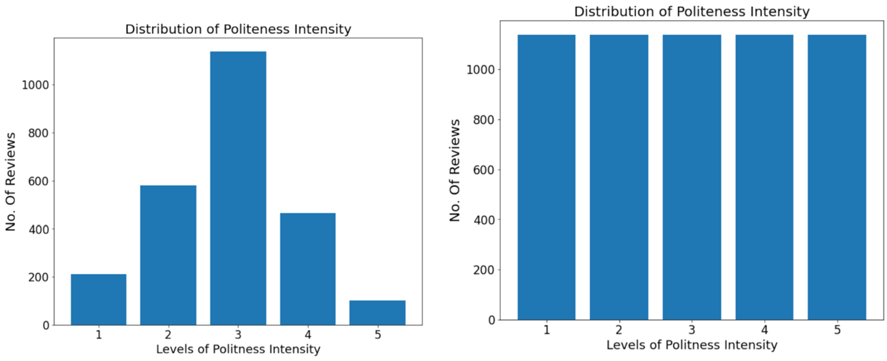


Fig. 1 Vanue-wase data distribution plot concerning politeness intensity. Here Tone Lavel (1, 2, 3, 4, and 5) indicates (highly impolite, impolite, neutral, polite, and highly polite)



(a) Data distribution plot before upsampling

(b) Data distribution plot after upsampling

Fig. 2 Data distribution plot with respect to politeness intensity (highly impolite, impolite, neutral, polite, and highly polite) before and after upsampling

valuable for research purposes as it provides a unique resource in the field, covering a diverse range of reviews from various sources.

To ensure the *PolitePEER* dataset's diversity, we scraped and annotated review sentences from ICLR, NeurIPS, ShitMyReviewersSay, and Publons websites, resulting in 1348, 58, and 888 annotated review sentences, respectively. Additionally, we handpicked and annotated 206 review sentences from the dataset shared by Publons. Each sentence in the dataset is associated with a review ID, enabling researchers to map it to the original research paper and conference in which it was published. Figure 1 displays the data distribution plot concerning politeness intensity for each venue.

Figure 2a depicts a significant imbalance in the number of instances across different politeness labels in the *PolitePEER* dataset. To address this issue, we upsampled

Table 2 Statistics and distribution of annotation labels across the proposed *PolitePEER* dataset

Venues	ICLR, NeurIPS, ShitMyReviewersSay and Publons	
Number of sentences	2500	
Average length in terms of words	21	
Venue wise distribution	ICLR	1348
	NeurIPS	58
	Publons	206
	ShitMyReviewersSay	888
Politeness intensity-wise distribution	Highly impolite	210
	Impolite	582
	Neutral	1140
	Polite	465
	Highly polite	103

the extreme politeness labels, resulting in multiple copies of the same sentences. This approach helps in balancing the dataset and improving the model's performance on the minority classes. Figure 2 shows the data distribution plot before and after upsampling for each politeness category, highlighting the impact of upsampling on balancing the dataset. We hope that our *PolitePEER* dataset will enable researchers to gain deeper insights into the language used in peer reviews and lead to the development of more accurate models for predicting politeness in this domain.

3.1 Annotation labels and guidelines

In our study, we classified each sentence in the review corpus into one of five politeness classes: (1) highly impolite, (2) impolite, (3) neutral, (4) polite, and (5) highly polite based on various factors such as the tone, attitude of expressing opinions, writing etiquettes, and overall articulation of the review. To ensure consistency in our annotations, we conducted a preliminary study on Lauscher et al. (2018) and formulated our annotation guidelines based on ideas extrapolated from several related studies, including Hyland and Jiang (2020), Ghosal et al. (2022), Bharti et al. (2022a), and Verma et al. (2022). During the annotation process, we encountered discrepancies in determining the politeness level for certain review sentences. To address this, we employed an approach that involved expert guidance, iterative annotation, providing examples, and feedback. This approach helped to reduce inter-annotator disagreements and ensured that the final annotation was based on a majority approach. We also ensured that the annotations were performed by trained annotators with expertise in the domain of peer review. The resulting dataset with five politeness classes will allow for more nuanced analyses of the language used in peer review and aid in the development of models for automatic politeness classification. Table 2 shows the number of reviews for each politeness level, demonstrating the distribution of the dataset across the different classes. Our annotation guidelines are defined as follows:

1. Highly Impolite: In highly impolite review sentences, the reviewer hurls abuses or uses explicit language like *"noobs"* or *"a bunch of rookies"* at the authors or personal assessment of their character or intellect. Also, it includes racial or gender discrimination written in a demeaning way, as seen in, E.g., 1 from ShitMyReviewersSay. The reviewers might mention how the paper negatively affected them, as seen in, E.g., 2 from ShitMyReviewersSay.

   (a) E.g. 1This paper reads like a woman's diary, not like a scientific piece of work.
   (b) E.g. 2 Sorry guys, I'm throwing in the towel.

   In, E.g. 1, the reviewer has written a discriminatory sentence against women by comparing the substandard paper to a woman's diary. Such a statement is directly demeaning to women and is not at all related to the paper. Similarly, in, E.g. 2, the reviewer rudely expressed his distaste for the paper rather than formally writing his opinion.
2. Impolite: In impolite review sentences, the reviewer uses a sarcastic or mocking tone with a negative connotation. The review generally has informal writing. Refers to the authors in the second-person, like *"You"* or *"you'll"* or contains self-mention as seen in, E.g. 3 from ShitMyReviewersSay. The reviewer blatantly criticizes the authors for mistakes, writing styles, or other aspects of the paper. Hedges and discern having no support and grounded reasons for rejecting the papers as observed in, E.g., 4 from ShitMyReviewersSay.

