World Wide Web (2023) 26:683–708
https://doi.org/10.1007/s11280-022-01109-z

What have we learned from OpenReview?

Gang Wang[1] · Qi Peng[1] · Yanfeng Zhang[1,2] · Mingyang Zhang[1]

Received: 14 April 2022 / Revised: 26 August 2022 / Accepted: 22 September 2022 /
Published online: 9 November 2022

Abstract
Anonymous peer review is used by the great majority of computer science conferences. OpenReview is such a platform that aims to promote openness in peer review process. The paper, (meta) reviews, rebuttals, and final decisions are all released to public. We collect 11,915 submissions and their 41,276 reviews from the OpenReview platform. We also collect these submissions' citation data from Google Scholar and their non-peer-reviewed versions from arXiv.org. By acquiring deep insights into these data, we have several interesting findings that could help understand the effectiveness of the public-accessible double-blind peer review process. Our results can potentially help writing a paper, reviewing it, and deciding on its acceptance.

Keywords Peer review · OpenReview · Opinion divergence

1 Introduction

Peer review is a widely adopted quality control mechanism in which the value of scientific paper is assessed by several reviewers with a similar level of competence. The primary role of the review process is to decide which papers to publish and to filter information, which is particularly true for a top conference that aspires to attract a broad readership to

Qi Peng contributed equally to this work.

This article belongs to the Topical Collection: *APWeb-WAIM 2021*
Guest Editors: Yi Cai, Leong Hou U, Marc Spaniol, Yasushi Sakurai

✉ Yanfeng Zhang
zhangyf@mail.neu.edu.cn

Gang Wang
wanggangneu@stumail.neu.edu.cn

Qi Peng
ffpengqi@stumail.neu.edu.cn

Mingyang Zhang
theremay@outlook.com

[1] School of Computer Science and Engineering, Northeastern University, Shenyang 110169, China

[2] Key Laboratory of Intelligent Computing in Medical Image, Ministry of Education, Shenyang 110169, China

its papers. The novelty, significance, and technical flaws are identified by reviewers, which can help PC chair make the final decision.

Anonymous peer review (no matter single-blind or double-blind), despite the criticisms often leveled against it, is used by the great majority of computer science conferences, where the reviewers do not identify themselves to the authors. It is understandable that some authors are uncomfortable with a system in which their identities are known to the reviewers while the latter remain anonymous. Authors may feel themselves defenseless against what they see as the arbitrary behavior of reviewers who cannot be held accountable by the authors for unfair comments. On the other hand, apparently, there would be even more problems if letting authors know their reviewers' identities. Reviewers would give more biased scores for fear of retaliation from the more powerful colleagues. Given this contradiction, opening up the reviews to public seems to be a good solution. The openness of reviews will force reviewers to think more carefully about the scientific issues and to write more thoughtful reviews, since PC chairs know the identities of reviewers and bad reviews would affect their reputations.

OpenReview[1] is such a platform that aims to promote openness in peer review process. The paper, (meta) reviews, rebuttals, and final decisions are all released to public. Colleagues who do not serve as reviewers can judge the paper's contribution as well as judge the fairness of the reviews by themselves. Reviewers will have more pressure under public scrutiny and force themselves to give much fairer reviews. On the other hand, previous works on peer-review analysis [1–6] are often limited due to the lack of rejected paper instances and their corresponding reviews. Given these public reviews (for both accepted papers and rejected ones), studies towards multiple interesting questions related to peer-review are made available.

Given these public reviews, there are multiple interesting questions raised that could help us understand the effectiveness of the public-accessible double-blind peer review process: a) As known, AI conferences have extremely heavy review burden in 2020 due to the explosive number of submissions [7]. These AI conferences have to hire more non-experts to involve in the double-blind review process. How is the impact of these non-experts on the review process (Section 3.1)? b) Reviewers often evaluate a paper from multiple aspects, such as motivation, novelty, presentation, and experimental design. Which aspect has a decisive role in the review score (Section 3.2)? c) The OpenReview platform provides not only the submission details (e.g., title, keywords, and abstract) of accepted papers but also that of rejected submissions, which allows us to perform a finer-grained cluster analysis. Given the fine-grained hierarchical clustering results, is there significant difference in the acceptance rate of different research fields (Section 3.3)? d) A posterior quantitative method for evaluating papers is to track their citation counts. A high citation count often indicates a more important, groundbreaking, or inspiring work. OpenReview releases not only the submission details of accepted papers but also that of rejected submissions. The rejected submissions might be put on arXiv.org or published in other venues to still attract citations. This offers us opportunities to analyze the correlation between review scores and citation numbers. Is there a strong correlation between review score and citation number for a submission (Section 3.4)? e) Submissions might be posted on arXiv.org before the accept/reject notification, which might be the rejected ones from other conferences. They are special because they could be improved according to the rejected reviews and their authors are not anonymous. Are these submissions shown higher acceptance rate (Section 3.5)? f) The rebuttal is an opportunity provided by the OpenReview platform

[1] https://openreview.net/

for authors and reviewers to communicate. A good rebuttal may improve the score of the paper. How to write a rebuttal to boost the review score (Section 3.6)?

In this paper, we collect 11,915 (accepted and rejected) submissions and their 41,276 reviews from ICLR 2017-2022 venues[2] on the OpenReview platform as our main corpus. By acquiring deep insights into these data, we have several interesting findings and aim to answer the above raised questions quantitatively. Our submitted supplementary file also includes more data analysis results. We expect to introduce more discussions on the effectiveness of peer-review process and hope that treatment will be obtained to improve the peer-review process.

2 Dataset

ICLR has used OpenReview to launch double-blind review process for 10 years (2013-2022). Similar to other major AI conferences, ICLR adopts a reviewing workflow containing double-blind review, rebuttal, and final decision process. After paper assignment, typically three reviewers evaluate a paper independently. After the rebuttal, reviewers can access the authors' responses and other peer reviews, and accordingly modify their reviews. The program chairs then write the meta-review for each paper to make the final accept/reject decision according to the three anonymous reviews. Each official review mainly contains a review score (integer between 1 and 10), a reviewer confidence level (integer between 1 and 5), and the detailed review comments. The official reviews and meta-reviews are all open to the public on the OpenReview platform. Public colleagues can also post their reviews on OpenReview. We will present the collected dataset of submissions and reviews from OpenReview, these submissions' citation data from Google Scholar, and their non-peer-reviewed versions from arXiv.org.[3]

Submissions and reviews We have collected 11,939 submissions and 41,276 official reviews from ICLR 2017-2022 venues on the OpenReview platform. We only use the review data since 2017 because the submissions before 2017 is too few. Though a double-blind review process is exploited, the authors' identities of the rejected submissions are also released after decision notification. Thus, we can also access the identity information for each rejected submission, which is critical in most of our analysis. Some statistics of the reviews data are listed in Table 1, in which review len. indicates the average word count of the review.

Citations In order to investigate the correlation between review scores and citation numbers, we also collect the citation information from Google Scholar for all the 3,685 accepted papers from 2017 to 2022. Since the rejected submissions might be put on arXiv.org or published in other venues, they might also attract citations. We also collect the citation information for 8,230 rejected submissions that have been published elsewhere (210 for 2017, 324 for 2018, 493 for 2019, 955 for 2020, 474 for 2021, 393 for 2022, and totally 2849 rejected papers). All the citation numbers are gathered up to 31 Mar. 2022.

[2] International Conference on Learning Representations. https://iclr.cc/

[3] These datasets and the source code for the analysis experiment are available at https://github.com/Seafoodair/Openreview/

Table 1 Statistics of ICLR reviews dataset

year	#papers	#authors	accept rate	#reviews	review len.
2017	489	1,417	50.1%	1,495	295.11
2018	939	2,882	49.0%	2,849	372.07
2019	1,541	4,332	32.5%	4,733	403.22
2020	2,558	7,765	26.5%	7,766	407.08
2021	2,966	8,751	29.0%	11,291	465.51
2022	3,422	10,475	31.9%	13,142	355.34
total	11,915	35,622	36.5%	41,276	383.06

arXiv submissions In order to investigate whether the submissions that have been posted on arXiv.org before notification have a higher acceptance rate, we also crawl the arXiv versions of ICLR 2017-2022 submissions if they exist. We record the details of an arXiv preprint if its title matches an ICLR submission title. Note that, their contents might be slightly different. We totally find 3,532 matched arXiv papers and 2,761 among them were posted before notification (178/150 for 2017, 103/79 for 2018, 420/303 for 2019, and 457/416 for 2020, 1093/787 for 2021, 1281/1026 for 2022) up to 24 Mar 2022.

3 Results learned from open reviews

3.1 How is the impact of non-expert reviewers?

Due to the extensively increasing amount of submissions, ICLR 2020 hired much more reviewer volunteers. There were complaints about the quality of reviews (47% of the reviewers have not published in the related areas [7]). Similar scenarios have been observed in other AI conferences, such as NIPS, CVPR, and AAAI. Many authors complain that their submissions are not well evaluated because the assigned "non-expert" reviewers lack of enough technical background and cannot understand their main contributions. How is the impact of these "non-experts" on the review process? In this subsection, we aim to answer the question through quantitative data analysis (Table 2).

