What Have We Learned from OpenReview?

Gang Wang  
Northeastern University

Qi Peng  
Northeastern University

Yanfeng Zhang (✉️ zhangyf@mail.neu.edu.cn)  
Northeastern University

Mingyang Zhang  
Northeastern University

Research Article

Keywords: Peer review, OpenReview, Opinion divergence

Posted Date: April 25th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1558902/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.  
Read Full License
What Have We Learned from OpenReview?

Gang Wang\textsuperscript{1}, Qi Peng\textsuperscript{1†}, Yanfeng Zhang\textsuperscript{1,2*} and Mingyang Zhang\textsuperscript{1}

\textsuperscript{1}School of Computer Science and Engineering, Northeastern University, Shenyang, 110169, China.
\textsuperscript{2}Key Laboratory of Intelligent Computing in Medical Image, Ministry of Education, Shenyang, 110169, China.

*Corresponding author(s). E-mail(s): zhangyf@mail.neu.edu.cn;
Contributing authors: wanggangneu@stumail.neu.edu.cn; ffpengqi@stumail.neu.edu.cn; theremay@outlook.com;

\textsuperscript{†}These authors contributed equally to this work.

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 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\textsuperscript{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\textsuperscript{[1]–[5]} 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\textsuperscript{[6]}. 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 (Sec. 3.1)? b) Reviewers often evaluate a paper from multiple aspects, such as motivation, novelty, presentation, and experimental design. Which aspect has a

\textsuperscript{1}\url{https://openreview.net/}
decisive role in the review score (Sec. 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 more fine-grained cluster analysis. Given the fine-grained hierarchical clustering results, is there significant difference in the acceptance rate of different research fields (Sec. 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 inspired 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 (Sec. 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 (Sec. 3.5)? f) The rebuttal is an opportunity provided by the openreview platform 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 (Sec. 3.6)?

In this paper, we collect 11,915 (accepted and rejected) submissions and their 41,276 reviews from ICLR 2017-2022 venues 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

---

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

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.

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 length of reviews.

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.

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.

---

3These datasets and the source code for the analysis experiment are available at https://github.com/Seafoodair/Openreview/
Table 2 Statistics of different confidence level reviews

| Time       | 2017-2019 |          |          |          |          |
|------------|-----------|----------|----------|----------|----------|
| Level      | level1    | level2   | level3   | level4   | level5   |
| #reviews   | 74        | 455      | 2,330    | 4,612    | 1,600    |
| fraction   | 0.80%     | 5.01%    | 25.67%   | 50.81%   | 17.71%   |

(a) 2017-2019

| Time   | 2020 |          |          |          |          |
|--------|------|----------|----------|----------|----------|
| Level  | level1 | level2   | level3   | level4   | level5   |
| #reviews | 1,104 | 2,554    | 2,659    | 1,449    | —        |
| fraction | 14.22% | 32.89%   | 34.24%   | 18.66%   | —        |

(b) 2020

| Time       | 2021-2022 |          |          |          |          |
|------------|-----------|----------|----------|----------|----------|
| Level      | level1    | level2   | level3   | level4   | level5   |
| #reviews   | 110       | 1,404    | 7,265    | 12,362   | 3,476    |
| fraction   | 0.45%     | 5.70%    | 29.51%   | 50.22%   | 14.12%   |

(c) 2021-2022

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 [6]). 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.

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. Fig. 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
What Have We Learned from OpenReview?

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

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 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
What Have We Learned from OpenReview?

Fig. 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.

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

| Samples     | 2017-2019 | 2020   | 2021   | 2022   |
|-------------|-----------|--------|--------|--------|
| P-value     | $6.79 \times 10^{-7}$ | $2.45 \times 10^{-30}$ | $1.84 \times 10^{-20}$ | $4.67 \times 10^{-20}$ |
| Cohen’s $d$ | 0.21      | 0.26   | 0.34   | 0.32   |

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 [6]. 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
What Have We Learned from OpenReview?