   (a) E.g. 3 Have you no command of the English language?.
   (b) E.g. 4 The writing is often arrestingly pedestrian.

   In, E.g. 3, the reviewer informs the authors in the second person, " you". In addition, the reviewer has informally criticized the authors for weak English rather than professionally offering suggestions to rectify the mistakes. Similarly, in, E.g., 4, the author impolitely criticizes the writing style.
3. Neutral: In neutral review sentences, the review is bland and does not lean to either side of the scale. It is generally related to work, as seen in, E.g., 5 from ICLR. Proper writing etiquettes are followed without unprofessional expressions or comments, as seen in, E.g., 6 from ICLR. The reviewer mainly focuses on the works and technical aspects of the research papers. The overall tone of the review sentence could be more apparent on which side of politeness it belongs.

   (a) E.g. 5 The matrices of unknown variances are considered as parameters and are learned with a standard gradient descent.
   (b) E.g. 6 This paper proposes a new object counting module which operates on a graph of object proposals.

   In E.g., 5, the reviewer has summarised what he has understood about the key contribution in the paper and has maintained it in a neutral tone. The review mentions that the presentation of the paper is difficult, but it does not insult the author. Similarly, in E.g., 6, the reviewer summarises the author's proposition in the paper without making derogatory or praising comments.

4. Polite: In polite review sentences, the reviewer uses semi-formal writing. The reviews contain moderate praise (slightly positive), as seen in, E.g., 7 from ICLR, or criticism but only regarding the paper in a semi-professional manner, as seen in, E.g., 8 from ICLR. No harsh language is used.

   (a) E.g. 7 Overall, the proposed approach is novel and achieves good results on a range of tasks.
   (b) E.g. 8 Results are not always too impressive, but authors seem intent on making them useful for pathogists in practice (an intention that is always worth the effort).

   In, E.g., 7, the reviewer maintains a positive tone and praises the author for the approach and results. Similarly, in, E.g., 8, the reviewer still needs to be satisfied with the results, yet the reviewer still praises the authors for their efforts and motivates them to research further.
5. Highly Polite: In highly polite review sentences, the reviewer uses formal writing. One of the distinguishing features of highly polite is the presence of courtesy markers like *"please," "sorry,"* and *"kindly"* as seen in, E.g., 9 from ICLR. The sentences have an overall positive tone (praising) or criticizing but only regarding the paper in a proper professional manner, as observed in, E.g., 10 from ICLR.

   (a) E.g. 9 Another issue is the fact that, on my humble opinion, the main text looks like a long proof.
   (b) E.g. 10 In the interest of being helpful, my suggestion is that the authors go back and review what is involved in the scientific method. An essential step in the scientific method is posing hypotheses and/or asking questions, and that step is completely lacking in this study.

   In, E.g., 9, the reviewer points out an issue in the paper that the main text looks like a *"long proof"* but writes it in a very generous way *("my humble opinion")*. Similarly, in, E.g., 10, the reviewer offers an insightful comment and courteously advises the authors to work on them.

3.1.1 Data quality

Four annotators are assigned to the task. One of the annotators (also an author) is a Ph.D. student familiar with NLP/ML paper discourse and review structure. The other three annotators are hired duly paid according to the annotation payment standards in India. Of these three, two hold graduate degrees in Linguistic and English Literature, and one holds a Bachelor's degree in computer science and engineering (CSE). We asked each annotator to read (Hyland & Jiang, 2020) paper to understand the critical stance dimension. Each primary investigator participates in team meetings to resolve confusing cases.

Inter-Annotator Agreement (IAA): we assessed the inter-annotator agreement (IAA) of a subset of our dataset consisting of 1500 review comments. We found substantial agreement among the annotators, with IAA values ranging from 90.74% to 93.17% using Cohen's Kappa, Kendall's Tau, and Krippendorff's Alpha measures. Encouraged by these results, we increased the dataset size to include a more diverse set of cases, including edge cases, syntactic and semantic ambiguities, and grammatical inconsistencies. Despite these challenges, we confidently conclude that the IAA values obtained from the 1500 review comments are representative of the larger dataset. To validate this claim, we measured the IAA of the expanded dataset using (Cohen, 1960), Kendall and Smith (1939), and Krippendorff (2004) Alpha measures. We found that all three measures returned high agreement scores, with Cohen's Kappa and Fleiss' Kappa both returning a score of 90.70% and Krippendorff's Alpha returning a score of 93.17%. These results provide strong evidence that the high IAA values observed in the smaller dataset are robust and generalizable to a larger and more diverse set of review comments.

4 Experiments

Our *PolitePEER* dataset comprises 5 discrete levels of politeness tone, enabling us to perform multiclass classification on the review sentences to identify the nature of the reviews based on the *PolitePEER* taxonomy. When a review sentence is fed to the model, it returns the politeness score in the range of 1 (highly impolite) to 5 (highly polite).

4.1 Baselines models

Most of the recent works on linguistic classification leverage state-of-the-art transformers for extracting the embedding vectors. We also experimented with the avant-garde sentence embedding transformers for better performance. The baseline model consists of 3 main sections: Input Layer, Embedding Layer, and Output Layer as shown in Fig. 3.