Table 2 Statistics of different confidence level reviews

Time Level	level1	level2	level3	level4	level5
(a) 2017-2019					
#reviews	74	455	2,330	4,612	1,600
fraction	0.80%	5.01%	25.67%	50.81%	17.71%
(b) 2020					
#reviews	1,104	2,554	2,659	1,449	—
fraction	14.22%	32.89%	34.24%	18.66%	—
(c) 2021-2022					
#reviews	110	1,404	7,265	12,362	3,476
fraction	0.45%	5.70%	29.51%	50.22%	14.12%

Review score distribution For ICLR 2017-2019, reviewer gives a review score (integer) from 1 to 10, and is asked to select a confidence level (integer) between 1 and 5. For ICLR 2020, reviewer gives a rating score in {1, 3, 6, 8} and should select an experience assessment score (similar to confidence score) between 1 and 4. For ICLR 2021-2022, the same mechanism is adopted as ICLR 2017-2019. We divide the reviews into multiple subsets according to their confidence levels. Figure 1 shows the smoothed review score distributions for each subset of reviews. For ICLR 2018 and 2021, we consistently observe that the scores of reviews with confidence level 1 and 2 are likely to be higher than those reviews with confidence level 4 and 5. For ICLR 2020, we can observe that in low-score areas, the fraction of the scores of reviews with confidence level 4 and 5 is higher than those reviews with confidence level 1 and 2. The trend of ICLR 2017 is not clear because it contains too few samples to be statistically significant (e.g., only 7 level-1 reviews). In 2017-2019, the lowest confidence level reviews has an average review score 5.675, while the highest confidence level reviews has an average review score 4.954. In 2020, the numbers for the lowest and highest confidence level reviews are 4.726 and 3.678, respectively. In 2021, the numbers for the lowest and highest confidence level reviews are 5.663 and 5.214, respectively. In 2022, the numbers for the lowest and highest confidence level reviews are 5.529 and 5.001, respectively. Our results show that the low-confidence reviewers (e.g., level 1 and 2) tend to be more tolerant because they may be not confident about their decision, while the high-confidence reviewers (e.g., level 4 and 5) tend to be more tough and rigorous because they may be confident in the identified weakness.

Significant difference analysis In order to evaluate the mean difference between non-expert and expert reviewers, we use the method of hypothesis testing. 'T Test' is one of the most widely used techniques for testing a hypothesis based on a difference between sample means. We perform a t-test with observed scores and compute the effect size to examine if

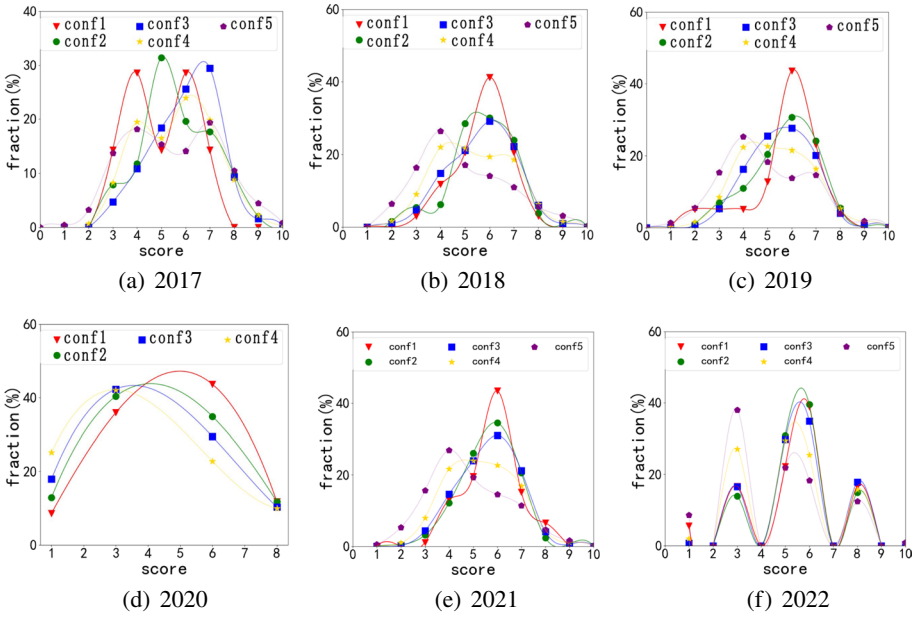


Figure 1 The review score distributions of different confidence level (conf) reviews

there is a statistically significant difference in the underlying means of the scores provided by different confidence levels of reviewers. Firstly, we propose a null hypothesis that there is no difference between non-expert and expert reviewers. Secondly, we propose an alternative hypothesis (H1) that there are differences among different types of reviewers. In other words, many people think that junior reviewers are often perceived to be more critical than senior reviewers. In this section, we measure the 'T Test' in the different samples and show the result in Table 3. The results indicate that all P-values are less than 0.05 in the table. In many fields of scientific research, P-value less than 0.05 is equivalent to a significant difference. Note that P-value only represents statistical significance. Therefore, we use the effect size to measure the significance of the differences between groups. A large effect size means a research finding has practical significance, while a small effect size indicates limited practical applications. As shown in the table, those results show that the d value of effect size is between [0.2, 0.5], which is the influence of a small effect. From 2017 to 2019, the effect size d value is the lowest due to the small base of non-experts, so their impact is relatively small. In 2021, the effect size d value was the largest. It shows that the opinions of non-experts have a greater impact on the papers. Our experiments show that there are different opinions between non-experts and expert reviewers, but the effect size is insignificant.

Divergence reflected by Euclidean distance On the other hand, peoples are worrying that non-expert reviewers are not competent to give a fair evaluation of a submission (e.g., fail to identify key contributions or fail to identify flaws) and will ruin the reputation of top conferences [7]. Particularly, these non-expert reviewers may have different opinions with the expert reviewers regarding the same paper. Actually, opinion divergence commonly exists between reviewers in the peer-review process. Each paper is typically assigned to 3 reviewers. These 3 reviewers may have significantly different review scores. In order to illustrate the difference between the reviews with different confidence scores, we first compute the euclidean distance $DIS(l_i, l_j)$ between between group l_i and group l_j as follows. Let R_{l_i,l_j} be the set of paper IDs, where each paper concurrently has both confidence-l_i review(s) and confidence-l_j review(s). Let $\overline{s_p^i}$ be paper p's average review score from l_i-confidence reviews. Then, the distance between the group of confidence-l_i reviews and that of confidence-l_j reviews is:

$$DIS(l_i, l_j) = \sqrt{\sum_{p \in R_{l_i,l_j}} \left(\overline{s_p^i} - \overline{s_p^j}\right)^2}. \tag{1}$$

After computing the distance betwenn each pair of groups, we can construct a distance matrix. According to the distance matrix, we use t-SNE [8] to plot the visualized layout of different groups of reviews of each year in Figure 2. For ICLR 2017-2019, we can see similar layout, where conf1 reviews are close to conf2 reviews in the central part and the groups of conf3, conf4, and conf5 locate around. The group of conf4 reviews is far apart from the most professional reviews (conf5). In ICLR 2020, there are 4 confidence levels. Surprisingly, we observe that the most professional reviews (conf4) and least professional

Table 3 Calculate the difference p-value and effect size d value between different confidence reviews

Samples	2017-2019	2020	2021	2022
P-value	6.79×10^{-7}	2.45×10^{-30}	1.84×10^{-20}	4.67×10^{-20}
Cohen's d	0.21	0.26	0.34	0.32

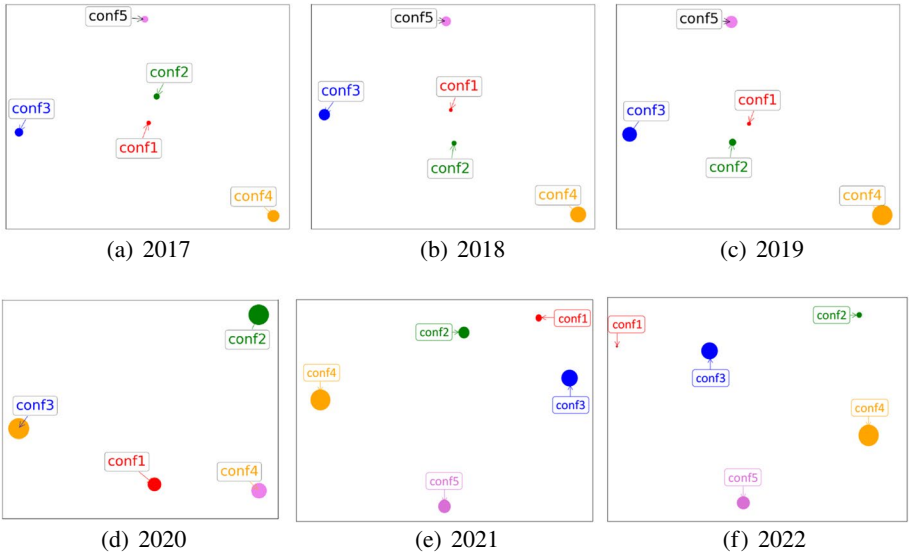


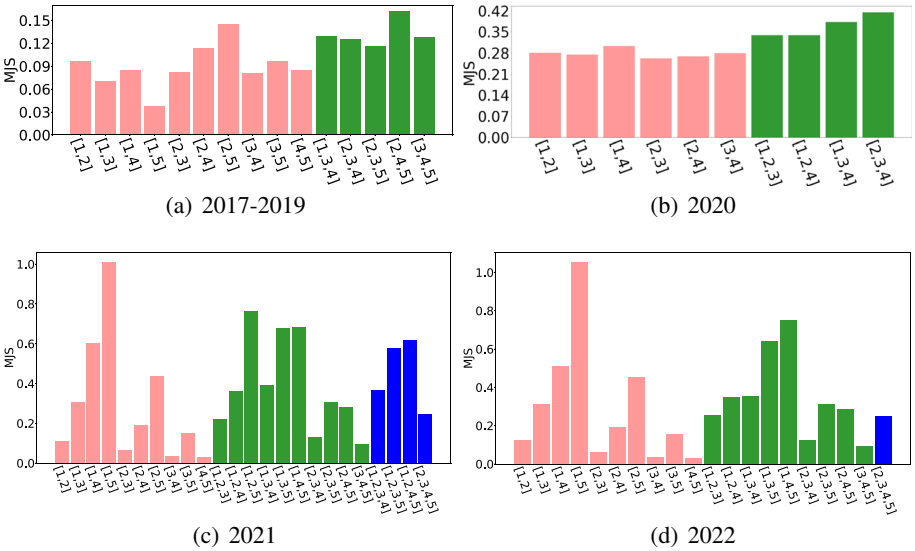
Figure 2 The visualized layout of groups of reviews with different confidence scores. Each point indicates a group of reviews with a specific confidence level (abbrv. conf). The size of point indicates the relative number of reviews in that group. The distance between two points indicates the divergence of review scores between two groups

reviews (conf1) are closest to each other. Conf2 reviews and conf3 reviews are both far apart from conf4 reviews. For ICLR 2021-2022, conf1 reviews are close to conf3 reviews in relative position. The group of conf4 reviews is far apart from the most professional reviews (conf5). Our results show that conf1 reviews are far apart from conf5 reviews in recent years. This is because many non-professional reviewers have been introduced and there are gaps in the knowledge of professional fields.