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 $s^i_p$ 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:

$$DIS(l_i, l_j) = \sqrt{\sum_{p \in R_{l_i, l_j}} (s^i_p - s^j_p)^2}. \quad (1)$$

After computing the distance between each pair of groups, we can construct a distance matrix. According to the distance matrix, we use t-SNE [7] to plot the visualized layout of different groups of reviews of each year in Fig. 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 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).
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) [8], 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:

\[
MJS(l_1, \ldots, l_m) = \frac{1}{m} \sum_{i \in \{l_1, \ldots, l_m\}} \left( \frac{1}{R_{l_1, \ldots, l_m}} \sum_{p \in R_{l_1, \ldots, l_m}} \frac{s^i_p \cdot \log \left( \frac{s^i_p}{s^{[1,m]}_p} \right)}{s^i_p} \right),
\]

(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, \( s^i_p \) is paper \( p \)'s average review score of \( l_i \)-confidence reviews, and \( s^{[1,m]}_p \) 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 Fig. 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 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.

\[
VAR(l_1, \ldots, l_m) = \frac{1}{m} \left( \sum_{p \in R^n_{l_1, \ldots, l_m}} \text{var}(s^1_p, s^2_p, \ldots, s^n_p) \right),
\]

(3)

where \( R^n_{l_1, \ldots, l_m} \) is a set of paper IDs, where each paper concurrently has reviews with confidence levels \( l_1, \ldots, l_m \) and the number of reviews with confidence
What Have We Learned from OpenReview?

(a) 2017-2019

(b) 2020

(c) 2021

(d) 2022

Fig. 4 Average variance of different combinations of confidence levels

levels $l_1, \ldots, l_m$ is $n$ ($n \leq m$), and $\text{var}(s^1_p, s^2_p, \ldots, s^n_p)$ 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 Fig. 4. We do 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 Fig. 4 than in Fig. 3. In 2017-2019, it is surprising 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]).

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 rather 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
What Have We Learned from OpenReview?

Fig. 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.

Sentiment Analysis. Given these five datasets including the labeled data, we perform sentiment analysis for each aspect using a pretrained text model ELECTRA [9] 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[10] 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 (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. 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

---

4 All of the annotated data including manually annotated ones are publicly available at at https://github.com/Seafoodair/Openreview/.
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 Fig. 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 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 [11]. 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 [12] 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 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 [13] 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 [14] 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 number of clusters and plot a dendrogram to visualize the hierarchical clustering result as shown in Fig. 6.

From Fig. 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
What Have We Learned from OpenReview?

Fig. 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.

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

The citation number quantitatively indicates a paper’s impact. 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 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 Fig. 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 Fig. 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
What Have We Learned from OpenReview?

![Distribution of citation numbers of individual papers](image)

**Fig. 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.

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 excellent original papers are rejected by the non-expert reviewers but they can still attract great attentions after published on arXiv.

**Do Great Papers Gain More Diverse Review Scores?** People always have diverse opinions on the breakthrough works. Reviewer A thinks it novel and is happy to give higher scores, but reviewer B may think it too crazy or unrealistic and reject it. There might be a big debate between reviewers. But it is usually hard to reach consensus. 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. Fig. 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.
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, 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 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. Fig. 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.

---

5. We compare paper creation date on arXiv with ICLR official notification date.
rate. The reason might be that the authors cannot wait to share their research results after their papers are accepted.

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, for ICLR 2020, the reviewer gives a rating score in \{1, 3, 6, 8\}, so the changed scores 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 length and times of rebuttals. The average length of rebuttals is 2735.8 words. The average length of a score-changed rebuttal is 4507.5 words. 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 rebuttal multiple 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. Therefore, we propose a model
DBERT to predict the after-rebuttal scores according to author response. 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 pair of review-rebuttal, and there is a sequence. The change of scores, which can be divided 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 $Rev_1 = [s_1^1, s_2^1, ..., s_x^1]$ and corresponding rebuttal with y sentence $Reb_1 = [s_1^2, s_2^2, ..., s_y^2]$. 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. Fig.11 shows double BERT[15](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 in Fig 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:

$$S = [W_0, W_1, \cdots, W_{n-1}]$$ (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
What Have We Learned from OpenReview?

Fig. 11 Overview of DBERT model architecture.