4.1.1 Input layer

The input to our model is review sentences. Let us denote the review $R = \left(s_1, s_2, \dots, s_N\right)$. Once a peer review is fed to the input layer, it undergoes a series of data-cleaning and pre-processing processes. We first remove all irrelevant characters (like white spaces, newline characters, underscores) and then feed sentences to the

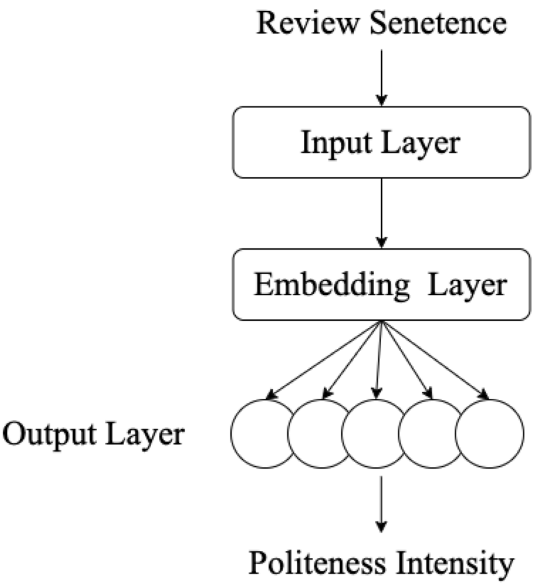


Fig. 3 Flowchart of the proposed baseline models

embedding layer as shown in Fig. 3. The role of the input layer is to take in the input and make it ready for the embedding layer.

4.1.2 Embedding layer

The following variations of the state-of-the-art sentence embeddings are used:

1. SciBERT: SciBERT (Beltagy et al., 2019) is trained on scientific texts. The variant uses SciBERT pre-trained model from Hugging Face[8] in the embedding layer which encodes the incoming processed input into a 768-dimensional embedding vector.
2. HateBERT: Hate-BERT proposed by Caselli et al. (2020), is built on BERT transformer and fine-tuned on RAL-E, a dataset of abusive, offensive and hateful language in comments on the Reddit platform. The variant uses HateBERT pre-trained model from Hugging Face[9] in embedding layer which encodes the incoming processed input into a 768-dimensional embedding vector.
3. ToxicBERT: ToxicBERT proposed by Luu and Nguyen (2021), is a further adaptation of Hate-BERT transformer, focusing on types of toxicity like threats, obscenity, insults, and identity-based hate in the reviewer's comments. The variant uses the ToxicBERT pre-trained model from Hugging Face[10] in embedding layer which encodes the incoming processed input into a 768-dimensional embedding vector.

[8] https://huggingface.co/gsarti/scibert-nli.

[9] https://huggingface.co/GroNLP/hateBERT.

[10] https://huggingface.co/unitary/toxic-bert.

Table 3 The hyperparameter details used for all baseline models

Embedding dimension (For SciBert, Hate-BERT and Toxic-BERT)	768
Batch size	32
Epochs	30
Loss function	Categorical crossentropy
Activation function	Softmax
Optimizer	Adam
Learning rate	0.001

4. Custom Embedding: In the custom embedding layer, we first feed the tokenized input to the word2vec embedding layer and pass the resulting 300-dimensional embedding vector to Bidirectional LSTM. Bidirectional LSTM efficiently captures the data flow in both the directions i.e. forward and reverse flows, thereby enhancing its context preserving ability, even for the long sentences. After experimenting and hyperparameter tuning, we consider 256 units in each forward and backward LSTM flow for the Bi-LSTM layer.

4.2 Output layer

The 768-embedding vector is then passed to the output layer. It consists of 5 neurons, each with a politeness intensity label. We use the softmax activation function in the output layer to normalize the output values, such that the sum of the probabilities of all the politeness labels is 1. This is beneficial for multi-class classification. Finally, the output layer returns an array of probabilities of all 5 politeness tone intensities, and the label with the highest probability is selected.

4.3 Experimental setup

All the experiments were performed on the *PolitePEER* using 80, 15, and 5 train, test, and validation split respectively. In order to ensure equal representation of all the labels in all the 3 splits we used stratified splitting of the corpus. The experiment was performed using Python and Keras package with Tensorflow-GPU on Kaggle having 16 GB RAM, along with Tesla P100 16GB VRAM GPU. *For fair comparison we have kept the hyperparameter values the same for all the baseline models. Batch Size: 32, epochs: 30, Learning Rate: 0.01.* The hyper-parameters used in the learning process for all pre-trained models are shown in Table 3.