Divergence reflected by Jensen-Shannon divergence By using euclidean distance, we can only measure the divergence of two sets of different level reviews. Inspired by Jensen-Shannon Divergence for multiple distributions (MJS) [9], we design the MJS metric to measure the divergence between multiple sets of reviews. The MJS of m sets ($m \geq 2$) of different confidence reviews is defined as follows:

$$MJS(l_1, \ldots, l_m) = \frac{1}{m} \sum_{i \in \{l_1, \ldots, l_m\}} \left(\frac{1}{| R_{l_1, \ldots, l_m} |} \sum_{p \in R_{l_1, \ldots, l_m}} \overline{s_p^i} \cdot log \left(\frac{\overline{s_p^i}}{s_p^{[1,m]}} \right) \right) \tag{2}$$

where $R_{l_1, \ldots, l_m}$ is the set of paper IDs, where each paper concurrently has reviews with confidence levels $l_1, \ldots, l_m$, $\cdot$ returns the size of a set, s_p^i is paper p's average review score of l_i-confidence reviews, and $s_p^{[1,m]}$ is paper p's average review score of reviews with confidence levels $l_1, \ldots, l_m$. The bigger the *MJS* is, the significant the opinion divergence is. A nice property of MJS metric is that it is symmetric, e.g., $MJS(i, j) = MJS(j, i)$ and $MJS(i, j, k) = MJS(k, j, i)$. We measure the MJS divergence of different combinations of confidence levels and show the results in Figure 3. Note that, the results of combinations that contain less than 10 reviews are not shown since they are too few to be statistically significant. In



(a) 2017-2019 (b) 2020

(c) 2021 (d) 2022

Figure 3 MJS divergence of different combinations of different confidence level reviews

2017-2019 the MJS divergence between conf1 reviews and conf5 reviews is the smallest. In 2020, it shows bigger divergence than 2017-2019 on different combinations but relatively similar divergence results among different combinations. In addition, a combination of three different confidence levels is likely to result in bigger divergence than a combination of two confidence levels. We observe that the MJS divergence between different confidence-level reviews is the more significant in 2021-2022. It shows a bigger divergence than 2017-2020 on different combinations. After further analysis, we find that divergence difference mainly exists between non-expert and expert reviewers. The reason behind this might be the extensively increasing amount of submissions. ICLR hired more and more reviewer volunteers. There exist biases among different reviewers.

Divergence reflected by average variance Opinion divergence also exists within the same confidence-level reviews. The above measurements cannot depict the intra-level opinion divergence. Here, we use average variance to measure the intra-level opinion divergence and the inter-level opinion divergence. The average variance of m sets ($m \geq 1$) of different confidence reviews is defined as follows.

$$VAR(l_1, \ldots, l_m) = \frac{1}{m}\left(\sum_{p \in R^n_{l_1,\ldots,l_m}} var\left(s_p^1, s_p^2, \ldots, s_p^n\right)\right), \tag{3}$$

where $R^n_{l_1,\ldots,l_m}$ is a set of paper IDs, where each paper concurrently has reviews with confidence levels l_1,…, l_m and the number of reviews with confidence levels l_1,…, l_m is n ($n \leq m$), and $var(s_p^1, s_p^2, \ldots, s_p^n)$ is the variance of paper p's n review scores. Since we will compare the VAR values of different combinations of different confidence level reviews, we have to make sure that the number of samples for variance computation are equal to each other, which can be achieved by introducing the fixed number n. ICLR papers typically have 3 reviews, so we set $n = 3$. The average variance results are shown in Figure 4. We do

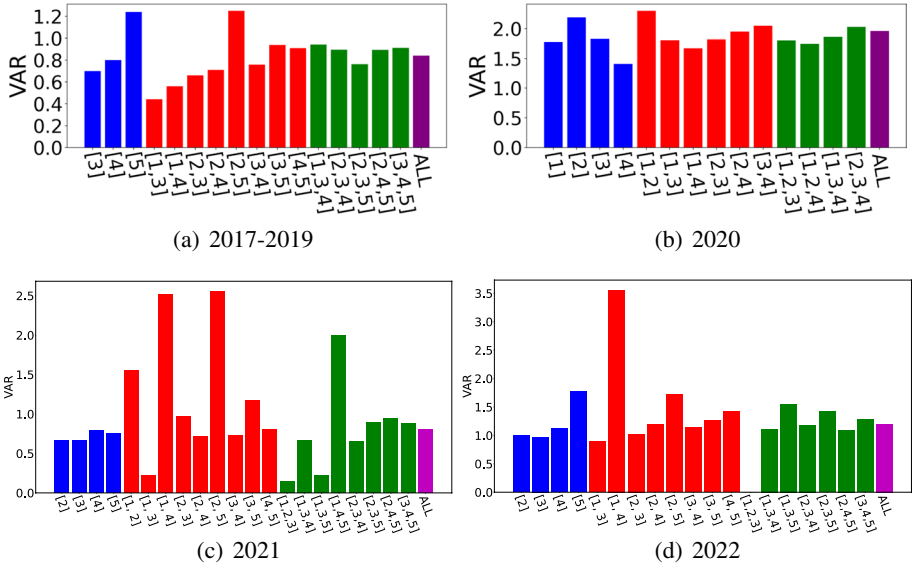


Figure 4 Average variance of different combinations of confidence levels

not show the results of combinations that contain less than 10 samples. Since there is one more constraint that each combination has to include 3 reviews (refer to the definition of $R^n_{l_1,\ldots,l_m}$), less bars are shown in Figure 4 than in Figure3. In 2017-2019, it is surprised that the maximum variance appears among the most professional reviews (i.e., conf5). While in 2020, the most professional reviews (i.e., conf4) have the minimum variance. It also shows that the variance between the professional reviews and the non-professional reviews is relatively small no matter in 2017-2019 (e.g., conf[1,4]) or in 2020 (e.g., conf [1,4]). In 2021-2022, It also shows that the variance between the professional reviews and the non-professional reviews is relatively large (e.g., conf[1,4], conf[2,5]). Our results show that opinion divergence also exists within the same confidence-level reviews. Maybe the reason is that different reviewers have different interpretations of the papers.

How is the impact of non-expert reviewers? All these facts demonstrate that after the introduction of non-professional reviewers, differences of opinion divergence exist but have little impact. We also observe that the opinion divergence between non-expert reviewers and other reviewers is often relatively larger in recent year. The reason behind might be that the expert reviewers often have a more reject opinion than non-expert reviewers. They have enough confidence in the reviewed papers. On the contrary, non-professional reviewers are are more cautious to give positive or negative recommendations.

3.2 Which aspects play important roles in review score?

Reviewers often evaluate a paper from various aspects. There are five most important aspects, i.e., novelty, motivation, experimental results, completeness of related workers, and presentation quality. Some conferences provide a peer-review questionnaire that requires reviewer to evaluate a paper from various aspects and give a score with respect to

each aspect. Unfortunately, ICLR does not ask reviewers to answer such a questionnaire. Then a question arises accordingly. Which aspects play more important roles in determining the review score? We aim to answer this question by analyzing the sentiment of each aspect.

Corpus creation For each review, we first extract the related sentences that describe different aspects by matching a set of predefined keywords. The keywords "novel, novelty, originality, and idea" are used to identify a sentence that describes novelty of the paper, "motivation, motivate, and motivated" are used to identify a sentence related to motivation, "experiments, empirically, empirical, experimental, evaluation, results, data, dataset, and data set" are used to identify a sentence related to experiment results, "related work, survey, review, previous work, literature, cite, and citation" are used to identify a sentence related to the completeness of related work, and "presentation, writing, written, structure, organization, structured, and explained" are used to identify a sentence related to presentation quality. We have collected a corpus containing 95,208 sentences which are divided into five subsets corresponding to the five aspects. Specifically, we have 11,916 sentences related to "novelty", 5,107 sentences related to "motivation", 62,446 sentences related to "experimental results", 8,710 sentences related to "completeness of related work", and 7,029 sentences related to "presentation quality".

Automatic annotation In order to train a sentiment analysis model, we need to first annotate enough number of sentences with sentiment label (i.e., positive, negative, and neutral). However, this workload of manual annotation is huge due to the large size of review corpus. Fortunately, we find a possibility of automatic annotation after analyzing the reviews. A large number of reviewers write their positive reviews and negative reviews separately by using the keywords such as "strengths/weaknesses", "pros/cons", "strong points/weak points", "positive aspects/negative aspects", and so on. We segment the review text and identify the positive/negative sentences by looking up these keywords. The boundaries are identified when meeting an opposite sentiment word for the first time. By intersecting the set of positive/negative sentences with the set of aspect-specific sentences, we obtain a relatively large set of sentiment-annotated corpus for each aspect. Particularly, we have 2,893 sentiment-annotated sentences for "novelty", 1,057 for "motivation", 8,956 for "experimental results", 1,402 for "completeness of related work", 1,644 for "presentation quality", and 15,952 in total. We also manually annotate 6,095 sentences including 1,227 corrected automatically annotated sentences since some neutral sentences are incorrectly annotated with positive or negative sentiment. Finally, we have 20,820 labeled sentences,[4] i.e., 21.87% of the total number of sentences (95,208) in corpus. Note that, there might be more than one sentences describing one aspect but having different sentiments. In such a case, we label the sentiment by a majority vote.

Sentiment analysis Given these five datasets including the labeled data, we perform sentiment analysis for each aspect using a pretrained text model ELECTRA [10] which was recently proposed in ICLR 2020 with state-of-the-art performance. The results demonstrate that ELECTRA achieves better contextual sentiment analysis compared to the CSNN [11] model. The detailed hyper-parameter settings of ELECTRA are described in our support materials. We split the annotated dataset of each aspect into training/validation/test sets

[4] All of the annotated data including manually annotated ones are publicly available at at https://github.com/Seafoodair/Openreview/.