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:

$$i_t = \sigma (\omega_{ei}e_t + \omega_{hi}h_{t-1} + \omega_{ci}c_{t-1} + b_i)$$  \hspace{1cm} (5)
$$f_t = \sigma (\omega_{ef}e_t + \omega_{hf}h_{t-1} + \omega_{cf}c_{t-1} + b_f)$$  \hspace{1cm} (6)
$$c_t = f_t c_{t-1} + i_t \tanh (\omega_{ec}e_t + \omega_{hc}h_{t-1} + b_c)$$  \hspace{1cm} (7)
$$o_t = \sigma (\omega_{eo}e_t + \omega_{ho}h_{t-1} + \omega_{co}c_t + b_o)$$  \hspace{1cm} (8)
$$h_t = o_t \tanh (c_t)$$  \hspace{1cm} (9)

Let $S_e = [e_0, e_1, \ldots, e_{n-1}]$ represent the input information of LSTM. Where $\sigma$ 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, $\omega$ is a weight matrix (e.g., $\omega_{eh}$ 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 embedding. These two LSTM hiding layers have different directions, so they are named forward hiding layer and backward hiding layer. They are denoted
by denotations $h_i^f$ and $h_i^b$, respectively. The BiLSTM model is implemented with the following equations:

$$h_i^f = \tanh \left( \omega_{eh} e_t + \omega_{hh} h_{i-1}^f + b^f \right)$$  \hspace{1cm} (10)$$

$$h_i^b = \tanh \left( \omega_{eh} e_t + \omega_{hh} h_{i+1}^b + b^b \right)$$  \hspace{1cm} (11)$$

$$y_t = \omega_{hy} h_i^f + \omega_{hy} h_i^b + b_y$$  \hspace{1cm} (12)$$

$y_t$ is the combination of $h_i^f$ and $h_i^b$. 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 implemented with the following equations:

$$\alpha_{ki} = \frac{\exp(a_{ki})}{\sum_{j=i}^{T_x} \exp(a_{kj})}$$  \hspace{1cm} (13)$$

$$a_{ki} = v \tanh (W h_k + U h_i + b)$$  \hspace{1cm} (14)$$

$$C = \sum_{i=1}^{T_x} a_{ki} h_i$$  \hspace{1cm} (15)$$

Where $i, j, k \in n$, $k$ represents k-th sentences, $i$ represents i-th word in the sentence and $T_x$ is the number of words in a sentence. $\alpha_{ki}$ is attention score. The bigger $\alpha_{ki}$ 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 used 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
What Have We Learned from OpenReview?

We initialize the learning rates, epoch, and batch size as $2 \times 10^{-5}$, 10, and 6, respectively. We compare DBERT with the state-of-the-art pre-train model e.g., BERT and BERT+BiLSTM to show its superiority. Classification results are presented in Table 4. We can observe that the accuracy of double LSTM (DLSTM) is the lowest. A detailed description of DLSTM is in the appendix E. We analyzed the reasons why DLSTM has low accuracy. Because DLSTM applying one-hot encoding to words has no good effect. 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 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 will 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.

We perform a deep analysis on the rebuttal phase. Combining the analysis results with our rebuttal experience, We put forward some effective suggestions for writing a good academic rebuttal. First of all, we should take the time to examine the reviewer’s comments carefully. Secondly, We can concisely and completely quote the core of the reviewer’s question or concern. And then, for each question, we should give a clear response, including relevant background, a detailed description, or an explanation of the paper’s point of view. We can’t just focus on the problem and ignore the intention of the problem. Next, we utilize the DBERT model to evaluate rebuttal content. When the evaluation results meet our expectations, we proceed to the next step. Otherwise, we modify the contents of the rebuttal. Finally, We acknowledge the reviewer’s efforts and hope to continue communicating.

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 [16] can achieve automated paper reviewer assignment. Mrowinski et al. [17] exploit evolutionary computation to improve editorial strategies in peer review. Roos et al. [18] propose a method for calibrating the ratings of potentially biased reviewers via a maximum likelihood estimation (MLE) approach. Stelmakh et al. [19] discuss biases due to demographics in single-blind peer review and study associated hypothesis testing problems. Nihar B. Shah et al. [20] 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. [21] 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. [22] proposed a framework to nonparametrically estimate biases expressed in text.