Table 4 Politeness tone class-wise testing accuracy and F1 score of all the baseline models for politeness multiclass classification on our *PolitePEER*

Tasks		Competitive baselines							
		Custom embedding (word2vec + BiLSTM)		SciBERT		Hate-BERT		Toxic-BERT	
		ACC	F1	ACC	F1	ACC	F1	ACC	F1
Politeness	Highly impolite	0.961	0.92	0.353	0.49	0.759	0.69	0.843	0.77
	Impolite	0.804	0.78	0.373	0.43	0.765	0.58	0.588	0.61
	Neutral	0.746	0.81	0.559	0.54	0.441	0.53	0.458	0.53
	Polite	0.902	0.89	0.508	0.63	0.792	0.51	0.639	0.61
	Highly polite	1	0.98	0.489	0.38	0.887	0.77	0.919	0.87

5 Results and discussion

To better understand the adequacy of our proposed taxonomy, we trained the presented baseline models on our *PolitePEER*. Table 4 summarises the experimental results of all 4 baseline models for multiclass classification of politeness tone intensity. We trained all the models under the same experimental conditions for a fair comparison. Evaluation of the model's performance for all the 5 politeness classes were made against (1) accuracy and (2) F1 score and presented in Table 4. We noted that the accuracy of all the pre-trained based embedding models *viz.* SciBERT, Hate-BERT, and ToxicBERTbaselines yielded poor results compared to our custom embedding model (word2vec and BiLSTM). One of the primary reasons for the poor performance of the former three models was that our *PolitePEER* consisted of only 2500 review sentences.

As a result, our custom embedding layer overfitted (memorized over learning the features) on the limited corpus and, thus, could transform the given testing review sentence into embedding vectors without much loss of information, compared to the other three pre-trained transformer variants, which were generalized over large sentences and tokens. Therefore, we witnessed such high accuracy scores for the custom embedding baseline. Moreover, owing to the significant imbalance in our *PolitePEER*, as observed earlier in Fig. 2, we had to upsample the extreme politeness labels viz, highly impolite (210 review sentences) and highly polite (103 review sentences), resulting in multiple copies of the same sentences, thus overfitting the training dataset, especially for these classes. As a result of this, we observe eccentric results for these 2 classes, i.e., either too high accuracy scores like 96.1% (highly impolite) and 100% (high polite) for custom embeddings, 91.9% (high polite) for ToxicBERTand 88.7% (high polite) for Hate-BERT or very low accuracy of 35.3% (highly impolite) for SciBERT.

SciBERT had the worst performance, as it was trained to capture the linguistic features of scientific jargon and not hate or politeness-based features. Thus, we observe that SciBERT outperforms both Hate-BERT and ToxicBERTin neutral

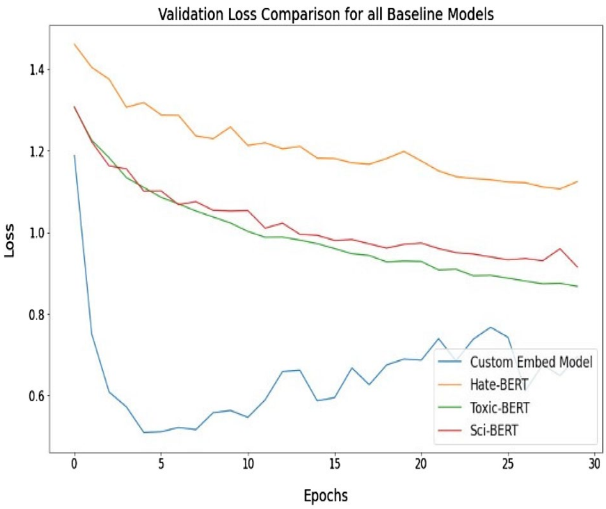


Fig. 4 Illustrates the comparison of validation loss of the 4 baseline models on our *PolitePEER* for multiclass classification

review sentences. Hate-BERT and ToxicBERTproduced decent results in other cases. ToxicBERToutperformed Hate-BERT at the extreme ends, i.e., highly impolite and highly polite, as ToxicBERTwas fine-tuned on dataset specifically focusing on threats, obscenity, insults, and identity-based scathing (e.g., targeting the authors) in the reviewer's comments. These were the basis for the highly impolite level and antithesis to its obverse highly polite. On the contrary, Hate-BERT marginally outperformed ToxicBERTfor the polite and impolite politeness levels as it was fine-tuned on hateful comments and inappropriate addressing or writing in the comments sharing resemblance to the key identifiers of the impolite level and complemented the polite level. Our custom embedding yielded exceptionally high accuracy values merely because it memorized the training data over learning its features.

5.1 Error analysis (misclassified statistics)

We critically examined all the baseline models' performance over varied testing samples and enlisted a few examples where the models efficiently captured the context and correctly predicted the politeness label in Table 5 and cases where we observed misclassification from some or all the baselines in Table 5. We noticed a general trend across all the baseline models, that the models tend to produce higher politeness scores (i.e., polite (4) or highly polite (5)) for long sentences > 30 words, irrespective of the content present in the sentences. It was mainly due to the fact that we just used the existing pre-trained models in the embedding layer and directly passed the embedding vector to the output layer, comprising of 5 neurons, one for each politeness tone level, without adding any context-preserving layers like LSTM or BiLSTM. Since the long review sentences > 30 words would naturally have at least some technical and paper-related phrases/words that overshadow the overall

Table 5 Few examples where the models correctly predicted the politeness classes. Here (1, 2, 3, 4, and 5) indicates (highly impolite, impolite, neutral, polite, and highly polite)

Sr. no.	Sentences	Actual label	Predicted by baseline model			
			Custom embed	ToxicBERT	HateBERT	SciBERT
1	This is (I'm sorry) utter nonsense	1	1	1	1	1
2	The image presentation is a minor extension based on a method of producing permutation invariant adjacency matrix	3	3	3	3	3
3	Sorry for our long silence, due to some perplexity on our side at reading your manuscript	5	5	5	5	5
4	You have put in a lot of effort answering a question that should have never been asked	2	2	2	2	2
5	The study is poorly conceived and inadequately conducted and the conclusions made by the authors do not necessarily follow from the results	4	4	4	4	4

context (even harsh), yielding higher politeness levels. On the contrary, we observed higher performance from our custom embedding model because it incorporated an intermediary BiLSTM layer. However, our custom embedding layer was just trained on 2500 review sentences, and the embedding matrix memorized those reviews rather than learning to extract features from them.