Table 4 Model accuracy results

Aspects	Novelty	Motivation	Experiment	Related work	Presentation
TextCNN	84.36%	81.25%	90.63%	83.33%	86.84%
BERT	95.15%	79.68%	91.70%	82.97%	94.01%
ELECTRA	93.96%	88.46%	94.99%	85.12%	93.38%
T5	85.99%	84.39%	92.24%	83.48%	89.22%

(8:1:1), and use 10-fold cross validation to train five sentiment prediction models for the five aspects. We obtain five accuracy results 93.96%, 88.46%, 94.99%, 85.12%, and 93.38% for novelty, motivation, experimental results, completeness of related workers, and presentation quality, respectively. Then, we conduct experiments with multiple different models. Table 4 shows the multiple model accuracy results of every aspect. We can see that the pretrained models BERT and ELECTRA show better results than the other two models. ELECTRA is slightly better than BERT, so we use ELECTRA in our text analysis task. Next, we then use the whole annotated dataset of each aspect to train the corresponding sentiment analysis model and use this model to predict the sentiment of the other unlabeled sentences of each aspect. Finally, for each review, we can obtain the sentiment score of each aspect. Note that, some individual aspects might not be mentioned in a review, which are labeled with neutral.

Sentiment of each aspect vs. review score Given the sentiment analysis results of all aspects of each review and the review score, we perform the correlation analysis. We group the reviews with the same combination of aspect sentiments and compute the average review score of each group. The groups that receive less than 3 reviews are not considered since they have too few samples to be statistically significant. We visualize the result as shown in Figure 5. We can see that the higher review score often comes with more positive aspects from a macro perspective, which is under expectation. We observe that most of the reviews with score higher than 6 do NOT have negative comments on novelty, motivation, and presentation, but may allow some flaws in related work and experiment. The reviewers that have overall positive to the paper are likely to pose improvement suggestions on related work and experiment to make the paper perfect. The presentation quality and experiment seem to be mentioned more frequently than the other aspects, and the positive sentiment on presentation is distributed more evenly from high-score reviews to low-score

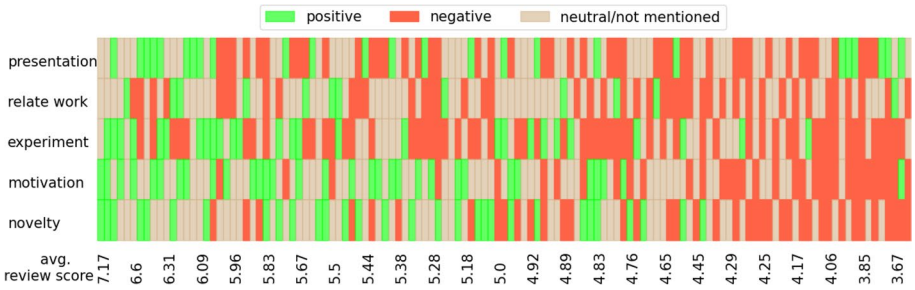


Figure 5 The sentiment of each aspect vs. the review score. Each column represents a group of reviews with the same combination of aspect sentiments. These groups are sorted in the descending order of the average review score of a group of reviews

reviews. This implies that presentation does not play important role in making the decision. It is also interesting that there is no review in which all aspects are positive or negative. It is unlikely that a paper is perfect in all aspects or has no merit. Reviewers are also likely to be more rigorous in ' papers and be more tolerant with poor papers.

Causality analysis In order to explore which aspect determines the final review score, we perform causal inference following [12]. Besides the above five aspects, we also include the factor of reviewer confidence. The process of causal analysis includes four steps: modeling, intervention, evaluation, and inference. In the modeling process, we use multivariate linear regression method [13] to perform regression task on the ICLR reviews dataset, where the six evaluated parameters are the sentiment scores of the five review aspects and a reviewer confidence score, and the regression label is the review score. Each parameter is standardized to [-1,1]. To avoid randomness of model training, we launch 1000 times of training and obtain the average MSE (Mean Square Error) 0.24. The intervention process removes each factor x one by one and performs multiple times of model evaluation to obtain multiple average MSE results, each corresponding to an x-absence model. In the absence of overfitting, the MSE value of any x-absence model should be larger than 0.24. The MSE value of the x-absence model implies the causality. A larger MSE value of an x-absence model implies that the factor x is more dominant in determining the final score, and vice versa. In the inference process, we compare the MSE values to infer the causality. The average MSE values of the reviewer confidence, novelty, motivation, experiment, related work, and presentation are 0.84, 0.77, 0.34, 0.86, 0.33, and 0.34, respectively. We observe that the factors of reviewer confidence, novelty, and experiment change the MSE greatly, so they are more dominant in determining the final score.

3.3 Which research field has higher/lower acceptance rate?

AI conferences consider a broad range of subject areas. Authors are often asked to pick the most relevant areas that match their submissions. Area chair could exist who makes decisions for the submissions of a certain research area. Different areas may receive different number of submissions and also may have different acceptance rates. Program chairs sometimes announce the number of submissions and the acceptance rate of each area in the opening event of a conference, which could somehow indicate the popularity of each area. But, the classification by areas is coarse. A more fine-grained classification that provides more specific information is desired. Thanks to the more detailed submission information provided by OpenReview, we utilize the title, abstract, and keywords of each submission to provide a more fine-grained clustering result and gather the statistics of acceptance rate of each cluster of submissions.

We first concatenate the title, abstract, keywords of each ICLR 2020 submission and preprocess them by removing stop words, tokenizing, stemming list, etc. We leverage an AI terminology dictionary [14] during the tokenizing process to make sure that an AI terminology containing multiple words is not split. We then formulate term-document matrix (i.e., AI term-submission matrix) by applying TF-IDF and calculate cosine distance matrix. The size of the term-document TF-IDF matrix for ICLR 2020 is 12436 x 2558, and the size of the cosine distance matrix is 2558 x 2558. We then apply the Ward clustering algorithm [15] on the matrix to obtain submission clusters. Ward clustering is an agglomerative hierarchical clustering method, meaning that at each stage, the pair of clusters with minimum between-cluster distance are merged. We use silhouette coefficient to finalize the

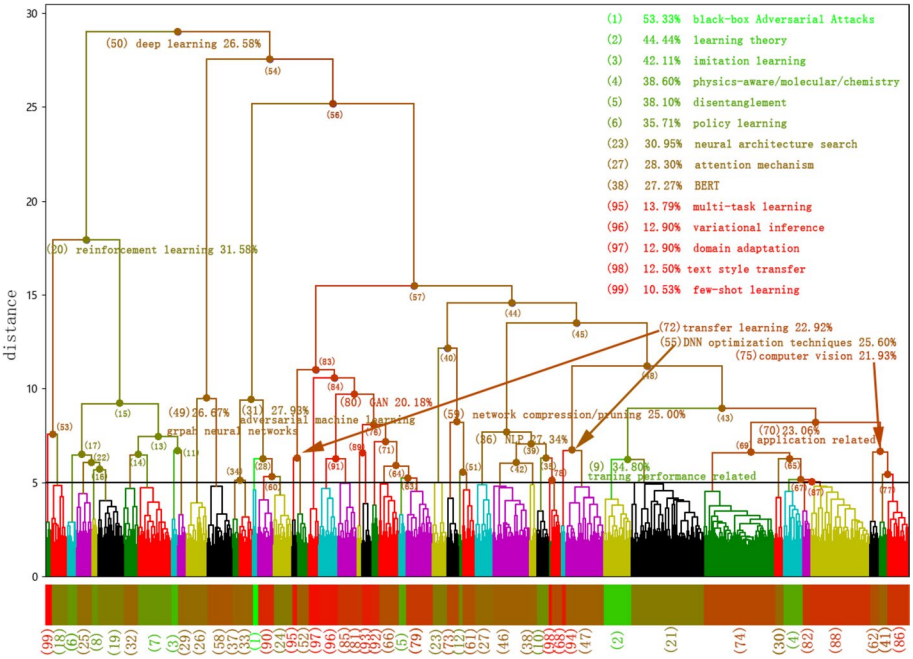


Figure 6 Visualized hierarchical clustering result of ICLR 2020 submissions. Each leaf node represents a submission. Cosine distance 5 is selected as the threshold to control the granularity of leaf-level clusters. There are 99 clusters in total, including both fine-grained clusters and coarse-grained clusters. Clusters are numbered in the order of their acceptance rate. The color of keywords indicates the acceptance rate of that cluster. Light green means a high acceptance rate, while light red means a low acceptance rate. The keywords of some typical clusters are labeled

number of clusters and plot a dendrogram to visualize the hierarchical clustering result as shown in Figure 6.

From Figure 6, we observe three aspects of insights. **(a) Overall Structure of Deep Learning Research**. We observe the correlation between research topics. For example, the submissions in the left part belong to reinforcement learning field (20), which is far apart from all the other research topics (because it is the last merged cluster and its distance to the other clusters is more than 27). Another independent research field is Graph Neural Networks (GNNs) (49), as a promising field, becomes really hot in only 2-3 years, which distinguishes itself from others by focusing on graph structure. Adversarial Machine Learning (31) is also an independent research field that attempts to fool models through malicious input and different from others. The next independent subject is Generative Adversarial Networks (GANs) (80). But GANs is not completely independent since we found that many submissions on NLP (36) and CV (75) are mixed with GANs as well. We also observe that Transfer Learning (72) is close to GANs, since some works have applied transfer learning to GANs. Most of the submissions in the right part are applications related (e.g., vision, audio, NLP, biology, chemistry, and robotics). They are mixed with DNN optimization techniques since many optimizations are proposed to improve DNN on a specific application field. **(b) Popularity Difference between Clusters**. We observe that multiple areas attract large amount of submission. For example, Reinforcement Learning (20), GNNs (49), GANs (80), NLP (36), and Computer Vision (75) have

attracted more than 50% of the submissions, which are really hot topics in today's deep learning research. **(c) Acceptance Rate Difference between Clusters**. There exists significant difference on acceptance rate between clusters, say ranging from 53.33% to 10.53%. The cluster of submissions on "Black-Box Adversarial Attacks" has the highest acceptance rate (53.33%), which is a subject belongs to "Adversarial Machine Learning" area. The top-6 highest acceptance rate topics are listed in the figure. The cluster of submissions on "Few-Shot Learning" has the lowest acceptance rate (10.53%), which is a subject belongs to "Reinforcement Learning" area. The top-5 lowest acceptance rate topics are listed in the figure. We also list some typical topics in the figure. For example, the cluster on "Graph Neural Networks (49)" has an acceptance rate of 26.67%. The cluster on "BERT (38)" has an acceptance rate of 27.27%. The cluster on "GANs (80)" has an acceptance rate of 20.18%. The cluster on "Reinforcement Learning (20)" has an acceptance rate of 31.58%.