Other Interesting Works of Peer-Review. Birukou et al. [23] 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. [24] 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 [25] 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 [26] 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. [27] 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. [28] 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. [29] 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. [30] investigate if such a citation bias in peer review actually exists. Guneet Singh Kohli et al. [31] proposed model to extract argument pair from peer review and rebuttal. Liying Cheng et al. [32] 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

We perform deep analysis on the dataset including review texts collected from OpenReview, 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.

Declarations

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

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

Human and Animal Ethics Not applicable

Consent for Publication All authors have read and agreed to the published version of the manuscript.
Availability of Supporting Data 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.

Conflicts of Interest/Competing Interests The authors declare no conflict of interest.

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

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

Authors’ 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.

Authors’ Information Gang Wang received the MS degree in software engineering from Liaoning University, China, in 2019. He is currently working toward the PhD degree in computer science at Northeastern University, China. His research interests include image recognition, blockchain and data mining.

Peng Qi obtained his Bachelor of engineering degree from Henan University in 2020. At present, He is studying at Northeast University (Shenyang, China) in the stage of master’s degree, and his main research direction is big data mining and analysis.

Yanfeng Zhang received the PhD degree in computer science from Northeastern University, China, in 2012. He is currently a professor with Northeastern University, China. His research consists of distributed systems and big data processing. He has published many papers in the above areas. His paper in SoCC 2011 was honored with “Paper of Distinction”.

Mingyang Zhang received her Bachelor degree in computer science from Northeastern University, China. Her main research interests include data mining and machine learning.

References

[1] Price, S., Flach, P.A.: Computational support for academic peer review: A perspective from artificial intelligence. Communications of the ACM 60(3), 70–79 (2017)
What Have We Learned from OpenReview?

[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] He, H.: Some Metadata for Those Curious About Their #ICLR2020. https://twitter.com/cHHillee/status/1191823707100131329

[7] Maaten, L.v.d., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research 9(Nov), 2579–2605 (2008)

[8] 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)

[9] 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)

[10] Ito, T., Tsubouchi, K., Sakaji, H., Yamashita, T., Izumi, K.: Contextual sentiment neural network for document sentiment analysis. Data Science and Engineering 5(2), 180–192 (2020)

[11] Gelman, A., Hill, J.: Causal inference using regression on the treatment variable. Analytical Methods for Social Research, pp. 167–198, Cambridge University Press (2006)

[12] Allison, P.D.: Multiple Regression: A Primer. Pine Forge Press, Pine Forge Press (1999)

[13] jiqizhixin: Artificial-Intelligence-Terminology. https://github.com/jiqizhixin/Artificial-Intelligence-Terminology

[14] Joe, H., Ward, J.: Hierarchical grouping to optimize an objective function. Journal of the American Statistical Association 58(301), 236–244 (1963)
[15] 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)

[16] Charlin, L., Zemel, R.: The toronto paper matching system: an automated paper-reviewer assignment system (2013)

[17] 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)

[18] 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)

[19] Stelmakh, I., Shah, N., Singh, A.: On testing for biases in peer review. Advances in Neural Information Processing Systems 32 (2019)

[20] Shah, N.B.: An overview of challenges, experiments, and computational solutions in peer review. Communications of the ACM (2021)

[21] Fallon, L., Grapin, S., Newman, D.S., Noltemeyer, A.: Promoting equity and social justice in the peer review process: Tips for reviewers. School Psychology International 43(1), 12–17 (2022)

[22] Manzoor, E., Shah, N.B.: Uncovering latent biases in text: Method and application to peer review. arXiv e-prints, 2010 (2020)

[23] 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. Frontiers in computational neuroscience 5, 56 (2011)

[24] 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)

[25] Cormode, G.: How not to review a paper: The tools and techniques of the adversarial reviewer. ACM SIGMOD Record 37(4), 100–104 (2009)

[26] Stelmakh, I.: Experiments with the ICML 2020 Peer-Review Process. https://blog.ml.cmu.edu/2020/12/01/icml2020exp/

[27] Shah, N.B., Tabibian, B., Muandet, K., Guyon, I., Von Luxburg, U.: Design and analysis of the nips 2016 review process. Journal of machine
What Have We Learned from OpenReview?

learning research 19, 1–34 (2018)