Therefore, this model would not generalize well with the unknown review sentences having new words unprecedented to the model. Validation loss of all the posited baselines for multiclass classification has been illustrated in Fig. 4.

Table 5 enumerates review sentences that are correctly classified by all the baseline models. For example, " *This is (I'm sorry) utter nonsense.*" Review sentence 1 in Table 5, all the models adeptly capture the strong derogatory word "*utter nonsense*" and classify it as highly impolite (1). Similarly, in " *Sorry for our long silence, due to some perplexity on our side at reading your manuscript*", review sentence 3 in Table 5, the model efficiently captures " *Sorry*" a strong politeness indicator, and the formal structure of the review sentence and correctly classifies the review sentence as highly polite (5). Moreover, in " *The image presentation is a minor extension based on a method of producing permutation invariant adjacency matrix.*" review sentence 2 in Table 5, the review sentence neither has any specific politeness indicators *(e.g., please, thank you, sorry)* nor extreme hateful and demeaning words, it just summarizes the work done in the paper, and so all the models classify it as neutral (3).

Table 6 iterates review sentences that are misclassified by at least one or all the baseline models. In " *Please, also perform spell checking and proof-reading. Too much of typos!*" review sentence 5 in Table 6, the presence of the politeness indicator "*Please*" and semi-formal writing style make it a polite (4) sentence. The custom embedding model might have focused on the latter segment, "*Too much of typos!*" and classified it as highly impolite (1). The absence of strong insulting or objectionable words confounds the ToxicBERT into classifying it as highly polite (5). Along similar lines, the absence of hateful comments in the review sentence leads Hate-BERT to misclassify it as neutral (3).

SciBERT, too, focuses mainly on technical details like "*spell checking*" and "*proof-reading*" and then the expression "*too much of typos!*" finally indicating a slightly impolite sentence; thus, SciBERT misclassifies it as impolite (2). In " *It is clear that the author has read way too much and understood way too little. Many of the most serious errors are more or less just copied out from what he has read, so it is hard to know how to deal with such cases. Sprinkled here and there are some things, also taken from his readings, that are more or less correct. But it is all very confused.*" review sentence 1 in Table 6, the sentence targets the authors *(mentions " he")* but in an indirect way. Even though the sentence does not contain any abusive or abasing words, it is still demeaning and informally written. Due to the absence of overt toxicity, ToxicBERTfails to capture the abhorrence in the sentence and thus misclassifies it as highly polite (5). Hate-BERT being trained on hateful comments from Reddit can identify the indirect insinuation; however, being a lengthy sentence, it dilutes the acerbity with other contextual words and misclassifies it as impolite (2). Being a long sentence, SciBERT and custom embeddings cannot preserve the overall context and misclassify it as polite (4) and highly impolite (5), respectively.

Table 6 Few examples where the models incorrectly predicted the politeness classes

Sr.No	Sentences	Actual label	Predicted by baseline model			
			Custom embed	ToxicBERT	HateBERT	SciBERT
1	It is clear that the author has read way too much and understood way too little. Many of the most serious errors are more or less just copied out from what he has read, so it is hard to know how to deal with such cases. Sprinkled here and there are some things, also taken from his readings, which are more or less correct. But it is all very confused	1	5	5	2	4
2	I think the authors have given adequate replies to all comments from the reviewers and the editor and made relevant changes to the manuscript	3	5	4	4	5
3	I have read this MS twice, which given the grammatical howlers in the Abstract would appear to be more times than it has been read by the authors. Given the lack of proof-reading and my concern over the methods used, I shall not comment beyond the end of the methods section, and shall comment selectively rather than exhaustively (which would indeed be exhausting)	2	5	5	4	2
4	This is all science done by wishful thinking	1	1	1	1	5
5	Please, also perform spell checking and proofreading. Too much of typos!	4	1	5	3	2

On the contrary, “ *This is all science done by wishful thinking.*” review sentence 4 in Table 6, all the models expect SciBERT proficiently capture the sarcastically mentioned “*wishful*” term in the review sentence and label it as highly impolite (1). Nonetheless, SciBERT, being fine-tuned to focus on scientific jargon, might have literally interpreted the sentence and thus classified it as polite (4). Therefore, our baselines can adeptly capture insinuation and sarcasm in short sentences but might fail in the case of long sentences, as observed in sentences 1 and 4 in Table 6.

To summarise, the custom embedding model outperforms all the other models for multiclass classification on the *PolitePEER*. However, it needs to be generalized with a larger dataset. Also, we observe an increase in the loss after 5 epochs indicating overfitting of the model as seen in Fig. 4. Toxic-BERT, Sci-BERT, and Hate-BERT too present decreasing loss values, as depicted in Fig. 4. In addition, we observed a common trend in all the baseline models. The models tend to predict higher politeness labels for long sentences (> 30 words), mainly because the embedding layer’s ability to preserve context fails for long sentences. Moreover, we have not used any context preservation layers (like the LSTM or BiLSTM layers) or attention layers in the pre-trained based baselines, which is another reason for poor performance for long sentences.