3.4 Review score vs. citation number

In this subsection, we show several interesting results on the correlation between review scores and citation numbers.

Is there a strong correlation between review score and citation number? OpenReview releases not only the submission details and reviews of the accepted papers but also that of the rejected submissions. These rejected submissions might be put on arXiv.org or published in other venues and still make an impact. We collect the citation

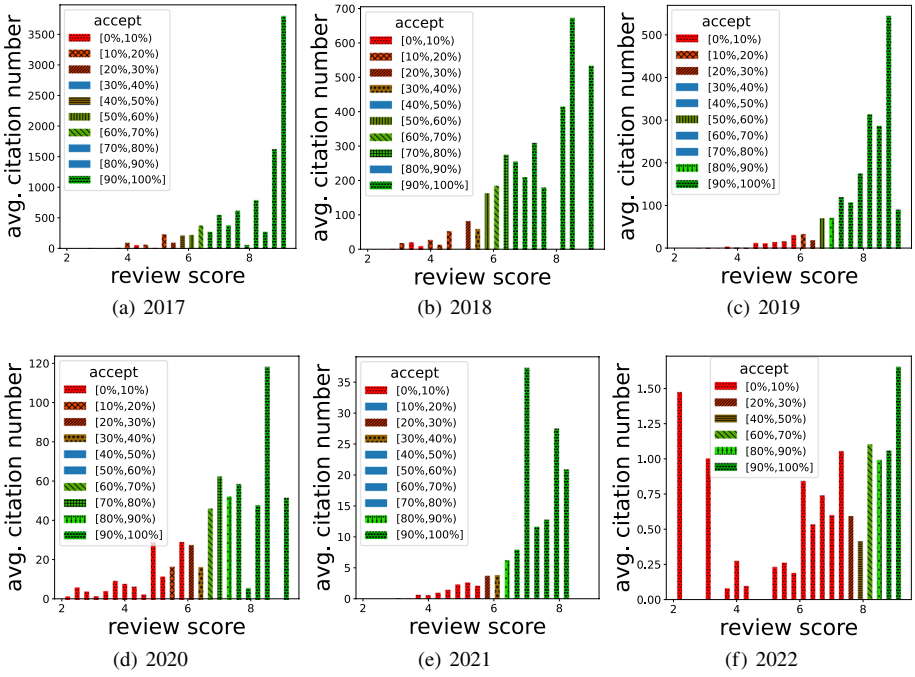


Figure 7 The histogram of citation numbers against 0.3-intervals of average review score

number information of both accepted papers and rejected papers and study the correlation between their review scores and their citation numbers. We plot the histogram of average citation numbers of ICLR 2017-2022 submissions as shown in Figure 7. The papers are divided into multiple subsets according to their review scores. Each bin of the histogram corresponds to a subset of papers with similar review scores (with an interval of 0.3). Then the average citation number of each subset is calculated. The color of bin indicates the acceptance rate of the corresponding subset of papers. From the figure, we can observe that the papers with higher review score are likely to have higher citation numbers, which is under expectation.

We further investigate the citation numbers of individual papers as shown in Figure 8. Each point represents a paper. Green color indicates an accepted paper and red color indicates a rejected one. The papers are sorted on the x-axis in the ascending order of their review scores. The distribution of citation numbers is messy. We can see that many rejected papers gain a large number of citations (i.e, red points in the top-left part), which is a bit surprised. Generally speaking, the accepted papers will attract more attentions since they are officially published in ICLR. However, the rejected papers may be accepted later at other venues and still attract attentions. In addition, a few papers with high review score are rejected (i.e., red points on the right side). We observe that the reject decision does not impact their citation numbers. Though rejected, the papers with higher review score are still likely to have higher citation numbers. An interesting finding that differs from that of ICLR 2017-2020 is that there are more rejected papers gain high citations. The possible reason could be that more papers are rejected by the reviewers but they can still attract great attentions after published on arXiv or other platforms.

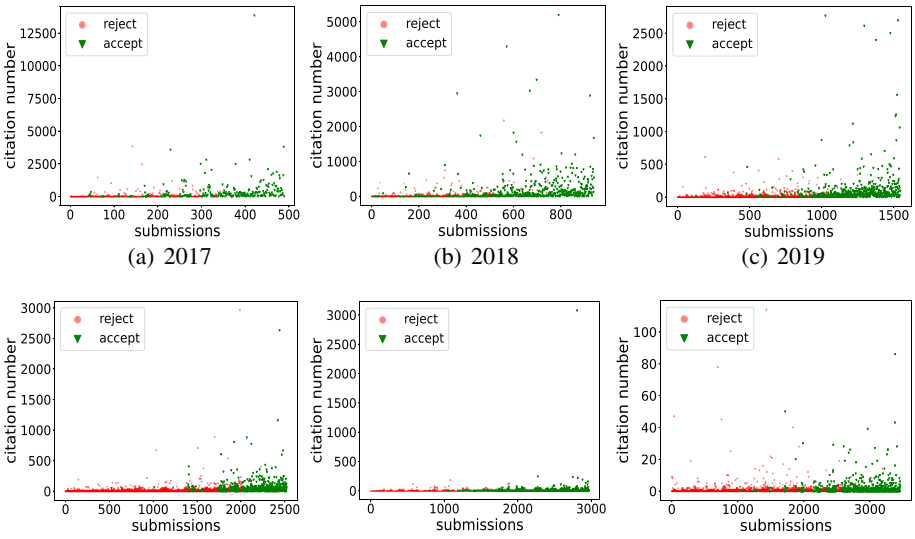
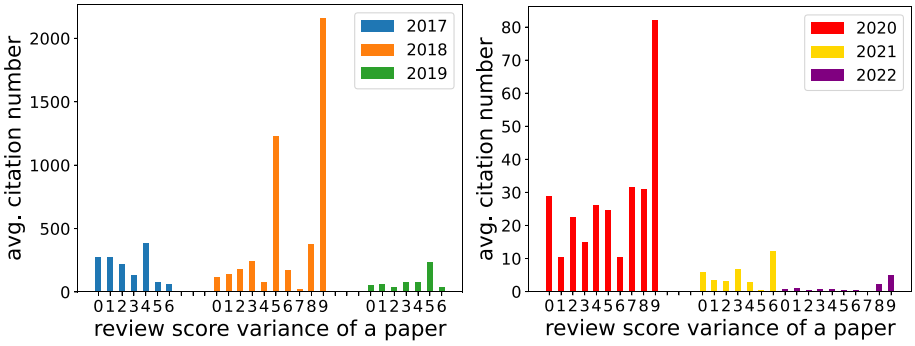


Figure 8 The distribution of citation numbers of individual papers, where the papers on the x-axis are sorted in the ascending order of their review scores



(a) 2017-2019 (b) 2020-2022

Figure 9 Review score variance of a paper vs. average citation number

Do highly cited papers gain more diverse review scores? We investigate the relationship between the variance of review scores of a submission and its citation number. We group papers according to their review score variances and calculate the average citation number of each group. Figure 9 shows the statistical results of the submissions of ICLR 2017-2022. We observe that the papers that have large number of citations are indeed more likely to gain diverse review scores. Note that a paper that has diverse review scores (big review score variance) does not necessarily have high review scores.

Causality analysis The causal analysis should meet two conditions: temporal precedence and correlation, accounting to [16]. For example, we cannot make a causal analysis between the reviewers' confidence level and "arXived submissions". Note that, we refer to the submissions that have been posed on arXiv before notification as "arXived submissions". Since we calculated their correlation coefficient at 0.0126, which is close to zero, it shows no correlation between them. Note that, as pointed out in [17], a correlation coefficient (r) of < 0.4 is often considered "weak". Correlation coefficients (r) of 0.4-0.7 as a moderate relationship, of 0.7-0.9 a strong or high relationship and > 0.9 as a "very high" relationship. In general, weak correlation is considered meaningless. Furthermore, we cannot analyze the influence of citation numbers on the review scores since they violate temporal precedence assumptions. In a word, temporal precedence and correlation are necessary and sufficient conditions for causal analysis.

In the article, we cannot conduct a causal analysis of the impact of review scores on citation numbers, since we calculated their correlation coefficient at 0.2448, which is weak correlation and does not satisfy the correlation condition.

3.5 Do submissions posted on arXiv have higher acceptance rate?

We found 2,761 submissions that have been posted on arXiv before accept/reject notification,[5] which account for about 23.17% of the total submissions. The arXiv versions are not anonymous, which bring unfairness to the double-blind review process. We refer to

[5] We compare paper creation date on arXiv with ICLR official notification date.

the submissions that have been posed on arXiv before notification as "arXived submissions". We investigate the acceptance rates of the arXived and non-arXived submissions. The acceptance rates of the arXived submissions in 2017, 2018, 2019, 2020, 2021, and 2022 are 59.33%, 62.39%, 45.36%, 30.48%, 44.28%, and 47.15% respectively. The acceptance rates of the non-arXived submissions in 2017, 2018, 2019, 2020, 2021, and 2022 are 45.88%, 41.23%, 26.37%, 17.22%, 20.07%, and 22.86%, respectively. We observe that the arXived submissions have significantly higher acceptance rate than the non-arXived submissions (48.16% vs. 28.94% on average).

We think the reason should be not only anonymity but also that the arXived ICLR submissions have higher quality. These arXived submissions might attract more feedbacks from colleagues, according to which the authors can improve their manuscripts. The arXived submissions might also be the rejected ones from other conferences and might have been improved according to the rejection reviews. We also observe that some arXived submissions are posted on arXiv one year before the submission deadline. Figure 10 shows the number of arXived submissions posted on arXiv by month, including both accepted ones and rejected ones. We can see that the papers posted on arXiv are more and more when approaching the submission deadline. There are also a large number of papers posted on arXiv between the submission date and the notification date. From the aspect of acceptance rate, we observe that the earlier the papers are posted on arXiv, the more likely they are accepted. In addition, the papers posted on arXiv after notification date have a higher acceptance rate. The reason might be that the authors cannot wait to share their research results after their papers are accepted.