[28] Bharadhwaj, H., Turpin, D., Garg, A., Anderson, A.: De-anonymization of authors through arxiv submissions during double-blind review. arXiv e-prints abs/2007.00177 (2020)

[29] 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 preprint arXiv:2010.05137 (2020)

[30] 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 preprint arXiv:2203.17239 (2022)

[31] Kohli, G.S., Kaur, P., Singh, M., Ghosal, T., Rana, P.S.: Arguably@ ai debater-nlpc 2021 task 3: Argument pair extraction from peer review and rebuttals. In: Natural Language Processing and Chinese Computing, pp. 590–602 (2021)

[32] 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)

A More Results on the Impact of Non-Expert reviewers

We show the additional results to Section 3.1 of the main manuscript in the following figures.

![Fig. 12](image_url) The number of submissions is increasing extensively from 2017 to 2020. This accordingly leads to heavy demand for reviewer volunteers, and at the same time leads to large number of low confidence reviews which will be shown in Fig. 13.
What Have We Learned from OpenReview?

Fig. 13 The numbers of high-confidence reviews and low-confidence reviews. For 2017-2019, reviews with confidence level 3, 4, and 5 are considered as high-confidence reviews, and reviews with confidence level 1 and 2 are considered as low-confidence reviews. For 2020, reviews with confidence level 3 and 4 are considered as high-confidence reviews, and reviews with confidence level 1 and 2 are considered as low-confidence reviews.

Fig. 14 The average review score of high-confidence reviews, low-confidence reviews, and mixed total. The low-confidence reviewer tend to be more tolerant because they are not confident about their decision, while high-confidence reviewer tend to be more tough and rigorous because they are confident in the weakness they identified.

B The Divergence between Different Confidence Level Reviews

We show more detailed results on the divergence between different confidence level reviews.
Fig. 15 The distribution of different reviewer scores of different confidence level reviews of 2017-2020.

Table 5 Euclidean distance between different confidence level reviews (ICLR 2017-2019)

|       | conf1 | conf2 | conf3 | conf4 | conf5 |
|-------|-------|-------|-------|-------|-------|
| num   | 74    | 455   | 2330  | 4612  | 1600  |
| fra(%)| 0.80  | 5.01  | 25.67 | 50.81 | 17.71 |
| conf1 | 2.45  | 6.78  | 10.63 | 11.75 | 5.83  |
| conf2 | 6.78  | 14.63 | 25.90 | 34.23 | 20.69 |
| conf3 | 10.63 | 25.90 | 34.23 | 75.02 | 45.40 |
| conf4 | 11.75 | 34.23 | 75.02 | 109.05| 69.92 |
| conf5 | 5.83  | 20.69 | 45.40 | 69.92 | 49.72 |

B.1 Covariance and Pearson Coefficient Between Different Confidence-Leveled Reviews

C The Impact of Low Confidence Reviews to Acceptance Rate

Suppose that we trust the opinions of reviewers and that reviewers of papers are distributed uniformly among papers of different levels. We divide reviewers into two categories: professional and non-professional. A paper is reviewed by at least three reviewers. Therefore, it can be divided into three cases. First, one situation is when all reviewer levels are non-professional. Second, all reviewer levels are professional. Finally, professional and non-professional
Fig. 16 Review score vs. length of review (i.e., number of words). The reviews with the same review score are grouped together to compute average review length. The reviews with higher score are likely to be short. Great paper is so good that it is not necessary to give more details about the reason for it.

Table 6 Euclidean distance between different confidence level reviews (ICLR 2020)

|         | conf1 | conf2 | conf3 | conf4 |
|---------|-------|-------|-------|-------|
| num     | 1104  | 2554  | 2659  | 1449  |
| fra(%)  | 14.22 | 32.89 | 34.24 | 18.66 |
| conf1   | 54.62 | 69.81 | 62.04 | 48.64 |
| conf2   | 69.81 | 106.22| 101.55| 76.71 |
| conf3   | 62.04 | 101.90| 101.55| 79.95 |
| conf4   | **48.64** | 76.71 | 79.95 | **59.03** |

reviewers jointly review a paper. The scoring variance can indicate that there is a difference between different confidence level reviews but sometimes does not. For example, the two review scores are 2 and 4, and the two review scores are 4 and 6. The variance between them is the same. The difference between non-professionals and professionals cannot be explained, and if the variance is used, the paper’s score cannot be related to the reviewer’s professionalism. Therefore, we use positive and negative differences when analyzing different levels of reviewers.