Our custom embedding model was chosen based on its strong performance on the specific dataset and task at hand, despite its ad hoc appearance. While we acknowledge the issue of overfitting in our model, our primary focus was on the creation of the *PolitePEER* dataset, rather than model optimization. The simple baseline model was intentionally chosen to provide a starting point for future research and facilitate comparisons. Our study aimed to contribute a valuable resource to improve the peer review process and inspire further research in the field.

6 Conclusion

Research is disseminated through peer review. However, it exhibits a power imbalance with review comments that are overly critical and can often cross the line into disparaging while also demonstrating poor review practices. It is the responsibility of the senior area chair or editor to moderate these review comments, especially for young researchers.

In this paper, we introduce a novel multidisciplinary *PolitePEER* dataset that consists of five levels of politeness: highly impolite, impolite, neutral, polite, and highly polite. Review sentences are collected from various venues, including ICLR, NeurIPS, ShitMyReviewersSay, and Publons. To ensure the quality of the annotated dataset, we formulated annotation guidelines and analyzed the dataset thoroughly, ensuring 93% inter-annotation agreement.

The *PolitePEER* dataset offers a valuable contribution towards improving the quality of peer review reports by providing annotations for five levels of politeness. While the number of impolite reviews may not be significant enough to warrant an automatic check, the impact of such reviews on the individuals involved in the review process cannot be overlooked. Impolite reviews can be demotivating and negatively impact the review process. Moreover, the automated reviewer assistant tool, which flags potentially impolite language, serves as a reminder for reviewers to maintain a respectful tone. This work ultimately aims to promote constructive communication in the review process and improve research outcomes.

A dataset of this kind to predict the politeness of review sentences would allow (senior) area chairs /editors to prevent such comments from reaching the public or authors. Additionally, an automated reviewer assistant tool could use such a predictor to flag or alert reviewers when they write harsh comments (or are repeat offenders). Of course, politeness and peer review are inherently subjective phenomena. However, we must strive to make the peer-review process more welcoming to maintain the objectivity of the fundamental process of scrutinizing science. In the future, we would like to broaden the scope of the dataset and study how reviewer confidence Bharti et al. (2022b) affects the moderating of peer-review text submissions.

Acknowledgements The third author, Asif Ekbal, has received the Visvesvaraya Young Faculty Award. He owes a debt of gratitude to the Indian government and the Ministry of Electronics and Information Technology for their assistance.

Author contributions PKB: Conceptualization, data curation, investigation, methodology, experiments, writing - original draft and review & editing. MN: Data curation, data cleaning. MA: Supervision, reviewing & editing. AE: Supervision, reviewing & editing.

Declarations

Conflict of interest It is declared that none of the authors have any conflicts of interest concerning the publication of this article.

Ethical approval We do not intend to attack specific individuals. Our purpose is to draw attention to the negative cultural zeitgeist in peer review, hoping that a discussion will stimulate improvement.

References

Andersson, L. M., & Pearson, C. M. (1999). Tit for tat? the spiraling effect of incivility in the workplace. *Academy of Management Review,* 24(3), 452–471.

Beaumont, L. J. (2019). Peer reviewers need a code of conduct too. *Nature,* 572(7769), 439–440.

Belcher, D. D. (2007). Seeking acceptance in an english-only research world. *Journal of Second Language Writing,* 16(1), 1–22.

Beltagy, I., Lo, K., Cohan, A. (2019). Scibert: A pretrained language model for scientific text. Preprint retrieved from http://arxiv.org/abs/1903.10676

Bharti, P.K., Ghosal, T., Agarwal, M., & Ekbal, A. (2022a). Adataset for estimating the constructiveness of peer review comments. In*International conference on theory and practice of digital libraries* (pp. 500–505). Springer.

Bharti, P.K., Ghosal, T., Agrawal, M., & Ekbal, A. (2022b). How confident was your reviewer? Estimating reviewer confidence from peer review texts. In *International workshop on document analysis systems* (pp. 126–139). Springer.

Bohannon, J. (2013). Who's afraid of peer review? American Association for the Advancement of Science

Bonn, N.A. (2020). Noémie aubert bonn

Bornmann, L., & Mutz, R. (2015). Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. *Journal of the Association for Information Science and Technology,* 66(11), 2215–2222.

Brennan, S. E., & Ohaeri, J. O. (1999). Why do electronic conversations seem less polite? the costs and benefits of hedging. *ACM SIGSOFT Software Engineering Notes,* 24(2), 227–235.

Brown, P., & Levinson, S.C. (1978). Universals in language usage: Politeness phenomena. In Questions and politeness: Strategies in social interaction, pp. 56–311. Cambridge University Press

Brown, P., Levinson, S.C., & Levinson, S.C. (1987). Politeness: Some universals in language usage. Cambridge University Press

Burke, M., & Kraut, R. (2008). Mind your ps and qs: the impact of politeness and rudeness in online communities. In: Proceedings of the 2008 ACM conference on computer supported cooperative work (pp. 281–284). ACM

Caselli, T., Basile, V., Mitrović, J., & Granitzer, M. (2020). Hatebert: Retraining bert for abusive language detection in english. Preprint retrieved from http://arxiv.org/abs/2010.12472

Choudhary, G., Modani, N., & Maurya, N. (2021). React: A review comment dataset for act ionability (and more). In: Web information systems engineering–WISE 2021: 22nd International conference on web information systems engineering, WISE 2021, Melbourne, VIC, Australia, October 26–29, 2021 (pp. 336–343). Springer

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement,* 20(1), 37–46. https://doi.org/10.1177/001316446002000104

Coniam, D. (2012). Exploring reviewer reactions to manuscripts submitted to academic journals. *System,* 40(4), 544–553.