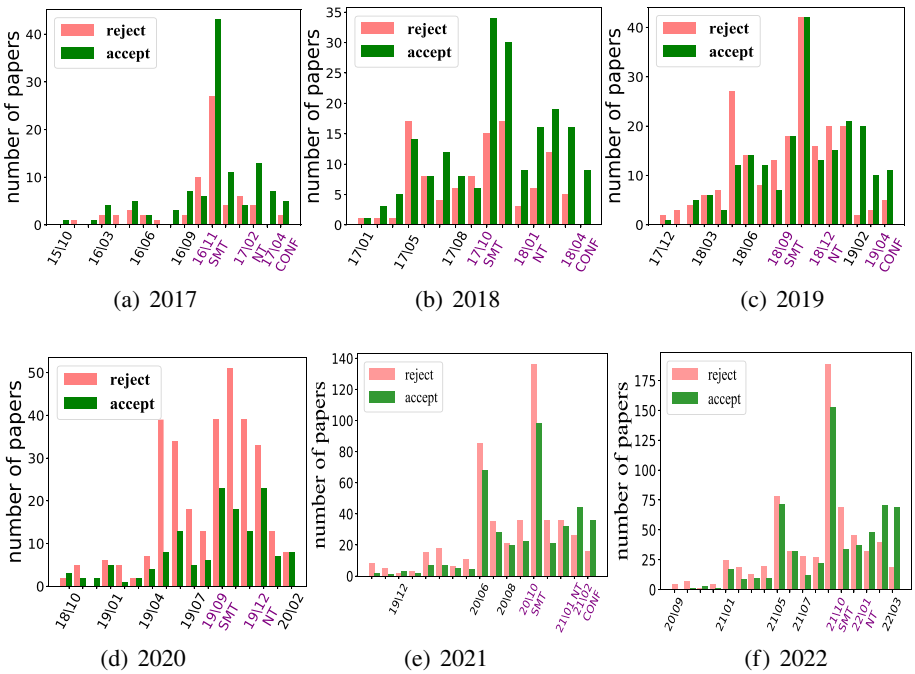


Figure 10 The review score distributions of different confidence level (conf) reviews

Causality analysis We observe that the arXived submissions have significantly higher acceptance rate than the non-arXived submissions. The "arXived submissions" and review scores meet the temporal precedence condition, and their correlation coefficient is 0.4799, which is a moderate relationship, thus we further determine whether causality exists. The process of causal analysis includes four steps: modeling, intervention, evaluation, and inference. In the modeling process, we use multivariate linear regression method to perform regression task on the ICLR reviews dataset, where the six evaluated parameters are the sentiment scores of the five review aspects and a reviewer confidence score, and the regression label is the review score. Each parameter is standardized to [-1,1]. To avoid randomness of model training, we launch 1000 times of training and obtain the average MSE (Mean Square Error) 0.24. Note that, the original model is the same as the model in Section 3.2. Then, we conduct interference experiments to analyze the causal relationship between "arXived submissions" and review scores. The intervention process removes "arXived submissions" factor and performs multiple times of model evaluation to obtain average MSE results. The average MSE of the interference model is higher than that of the original model (0.29 vs 0.24 on average MSE). Generally, in the absence of overfitting, the MSE value of "arXived submissions"-absence model will be larger than 0.24. The result indicates that there exists a causal relationship between "arXived submissions" and review scores.

3.6 How to write a rebuttal to boost the review score?

Rebuttal is commonly adopted in the paper review process and is widely used in peer review. We crawled ICLR reviews, rebuttals data, and the score changes after rebuttal. We collected 5790 (Reviews-Rebuttals) pairs from ICLR 2020. There are 623 papers with improved scores after rebuttal, accounting for about 10.76%. There is a note that the reviewer gives a rating score in {1, 3, 6, 8} for ICLR 2020, so the scores after rebuttal procedure are {2, 3, 5, 7}. The scores changed by 2, 3, 5 and 7, accounting for 3.61%, 6.41%, 0.71% and 0.02%, respectively. From the perspective of helping to receive papers, the effective rebuttal accounts for about 7.15% of the total (e.g., $\{3 \rightarrow 6\}$). We also count the sentence length and times of rebuttals. The average length of rebuttals is 2735.8 words. The average word length of papers' rebuttal that theirs scores changed is 4507.5. Obviously, the longer the rebuttal is, the more detailed the author answers the reviewer's questions. The reviewer will be able to better understand the paper's contribution and give appropriate recommendations. To the analysis of rebuttal times, the more rebuttal times, the easier it is to improve the review score (2.31 vs 1.31 on average). The reason behind this might be that multiple rebuttal times can make the reviewer understand your work well. It greatly increases the chances of improving scores.

All in all, rebuttal is an important phase to save your paper, so authors must pay much attention on the rebuttal phase. In this section, we mainly talk about these four aspects: question description, dataset description, model architecture, and experiments and results.

Question description and dataset description Each article has a review-rebuttal sequence. According to the change of score, we can divide it into five categories. Therefore, we transform the prediction problem into a classification problem. We formally state the problem as, given a pair of review-rebuttal. A review with x sentences $Rev1 = \left[s_1^1, s_1^2, ..., s_1^x\right]$ and corresponding rebuttal with y sentence $Reb1 = \left[s_2^1, s_2^2, ..., s_2^y\right]$. The goal is to predict the change in the review score. One review may correspond to multiple rebuttals, and the times

of rebuttal affect score changes. In order to make the data set have a unified format, we concatenate multiple rebuttals together. The Pearson correlation coefficient between rebuttal times and length is 0.76, which belongs to a strong correlation. Therefore, when we standardize the data set, it will not have a negative impact on the prediction results. We label five categories of score changes, and the mapping relationship between score change and label is {0: '0', 1: '2', 2: '3', 3: '5', 4: '7'}.

Model architecture This section introduces our architecture to predict the change of score after rebuttal. Figure 11 shows double BERT [18] (DBERT) architecture. Double BERTs are named DBERT. One BERT learns the review content, and another learns the rebuttal content. Inputs are encoded, producing two tuples of matrices (token, mask, sequence ids), one for each input. We use pre-trained BERT to generate token embedding. Further, these embeddings are fed as input to BiLSTM to generate sentence embedding. These encoded sentences are fed to the concatenate layer and concatenate together. After that, these encoded concatenate sentences are passed to the BiLSTM to encode review-rebuttal passages embedding. In addition, we introduce the attention mechanism, which can dynamically capture the relevant features from the review-rebuttal paragraph. Finally, the prediction results are output after being processed by the Dense layer. Next, we will describe the components of the framework in further detail.

Word embedding In order to extract the semantic information of review and rebuttal pairs, each sentence firstly is represented as a sequence of word embedding. As shown

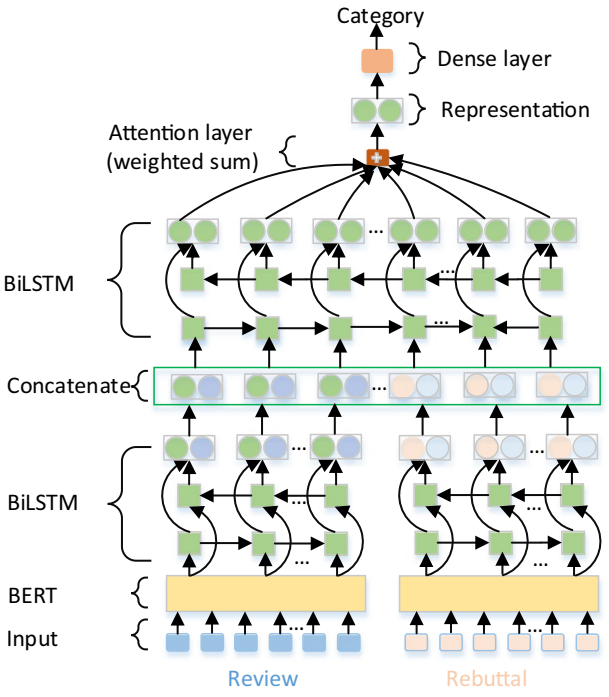


Figure 11 Overview of DBERT model architexture

in Figure 11, a review or rebuttal may contain multiple sentences. Take a sentence as an example to introduce the embedding process. Let's assume that a sentence has n-words. Our formal definition is as follows:

$$S = \left[W_0, W_1, \cdots, W_{n-1}\right] \tag{4}$$

where W_{n-1} represents the word of n-th with a serial number of n-1. In order to get a better embedding effect, we use the pre-training model (BERT) to generate word embedding. Each word is mapped to an embedding vector, and then we have $S_e = [e_0, e_1, \cdots, e_{n-1}]$. Where vector e_i represents the vector of i-th word with a dimension of d, $i \in (0, n-1)$. In this article, we set the dimension value d to 300. After that, we feed the token-level embedding into the BiLSTM network. In order to be able to learn sentence-level embedding.