Positive and negative difference: The calculation method of positive and negative difference is as follows. For example, an article is assigned three reviewers, a professional reviewer is divided into 4, and two non-professional reviewers are divided into 6 and 5. Then the positive and negative difference is (4)/1-(6+5)/2.

The reason for using positive and negative differences is that the average score of a paper can measure whether a paper is accepted or not, which is obvious. However, sometimes the reason for the high average may be highly professional scores, low non-professional scores, low professional scores, high non-professional scores, or high professional and non-professional reviewers. At this time, you need to use an indicator to see if the professional reviewer’s scoring plays a greater role. That is, if the average reviewer is equal, whether a high
Fig. 17 Acceptance rate vs. length of paper title (i.e., number of characters). The papers are grouped according to the length of title, and the average acceptance rate of each group is depicted. The papers with a long title are likely to be accepted.

professional reviewer’s scoring is more conducive to the paper’s admission. In other words, whether the scores of professional reviewers have a greater impact on the results of papers when the scores are quite different. This is why the design difference between positive and negative is used to judge the scoring differences between professional and non-professional reviewers. It is important to note that when all reviewers of a paper are professionals or non-professionals, the positive and negative difference is zero. Because the difference between positive and negative represents the differences between professional and non-professionals in giving a thesis. When the reviewers are all professionals or non-professionals, this is no longer a condition for calculating the positive and negative difference, so the positive and negative difference is set to 0 at this time.

D Text Sentiment Analysis Details

This is an example taken randomly from the 2020 review. Break this review into periods. It is divided into 8 sentences. First, match these sentences with words in 5 dimensions. The match was successful, for yes. No match is no. For the automatic labelling method, one method is to match angles and then perform emotional word matching. Another method is to use emotional words
Table 7 AVG review score variance of different combinations of different confidence level reviews (ICLR 2017-2019)

|     | avg var | paper num | review num |
|-----|---------|-----------|------------|
| (1) | no value| 0         | 0          |
| (2) | 0.39    | 4         | 12         |
| (3) | 0.70    | 84        | 252        |
| (4) | 0.80    | 408       | 1224       |
| (5) | 1.24    | 34        | 102        |
| (1,2)| 0.11    | 2         | 6          |
| (1,3)| 0.44    | 12        | 36         |
| (1,4)| 0.56    | 12        | 36         |
| (1,5)| 0.89    | 2         | 6          |
| (2,3)| 0.66    | 42        | 126        |
| (2,4)| 0.71    | 103       | 309        |
| (2,5)| 1.25    | 14        | 42         |
| (3,4)| 0.76    | 817       | 2451       |
| (3,5)| 0.94    | 128       | 384        |
| (4,5)| 0.91    | 573       | 1719       |
| (1,2,3)| 0.78   | 6         | 18         |
| (1,2,4)| 1.30   | 6         | 18         |
| (1,2,5)| 0.22   | 1         | 3          |
| (1,3,4)| 0.94   | 16        | 48         |
| (1,3,5)| 0.89   | 5         | 15         |
| (1,4,5)| 0.83   | 7         | 21         |
| (2,3,4)| 0.90   | 115       | 345        |
| (2,3,5)| 0.76   | 30        | 90         |
| (2,4,5)| 0.89   | 43        | 129        |
| (3,4,5)| 0.91   | 340       | 1020       |
| (all)| 0.84   |           |            |

first. For example, pros and cons in this paragraph. Think of a positive expression between pros and cons. After cons is a negative expression. These two methods can reduce the workload of manual labelling. In the second column, predictions are made for sentences where the human annotation and angle match. The results showed as expected.

Here we provide a typical review sample that contains “pros and cons”:

pros:

1. the proposed method shows the ability to learn the nested distributions with the help of hierarchical structure information. it is a reasonable way to model the general multimodal i2i. I think the authors work in the correct direction.