Danescu-Niculescu-Mizil, C., Lee, L., Pang, B., & Kleinberg, J. (2012). Echoes of power: Language effects and power differences in social interaction. In: Proceedings of the 21st international conference on world wide web (pp. 699–708)

Danescu-Niculescu-Mizil, C., Sudhof, M., Jurafsky, D., Leskovec, J., & Potts, C. (2013). A computational approach to politeness with application to social factors. Preprint retrieved from http://arxiv.org/abs/1306.6078

Dueñas, P. M. (2012). Getting research published internationally in english: An ethnographic account of a team of finance spanish scholars' struggles. *Ibérica, Revista de la Asociación Europea de Lenguas para Fines Específicos,* 24, 139–155.

Duthler, K. W. (2006). The politeness of requests made via email and voicemail: Support for the hyperpersonal model. *Journal of Computer-Mediated Communication,* 11(2), 500–521.

Falkenberg, L. J., & Soranno, P. A. (2018). Reviewing reviews: An evaluation of peer reviews of journal article submissions. *Limnology and Oceanography Bulletin,* 27(1), 1–5.

Fortanet, I. (2008). Evaluative language in peer review referee reports. *Journal of English for Academic Purposes,* 7(1), 27–37.

Gao, Y., Eger, S., Kuznetsov, I., Gurevych, I., & Miyao, Y. (2019). Does my rebuttal matter? insights from a major nlp conference. Preprint retrieved from http://arxiv.org/abs/1903.11367

Ghosal, T., Kumar, S., Bharti, P. K., & Ekbal, A. (2022). Peer review analyze: A novel benchmark resource for computational analysis of peer reviews. *Plos one,* 17(1), 0259238.

Gilbert, E. (2012). Phrases that signal workplace hierarchy. In: Proceedings of the ACM 2012 conference on computer supported cooperative work (pp. 1037–1046). ACM

Grice, H.P. (1975). Logic and conversation. In: Speech acts (pp. 41–58). Brill

Herring, S.C. (1994). Politeness in computer culture: Why women thank and men flame. In: Cultural performances: Proceedings of the third Berkeley women and language conference (pp. 278–294)

Hewings, M. (2004). An'important contribution'or'tiresome reading'? a study of evaluation in peer reviews of journal article submissions. *Journal of Applied Linguistics and Professional Practice*, 2004, 247–274.

Holmes, J. (2005). When small talk is a big deal: Sociolinguistic challenges in the workplace. *Second Language Needs Analysis,* 344, 371.

Hua, X., Nikolov, M., Badugu, N., Wang, L. (2019). Argument mining for understanding peer reviews. Preprint retrieved from http://arxiv.org/abs/1903.10104

Hyland, K. (2016). Academic publishing: Issues and challenges in the construction of knowledge-oxford applied linguistics

Hyland, K.(2018). Metadiscourse: Exploring interaction in writing. Bloomsbury Publishing

Hyland, K., & Jiang, F.K . (2020). "This work is antithetical to the spirit of research": An anatomy of harsh peer reviews. *Journal of English for Academic Purposes* 46, 10.

Hyland, K. (2005). Stance and engagement: A model of interaction in academic discourse. *Discourse Studies,* 7(2), 173–192.

Jefferson, T., Rudin, M., Folse, S.B., & Davidoff, F. (2006). Editorial peer review for improving the quality of reports of biomedical studies. *Cochrane Database of Systematic Reviews* 1, 4.

Kang, D., Ammar, W., Dalvi, B., Zuylen, M., Kohlmeier, S., Hovy, E.H., & Schwartz, R. (2018). A dataset of peer reviews (peerread): Collection, insights and NLP applications. In M. A. Walker, H. Ji, A. Stent (eds.) Proceedings of the 2018 conference of the North American chapter of the association for computational linguistics: Human language technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018 (pp. 1647–1661). Association for Computational Linguistics. https://doi.org/10.18653/v1/n18-1149 .

Kendall, M. G., & Smith, B. (1939). The problem of m rankings. *The Annals of Mathematical Statistics,* 10(3), 275–287. https://doi.org/10.1214/aoms/1177732140

Kourilová, M. (1996). Interactive functions of language in peer reviews of medical papers written by non-native users of english. *Unesco ALSED-LSP Newsletter,* 19(1), 4–21.