BiLSTM layer The BiLSTM layer captures the output S_e from the previous layer. After that, two different direction LSTMs are trained on the same input sequence. We first define an LSTM procedure and the output vector of LSTM o_t can be expressed by the following equations:

$$i_t = \sigma\left(\omega_{ei} e_t + \omega_{hi} h_{t-1} + \omega_{ci} c_{t-1} + b_i\right) \tag{5}$$

$$f_t = \sigma\left(\omega_{ef} e_t + \omega_{hf} h_{t-1} + \omega_{cf} c_{t-1} + b_f\right) \tag{6}$$

$$c_t = f_t c_{t-1} + i_t tanh\left(\omega_{ec} e_t + \omega_{hc} h_{t-1} + b_c\right) \tag{7}$$

$$o_t = \sigma\left(\omega_{eo} e_t + \omega_{ho} h_{t-1} + \omega_{co} c_t + b_o\right) \tag{8}$$

$$h_t = o_t tanh(c_t) \tag{9}$$

Let $S_e = [e_0, e_1, \cdots, e_{n-1}]$ represent the input information of LSTM. Where σ is a sigmoid function; c, f, i, and o are the cell state, forget gate, input, and output, respectively; and all b are biases, $t \in (0, n-1)$. h_t is the hidden state output, ω is a weight matrix (e.g., ω_{eh} is a weight connecting input (e) to hidden layer (h)). However, LSTM only considers the influence of past information on embedding. In order to overcome this shortcoming, the concept of BiLSTM was proposed, which can consider the impact of surrounding information on the embedding. These two LSTM hidden layers have different directions, so they are named forward hidden layer and backward hidden layer. They are represented by h_t^f and h_t^b, respectively. The BiLSTM model is expressed with the following equations:

$$h_t^f = \tanh\left(\omega_{eh}^f e_t + \omega_{hh}^f h_{t-1}^f + b_h^f\right) \tag{10}$$

$$h_t^b = \tanh\left(\omega_{eh}^b e_t + \omega_{hh}^b h_{t+1}^b + b_h^b\right) \tag{11}$$

$$y_t = \omega_{hy}^f h_t^f + \omega_{hy}^b h_t^b + b_y \tag{12}$$

y_t is the combination of h_t^f and h_t^b. In our architecture, we get sentence-level embedding for review and rebuttal. Then, we use S_c to represent concatenated the generated review and the rebuttal sentence embedding. S_c can represent the embedding of a review-rebuttal pair. After that, we feed the review-rebuttal pair embedding into the BiLSTM-ATTENTION layer. In order to be able to learn paragraph-level embedding.

BiLSTM-ATTENTION layer BiLSTM is specialized for sequential modelling and can extract the temporal relationship of review-rebuttal pairs. The attention mechanism is to assign different weights to words to enhance understanding of the sentiment of the entire context. In the section, we use the attention mechanism to capture the correlation between review and rebuttal (e.g., question and response are consistent). The attention mechanism can focus on the features of the keywords to reduce the impact of non-keywords on the text sentiment, and it can speed up the convergence of the model. The workflow of BiLSTM-ATTENTION is described in detail below. First, let $A = [a1, a2, \cdots, an]$ represent the output vector of BiLSTM hidden layer. Secondly, finding the relevant vectors for each embedding in the sequence. The attention model is expressed as follows:

$$\alpha_{ki} = \frac{\exp(a_{ki})}{\sum_{j=i}^{Tx} \exp(a_{kj})} \tag{13}$$

$$a_{ki} = v \tanh\left(Wh_k + Uh_i + b\right) \tag{14}$$

$$C = \sum_{i=1}^{Tx} a_{ki} h_i \tag{15}$$

where $i, j, k \in n$, α_{ki} is the attention score of the i-th word in the k-th sentence. The bigger α_{ki} is, the more important the i-th word in the sentence is. W and U are trainable matrices. Finally, we represent the sentence vector C as a weighted sum of the word annotations. When we get vector C, we can feed it to the full connection layer for classification.

Experiments and results In our experiments, we split the annotated dataset of each aspect into training/validation/test sets (8:1:1) and use two different BERTs. They have the same composition but are trained with different inputs. The first one receives review content, while the other uses rebuttal descriptions. We initialize the learning rates, epoch, and batch size as 2×10^{-5}, 10, and 6, respectively. We compare DBERT with the state-of-the-art pretrain model e.g., BERT and BERT+BiLSTM to show its superority. Classification results are presented in Table 5. We can observe that the precision of double LSTM (DLSTM) is the lowest. We analyzed the reasons why DLSTM has low precision, because DLSTM applying one-hot encoding to words has negative influence. It adds a massive number of dimensions to the dataset, but there really isn't much information. We analyze the performance of the baseline BERT model and BERT+BiLSTM model. BERT+BiLSTM has

Table 5 Model prediction results

Models	DLSTM	BERT	BERT+BiLSTM	DBERT
Precision	0.76	0.87	0.88	0.89
Recall	0.87	0.87	0.88	0.89
F1-score	0.81	0.84	0.84	0.85

better performance because BiLSTM captures paragraph embedding of review-rebuttal better than BERT. Our model with double BERT mechanisms performs consistently better than both BERT and BERT+BiLSTM. Because BERT only takes a fixed-length text as its input, the maximum length is 512. Therefore, a single BERT model lose a lot of important information. Our method not only solves the sentence length limitation problem of the pre-training model, but also provides an idea for conversational reasoning in NLP. In general, our results illustrate that DBERT can detect the correlation of review-rebuttal and can help reviewers make an appropriate evaluation.

Ablation study We conduct extensive ablation studies on Review-Rebuttal datasets. We define three alternatives to study the impact of independently training strategy. Here, DBERT1 does not consider double BERT. DBERT2 does not consider BiLSTM layer. DBERT3 does not consider attention layer. For a fair comparison, all these variants adopt the same settings and the evaluation metric. The mean average precision (MAP) of the model DBERT, DBERT1, DBERT2, and DBERT3 are 0.89, 0.87, 0.88, and 0.88, respectively. We can see that full DBERT performs best compared with other alternatives. Removing each component results in slight relative performance degeneration. It reflects the effectiveness of each component of DBERT, and shows the mutual promotion of our method.

Causality analysis We find that there are some papers with improved scores after rebuttal. The "rebuttal" and review scores meet the temporal precedence condition, and their correlation coefficient is 0.6325, which is a moderate relationship, so we further determine whether causality exists. Then, we conduct interference experiments to analyze the causal relationship between rebuttal and review scores. The intervention process adds rebuttal factor and performs multiple times of model evaluation to obtain average MSE results. This result is smaller than that of the original model (0.14 vs 0.24 on average MSE). Generally speaking, in the absence of overfitting, the MSE value of rebuttal-presence model should be smaller than 0.24. The result indicates that there is a causal relationship between rebuttal and review scores.

4 Related work

There exist many interesting works related to peer-review analysis. We list several related works as follows.

Review decision prediction Kang et al. [2] predict the acceptance of a paper based on textual features and the score of each aspect in a review based on the paper and review contents. They also contribute to the community a publicly available peer review dataset for research purpose. Wang and Wan [3] investigate the task of automatically predicting the overall recommendation/decision and further identifying the sentences with positive and negative sentiment polarities from a peer review text written by a reviewer for a paper submission. DeepSentiPeer [5] takes into account the paper, the corresponding reviews, and review's polarity to predict the overall recommendation score.

AI support for peer-review system Anonymous peer review has been criticized for its lack of accountability, its possible bias, and its inconsistency, alongside other flaws. With the recent progress in AI research, many researches put great efforts in improving the peer-review system with the help of AI. Price and Flach [1] survey the various means of

computational support to the peer review system. The famous Toronto Paper Matching system [19] can achieve automated paper reviewer assignment. Mrowinski et al. [20] exploit evolutionary computation to improve editorial strategies in peer review. Roos et al. [21] propose a method for calibrating the ratings of potentially biased reviewers via a maximum likelihood estimation (MLE) approach. Stelmakh et al. [22] discuss biases due to demographics in single-blind peer review and study associated hypothesis testing problems. Nihar B. Shah et al. [23] survey a number of challenges in peer review, understand these issues and tradeoffs involved via insightful experiments, and discuss computational solutions proposed in the literature. Lindsay Fallon et al. [24] provide manuscript reviewers with recommendations and self-reflection questions for monitoring biases and promoting equity and social justice in the peer review process. Emaad Manzoor et al. [25] proposed a framework to nonparametrically estimate biases expressed in text.

Conflict of interest in the peer-review The increasing relationship between academic research and external industry has left research vulnerable to conflicts of interest. COI can undermine the integrity of scientific research and threaten public trust in scientific findings. Mecca et al. [26] quantitatively analyzed the conflict of interest from the researcher's perspective and proposed best practices for resolving the conflict of interest. Nowadays, Detecting conflicts of interest (COIs) is key for guaranteeing the fairness of a peer-review process. The authors in [27] develop a novel interactive system called PISTIS that assists the declaration process in a semi-automatic manner. Aleman-Meza et al. [28] develop a Semantic Web application that detects Conflict of Interest (COI) relationships among potential reviewers and authors of scientific papers. The authors in [29] study a graphical declaration system that visualizes the relationships of authors and reviewers based on a heterogeneous co-authorship network. With the help of the declarations, we attempt to detect the latent COIs automatically based on the meta-paths of a heterogeneous network. In peer review process, it is prohibitively expensive for PC chairs with thousands of reviews to manage to double-check the accuracy and completeness of these manual declarations. Nor can reviewers reliably catch unreported conflicts. CLOSET [30] is a data-driven scalable solution to address the aforementioned challenges. Review scores and reviews may have biases induced by undetected COI violations. These biases may bring uncertainty of analysis results.

Other interesting works of peer-review Birukou et al. [31] analyzed ten CS conferences and found low correlation between review scores and the impact of papers in terms of future number of citations. Gao et al. [32] predict after-rebuttal (i.e., final) scores from initial reviews and author responses. Their results suggest that a reviewer's final score is largely determined by her initial score and the distance to the other reviewers' initial scores. Li et al. [4] utilize peer review data for the citation count prediction task with a neural prediction model. Cormode [33] outlines the numerous ways in which an adversarial reviewer can criticize almost any paper, which inspires us a future work on how to identify the adversarial reviewers based on the open review data. Ivan Stelmakh's Blog [34] shares a lot of interesting findings: First, reviewers give lower scores once they are told that a paper is a resubmission. Second, there is no evidence of herding in the discussion phase of peer review. Third, A combination of the selection and mentoring mechanisms results in reviews of at least comparable and on some metrics even higher-rated quality as compared to the conventional pool of reviews. Nihar B.Shah et al. [35] analyzed the influence of reviewer and AC bid, reviewer assignment, different types of reviewers, rebuttals and discussions, distribution across subject areas in detail. Homanga Bharadhwaj et al. [36] provide an analysis

on whether there is a positive impact if his/hers paper is upload on arXiv before the submission deadline. They suggest that the paper arXived will have a higher acceptance rate. David Tran et al. [37] analyzed ICLR conferences and quantified reproducibility/randomness in review scores and acceptance decisions, and examined whether scores correlate with paper impact. Their results suggest that there exists strong institutional bias in accept/reject decisions, even after controlling for paper quality. They analyzed the influence of scores among gender, institution, scholar reputation in detail. The authors leveraged the framework to accurately detect these biases from the review text without having access to the review ratings. Ivan Stelmakh et al. [38] investigate if such a citation bias in peer review actually exists. Guneet Singh Kohli et al. [39] proposed model to extract arguement pair from peer review and rebuttal. Liying Cheng et al. [40] propose a multitask learning framework based on hierarchical LSTM networks to extract argument pairs from peer review and rebuttal.