2. to model the partial order relation in the hierarchy, the authors borrow the thresholded divergence technique from the natural language processing field (athiwaratkun & wilson (2018)) and use the kl divergence.

cons:


Table 8  AVG review score variance of different combinations of different confidence level reviews (ICLR 2020)

|             | avg var | paper num | review num |
|-------------|---------|-----------|------------|
| (1)         | 1.78    | 23        | 69         |
| (2)         | 2.19    | 101       | 303        |
| (3)         | 1.83    | 118       | 354        |
| (4)         | 1.41    | 30        | 90         |
| (1,2)       | 2.30    | 168       | 504        |
| (1,3)       | 1.81    | 144       | 432        |
| (1,4)       | 1.67    | 47        | 141        |
| (2,3)       | 1.82    | 527       | 1581       |
| (2,4)       | 1.95    | 193       | 579        |
| (3,4)       | 2.05    | 258       | 774        |
| (1,2,3)     | 1.81    | 176       | 528        |
| (1,2,4)     | 1.75    | 100       | 300        |
| (1,3,4)     | 1.86    | 93        | 279        |
| (2,3,4)     | 2.03    | 310       | 930        |

Table 9  MJS-divergence of different combinations of different confidence level reviews (ICLR 2017-2019)

|             | MJS    | paper num | review num |
|-------------|--------|-----------|------------|
| (1,2)       | 0.10   | 18        | 38         |
| (1,3)       | 0.07   | 44        | 103        |
| (1,4)       | 0.08   | 48        | 115        |
| (1,5)       | 0.04   | 16        | 35         |
| (2,3)       | 0.08   | 215       | 492        |
| (2,4)       | 0.11   | 294       | 726        |
| (2,5)       | 0.14   | 101       | 224        |
| (3,4)       | 0.08   | 1374      | 3700       |
| (3,5)       | 0.10   | 537       | 1230       |
| (4,5)       | 0.08   | 1027      | 2717       |
| (1,2,3)     | 0.09   | 8         | 25         |
| (1,2,4)     | 0.19   | 8         | 25         |
| (1,2,5)     | 0.08   | 2         | 7          |
| (1,3,4)     | 0.13   | 20        | 64         |
| (1,3,5)     | 0.07   | 6         | 19         |
| (1,4,5)     | 0.13   | 7         | 21         |
| (2,3,4)     | 0.13   | 130       | 411        |
| (2,3,5)     | 0.12   | 38        | 120        |
| (2,4,5)     | 0.16   | 54        | 180        |
| (3,4,5)     | 0.13   | 366       | 1134       |

1. some figures are hard to understand without looking at the text. for example, in figure 1, the caption does not explain the figure well. what does each image, the order, and the different sizes mean? as to figure 3, the words “top left image”, “right purple arrows” are a bit confusing.

2. the “coarse to fine conditional translation” section describes the conditional translation in the shallow layers. i suggest mentioning it in previous sections for easy understanding.

3. as to the t-sne visualization in figure 9, different methods seem to use different n-d to 2-d mapping functions. this may lead to an unfair comparison.
Table 10 MJS-divergence of different combinations of different confidence level reviews (ICLR 2020)

|        | MJS | paper num | review num |
|--------|-----|-----------|------------|
| (1,2)  | 0.28| 535       | 1300       |
| (1,3)  | 0.27| 469       | 1110       |
| (1,4)  | 0.30| 298       | 675        |
| (2,3)  | 0.26| 1139      | 2908       |
| (2,4)  | 0.27| 693       | 1634       |
| (3,4)  | 0.28| 744       | 1787       |
| (1,2,3)| 0.34| 213       | 660        |
| (1,2,4)| 0.34| 138       | 434        |
| (1,3,4)| 0.38| 122       | 380        |
| (2,3,4)| 0.42| 364       | 1132       |

Table 11 Covariance (up) and Pearson coefficient (down) between different confidence level reviews (ICLR 2017-2019).