Krippendorff, K. (2004). Content analysis: An introduction to its methodology (2nd ed). Sage

Lakoff, R.(1973). The logic of politeness: Or, minding your p's and q's. In: Proceedings from the annual meeting of the Chicago linguistic Society (pp. 292–305). Chicago Linguistic Society

Lakoff, R. (1977). What you can do with words: Politeness, pragmatics and performatives. In: Proceedings of the Texas conference on performatives, presuppositions and implicatures 9pp. 79–106). ERIC

Lauscher, A., Glavaš, G., & Ponzetto, S.P. (2018). An argument-annotated corpus of scientific publications. Association for Computational Linguistics

Leech, G.N. (2016). Principles of pragmatics. Routledge

Lin, J., Song, J., Zhou, Z., Chen, Y., & Shi, X. (2022). Moprd: A multidisciplinary open peer review dataset. Preprint retrieved froms http://arxiv.org/abs/2212.04972

Luu, S.T., & Nguyen, N.L.T. (2021). Uit-ise-nlp at semeval-2021 task 5: Toxic spans detection with bilstm-crf and toxicbert comment classification. Preprint retrieved from http://arxiv.org/abs/2104.10100

Matsui, A., Chen, E., Wang, Y., & Ferrara, E. (2021). The impact of peer review on the contribution potential of scientific papers. *PeerJ,* 9, 11999.

Mulligan, A., Hall, L., & Raphael, E. (2013). Peer review in a changing world: An international study measuring the attitudes of researchers. *Journal of the American Society for Information Science and Technology,* 64(1), 132–161.

Mungra, P., & Webber, P. (2010). Peer review process in medical research publications: Language and content comments. *English for Specific Purposes,* 29(1), 43–53.

Obeng, S. G. (1997). Language and politics: Indirectness in political discourse. *Discourse & Society,* 8(1), 49–83.

Paltridge, B.(2017). The discourse of peer review (pp. 978–981). Palgrave Macmillan

Peterson, K., Hohensee, M., & Xia, F. (2011). Email formality in the workplace: A case study on the enron corpus. In: Proceedings of the workshop on language in social media (LSM 2011) (pp. 86–95). LSM

Plank, B., & Dalen, R. (2019). Citetracked: A longitudinal dataset of peer reviews and citations (pp. 116–122). BIRNDL@ SIGIR

Prabhakaran, V., Rambow, O., & Diab, M. (2012). Predicting overt display of power in written dialogs. In: Proceedings of the 2012 conference of the North American chapter of the association for computational linguistics: Human language technologies (pp. 518–522). ACL

Rogers, P. S., & Lee-Wong, S. M. (2003). Reconceptualizing politeness to accommodate dynamic tensions in subordinate-to-superior reporting. *Journal of Business and Technical Communication,* 17(4), 379–412.

Scholand, A.J., Tausczik, Y.R., & Pennebaker, J.W. (2010) Social language network analysis. In: Proceedings of the 2010 ACM conference on computer supported cooperative work (pp. 23–26).

Schwartz, S. J., & Zamboanga, B. L. (2009). The peer-review and editorial system: Ways to fix something that might be broken. *Perspectives on Psychological Science,* 4(1), 54–61.

Shema, H. (2022). The birth of modern peer review. Retrieved July 15, 2022, from https://blogs.scientificamerican.com/information-culture/the-birth-of-modern-peer-review/.

Shen, C., Cheng, L., Zhou, R., Bing, L., You, Y., & Si, L. (2022). Mred: A meta-review dataset for structure-controllable text generation. *Findings of the Association for Computational Linguistics: ACL,* 2022, 2521–2535.

Silbiger, N. J., & Stubler, A. D. (2019). Unprofessional peer reviews disproportionately harm underrepresented groups in stem. *PeerJ,* 7, 8247.

Singh, S., Singh, M., & Goyal, P. (2021). Compare: A taxonomy and dataset of comparison discussions in peer reviews. In: 2021 ACM/IEEE joint conference on digital libraries (JCDL) (pp. 238–241). IEEE

Spencer, S. J., Logel, C., & Davies, P. G. (2016). Stereotype threat. *Annual Review of Psychology,* 67(1), 415–437.

Stappen, L., Rizos, G., Hasan, M., Hain, T., & Schuller, B.W. (2020). Uncertainty-aware machine support for paper reviewing on the interspeech 2019 submission corpus

Swales, J. (1996). Occluded genres in the academy. Academic Writing 1996, 45–58

Verma, R., Roychoudhury, R., Ghosal, T. (2022). The lack of theory is painful: Modeling harshness in peer review comments. In: 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 (pp. 925–935). ACL

Voigt, R., Camp, N. P., Prabhakaran, V., Hamilton, W. L., Hetey, R. C., Griffiths, C. M., Jurgens, D., Jurafsky, D., & Eberhardt, J. L. (2017). Language from police body camera footage shows racial disparities in officer respect. *Proceedings of the National Academy of Sciences,* 114(25), 6521–6526.

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

Year's Best Peer Review Comments: Papers That "Suck the Will to Live" — discovermagazine.com. Retrieved January 02, 2023, https://www.discovermagazine.com/mind/years-best-peer-review-comments-papers-that-suck-the-will-to-live.

Yuan, W., Liu, P., & Neubig, G. (2022). Can we automate scientific reviewing? *Journal of Artificial Intelligence Research,* 75, 171–212.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Prabhat Kumar Bharti[1] · Meith Navlakha[2] · Mayank Agarwal[1] · Asif Ekbal[1]

✉ Prabhat Kumar Bharti
prabhat_1921cs32@iitp.ac.in

✉ Asif Ekbal
asif@iitp.ac.in

Meith Navlakha
meithnavlakha@gmail.com

Mayank Agarwal
mayank265@iitp.ac.in

[1] Department of Computer Science and Engineering, Indian Institute of Technology, Bihta, Patna, Bihar 801106, India

[2] Department of Computer Engineering, Dwarkadas J. Sanghvi College of Engineering, Mumbai, Maharashtra 400056, India