In this paper, we investigate ICLR 2017-2022's submissions and reviews data on OpenReview and show more different interesting results, e.g., the effect of low confidence reviews, the sentiment analysis of review text on different aspects, the hierarchical relationships of different research fields, etc, which have not been studied before.

5 Conclusion

Since different reviewers have different interpretations of the scores, different reviewers may share the same views but give different review scores. The bias caused by the reviewers' different interpretations of scores may affect our analysis results. Similarly, the bias also exists in sentiment analysis based on review text. On the other hand, review scores and reviews may have biases induced by undetected COI violations, topic bias, etc. These biases may bring uncertainty of analysis results.

We perform deep analysis on the dataset including review texts collected from OpenReivew, the paper citation information collected from GoogleScholar, and the non-peer-reviewed papers from arXiv.org. All of these collected data are publicly available on Github, which will help other researchers identify novel research opportunities in this dataset. More importantly, we investigate the answers to several interesting questions regarding the peer-review process. We aim to provide hints to answer these questions quantitatively based on our analysis results. We believe that our results can potentially help writing a paper, reviewing it, and deciding about its acceptance.

Acknowledgements We thank the anonymous reviewers of APWEB 2021 for their encouraging and constructive comments and suggestions.

Author contributions Yanfeng Zhang and Gang Wang wrote the main manuscript text and Qi Peng prepared figures 2-4. Yanfeng Zhang contributed to the overall design of the study and some experimental designs. Gang Wang designed the model and analysed the data. Qi Peng crawled ICLR 2017-2020 data and analysed some data. Mingyan Zhang implements the production of some data sets. All authors reviewed the manuscript.

Funding This work was supported by the National Natural Science Foundation of China (62072082, U1811261, 62202088, U2241212), the Fundamental Research Funds for the Central Universities (N2216015, N2216012), and the Key R&D Program of Liaoning Province (2020JH2/10100037).

Data availability The data used in the paper is available at https://github.com/Seafoodair/Openreview/tree/master/data.

Code availability The codes are available at https://github.com/Seafoodair/Openreview.

Declarations

Ethics approval This article does not contain any studies involving human participants and/or animals by any of the authors.

Consent for publication All authors have read and agreed to the published version of the manuscript.

Consent to participate All authors have agreed to participate in the research described in this manuscript.

Conflict of interests The authors declare no conflict of interest.

Human and animal ethics Not applicable

References

1. Price, S., Flach, P.A.: Computational support for academic peer review: A perspective from artificial intelligence. Commun. ACM **60**(3), 70–79 (2017)
2. Kang, D., Ammar, W., Dalvi, B., van Zuylen, M., Kohlmeier, S., Hovy, E.H., Schwartz, R.: A dataset of peer reviews (peerread): Collection, insights and NLP applications. In: North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp 1647–1661 (2018)
3. Wang, K., Wan, X.: Sentiment analysis of peer review texts for scholarly papers. In: International ACM SIGIR Conference, pp 175–184 (2018)
4. Li, S., Zhao, W.X., Yin, E.J., Wen, J.: A neural citation count prediction model based on peer review text. In: Natural Language Processing, pp 4913–4923 (2019)
5. Ghosal, T., Verma, R., Ekbal, A., Bhattacharyya, P.: Deepsentipeer: Harnessing sentiment in review texts to recommend peer review decisions. In: Association for Computational Linguistics, pp 1120–1130 (2019)
6. Stelmakh, I., Shah, N.B., Singh, A.: On testing for biases in peer review. In: NeurIPS, pp 5287–5297 (2019)
7. He, H.: Some metadata for those curious about their #ICLR2020. https://twitter.com/cHHillee/status/1191823707100131329
8. Maaten, L.v.d., Hinton, G.: Visualizing data using t-sne. J. Mach. Learn. Res. **9**, 2579–2605 (2008)
9. Aslam, J.A., Pavlu, V.: Query hardness estimation using jensen-shannon divergence among multiple scoring functions. In: European Conference on Information Retrieval, pp 198–209 (2007)
10. Clark, K., Luong, M.-T., Le, Q.V., Manning, C.D.: Electra: Pre-training text encoders as discriminators rather than generators. In: International Conference on Learning Representations (2020)
11. Ito, T., Tsubouchi, K., Sakaji, H., Yamashita, T., Izumi, K.: Contextual sentiment neural network for document sentiment analysis. Data Sci. Eng. **5**(2), 180–192 (2020)
12. Gelman, A., Hill, J.: Causal Inference using Regression on the Treatment Variable. Analytical Methods for Social Research, pp 167–198. Cambridge University Press, Cambridge (2006)
13. Allison, P.D.: Multiple Regression: A Primer. Pine Forge Press, Pine Forge Press (1999)
14. jiqizhixin: Artificial-Intelligence-Terminology. https://github.com/jiqizhixin/Artificial-Intelligence-Terminology
15. Joe, H., Ward, J.: Hierarchical grouping to optimize an objective function. J. Am. Stat. Assoc. **58**(301), 236–244 (1963)
16. Trivedi: Causality or causal inference or conditions for causal inference. https://conceptshacked.com/causal-inference/ (2020)
17. Alsaqr, A.M.: Remarks on the use of pearson's and spearman's correlation coefficients in assessing relationships in ophthalmic data. Afr. Vision Eye Health **80** (1), 10 (2021)
18. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp 4171–4186 (2019)

19. Charlin, L., Zemel, R.: The toronto paper matching system: an automated paper-reviewer assignment system (2013)
20. Mrowinski, M.J., Fronczak, P., Fronczak, A., Ausloos, M., Nedic, O.: Artificial intelligence in peer review: How can evolutionary computation support journal editors? PloS ONE **12**(9), 0184711 (2017)
21. Roos, M., Rothe, J., Scheuermann, B.: How to calibrate the scores of biased reviewers by quadratic programming. In: AAAI Conference on Artificial Intelligence, pp 255–260 (2011)
22. Stelmakh, I., Shah, N., Singh, A.: On testing for biases in peer review. Adv. Neural Inf. Process. Syst. **32** (2019)
23. Shah, N.B.: An overview of challenges, experiments, and computational solutions in peer review. Communications of the ACM (2021)
24. Fallon, L., Grapin, S., Newman, D.S., Noltemeyer, A.: Promoting equity and social justice in the peer review process: Tips for reviewers. Sch. Psychol. Int. **43**(1), 12–17 (2022)
25. Manzoor, E., Shah, N.B.: Uncovering latent biases in text: Method and application to peer review. arXiv e-prints 2010 (2020)
26. Mecca, J.T., Gibson, C., Giorgini, V., Medeiros, K.E., Mumford, M.D., Connelly, S.: Researcher perspectives on conflicts of interest: A qualitative analysis of views from academia. Sci. Eng. Ethics **21**(4), 843–855 (2015)
27. Wu, S., U, L.H., Bhowmick, S.S., Gatterbauer, W.: Pistis: A conflict of interest declaration and detection system for peer review management. In: Proceedings of the 2018 International Conference on Management of Data, pp 1713–1716 (2018)
28. Aleman-Meza, B., Nagarajan, M., Ramakrishnan, C., Ding, L., Kolari, P., Sheth, A.P., Arpinar, I.B., Joshi, A., Finin, T.: Semantic analytics on social networks: experiences in addressing the problem of conflict of interest detection, pp 407–416 (2006)
29. Wu, S., U, L.H., Bhowmick, S.S., Gatterbauer, W.: Conflict of interest declaration and detection system in heterogeneous networks, pp 2383–2386 (2017)
30. CLOSET: ConfLict Of IntereSt DEtection & ManagemenT System. https://personal.ntu.edu.sg/assourav/research/DARE/closet.html
31. Birukou, A., Wakeling, J.R., Bartolini, C., Casati, F., Marchese, M., Mirylenka, K., Osman, N., Ragone, A., Sierra, C., Wassef, A.: Alternatives to peer review: novel approaches for research evaluation. Front. Comput. Neurosci. **5**, 56 (2011)
32. Gao, Y., Eger, S., Kuznetsov, I., Gurevych, I., Miyao, Y.: Does my rebuttal matter? insights from a major NLP conference. In: North American Chapter of the Association for Computational Linguistics, pp 1274–1290 (2019)
33. Cormode, G.: How not to review a paper: The tools and techniques of the adversarial reviewer. ACM SIGMOD Rec. **37**(4), 100–104 (2009)
34. Stelmakh, I.: Experiments with the ICML 2020 Peer-Review Process. https://blog.ml.cmu.edu/2020/12/01/icml2020exp/
35. Shah, N.B., Tabibian, B., Muandet, K., Guyon, I., Von Luxburg, U.: Design and analysis of the nips 2016 review process. J Mach. Learn. Res. **19**, 1–34 (2018)
36. Bharadhwaj, H., Turpin, D., Garg, A., Anderson, A.: De-anonymization of authors through arxiv submissions during double-blind review. arXiv:2007.00177 (2020)
37. Tran, D., Valtchanov, A., Ganapathy, K., Feng, R., Slud, E., Goldblum, M., Goldstein, T.: An open review of openreview: A critical analysis of the machine learning conference review process. arXiv:2010.05137 (2020)
38. Stelmakh, I., Rastogi, C., Liu, R., Chawla, S., Echenique, F., Shah, N.B.: Cite-seeing and reviewing: A study on citation bias in peer review. arXiv:2203.17239 (2022)
39. Kohli, G.S., Kaur, P., Singh, M., Ghosal, T., Rana, P.S.: Arguably@ ai debater-nlpcc 2021 task 3: Argument pair extraction from peer review and rebuttals. In: Natural Language Processing and Chinese Computing, pp 590–602 (2021)
40. Cheng, L., Bing, L., Yu, Q., Lu, W., Si, L.: Ape: argument pair extraction from peer review and rebuttal via multi-task learning. In: Empirical Methods in Natural Language Processing, pp 7000–7011 (2020)