|        | conf1 | conf2 | conf3 | conf4 | conf5 |
|--------|-------|-------|-------|-------|-------|
| num    | 74    | 455   | 2330  | 4612  | 1600  |
| fra(%) | 0.80  | 0.45  | 0.33  | 0.48  | 1.22  |
| conf1  | 0.42  | 0.77  | 0.46  | 0.77  | 0.75  |
| conf2  | 0.45  | 0.68  | 0.36  | 0.43  | 0.47  |
| conf3  | 0.46  | 0.79  | 0.38  | 0.38  | 0.36  |
| conf4  | 0.95  | 0.86  | 0.44  | 0.38  | 0.36  |
| conf5  | 1.22  | 1.41  | 1.10  | 1.41  | 1.67  |

Table 12 Covariance (up) and Pearson coefficient (down) between different confidence level reviews (ICLR 2020).

|        | conf1 | conf2 | conf3 | conf4 |
|--------|-------|-------|-------|-------|
| num    | 1104  | 2554  | 2659  | 1449  |
| fra(%) | 14.22 | 32.89 | 34.24 | 18.66 |
| conf1  | 1.03  |       |       |       |
|        | 1.41  | 1.90  | 1.80  |       |
|        | 0.32  | 0.40  | 0.38  |       |
| conf2  | 1.41  | 1.50  | 1.86  | 2.05  |
|        | 0.32  | 0.32  | 0.40  | 0.41  |
| conf3  | 1.90  | 1.86  | 2.23  | 1.93  |
|        | 0.40  | 0.40  | 0.46  | 0.39  |
| conf4  | 1.80  | 2.05  | 1.90  | 1.87  |
|        | 0.38  | 0.41  | 0.39  | 0.41  |

Suggestions: 1. the authors use the pre-trained classification network vgg for feature extraction and then train dedicated translators based on these features. I wonder if the authors also tried finetuning vgg on the two domains or training an auto-encoder on the two domains. The domain-specific knowledge may help to improve the results and alleviate the limitations presented in the
| id | sentence |
|----|----------|
| 1  | It defines the samples whose average probability on assigned label in recent q iterations is largest among all labels as memorized samples, in the sense of the network memorize these samples. |
| 2  | Then authors proposed two stage method which firstly early-stops at minimum validation error (or memorized rate), and then trains on maximal safe set that gathers memorized samples. |
| 3  | The experiments compared several state-of-art approaches and showed that the proposed method benefits from early-stopping and safe set. |
| 4  | Authors also showed that the prestopping idea can also be used to improve other approaches. |
| 5  | Pros: The proposed method achieves better performance than state-of-art methods. Authors have good experiments which evaluate on multiple datasets and algorithms. |
| 6  | Authors also investigate the relation between model complexity and performance of co-teaching+ |
| 7  | Cons: Many recent papers indicate the “error-prone period”, authors should include related works about early-stopping on label noise training. |
| 8  | Although the method achieves good performance, since the idea is a bit straightforward especially after exploring above papers, I am slightly worried about novelty of the ideas. |

Fig. 18 The impact of low confidence reviews to the acceptance rate

We perform sentiment analysis for each aspect using a pre-trained text model ELECTRA [9]. The hyper-parameters of ELECTRA are listed as follows.
Table 13 Hyper-parameters of ELECTRA

|      | novelty | motivation | experiment | related work | presentation |
|------|---------|------------|------------|--------------|--------------|
| epoch|    7    |     7      |     3      |    10        |    10        |
| batchsize | 64      |   64       |   64       |   64         |   64         |
| maxlen|   256   |   256      |   256      |   256        |   256        |

E Model Architecture of DLSTM

Inspired by transfer learning, We propose the DLSTM method to solve the problem of few-shot learning. DLSTM’s model consists of two parallel LSTMs. As shown in Fig 19. The lstm in the red dotted box on the left uses the (review, original score) data to train a model. Review is the training data, and the original score is the label. The higher the model’s accuracy, the better the embedding performance of the review. We trained another lstm model with data (rebuttal, final score). The rebuttal is part of the training data, and the final score is the label. To train the review-rebuttal in pairs, we concatenate two lstms. The review embedding is generated by the model trained on the left. Rebuttal’s embedding is generated by the model being trained on the right. We splice the embedding of review and rebuttal (The red arrow in the figure indicates splicing). Take the embedding of the review-rebuttal as the training data and the final score as the label. The advantage of this model is that two lstms can train unequal amounts of data. Due to the small amount of data for
review-rebuttal pairs. We can use review data from other years to train the review model.