In Peer Review (Should) We Trust? A Meta Analysis of ICLR 2017

Atahan Ozer
Matrikelnummer 6317973
ataoz_hotmail.com

Yavuz Durmazkeser
Matrikelnummer 6401034
yavuz.durmazkeser@hotmail.com

Abstract

Peer reviewing holds an important role in the scientific publication process. In this project, we investigate whether peer-reviewing scores show any sign of future impact. To achieve that we first compare the key features of accepted and rejected papers. Then, we cluster the papers by using their abstract embeddings to check their future impact while being agnostic to topic popularity. After obtaining the embeddings, we compare the correlation between citation and review scores of the same cluster papers. Finally, we report the statistical properties of our findings.

1 Introduction

The modern scientific process is heavily dependent on peer reviewing. The role of peer reviewing on the quality and the rigor of scientific publication can not be articulated yet with a growing number of journals and conferences, the burden of peer reviewing is increasing enormously. This can be especially observed in the area of machine learning. For example, one of the premier conferences in machine learning International Conference on Learning Representations(ICLR) got 427 submissions in 2017 but got 4881 submissions in 2023. The increase is nearly tenfold and this trend can be also observed with other top-tier conferences. At this scale with the time limitations of conferences, a drop in the quality of the peer-reviewing process might be inevitable. An investigation of this drop could be challenging because, without significant human labor, it is not possible to read all the papers and reviews for quality purposes. Furthermore, even if there is a control mechanism over peer reviewing, that mechanism itself requires another control mechanism hence it creates an endless loop.

In this project as another approach to quality control, we try to examine whether there is a correlation and alignment between the reviews and the scientific impact. As a general metric, we use citation score for the scientific impact. However, using this metric requires an amount of time to be evaluated. For that purpose, we select ICLR 2017 as our dataset and we summarize our work as follows: ***(i)*** We conduct an exploratory data analysis of the data we collected and cleaned. ***(ii)*** We cluster the papers using their abstract so that we make them agnostic to sub-field citation changes to use in our further analysis. ***(iii)*** Finally, we report our statistical findings over the analysis of accepted and rejected papers.

2 Exploratory Data analysis

2.1 Data Collection and Preprocessing

Our literature search showed us there exists a work PeerRead [3] that already collected the ICLR 2017 from OpenReview. These reviews include the abstract of the paper, reviews from official reviewers, recommendation scores from official reviewers, scores confidence of reviewers, comments from area chairs, and the acceptance status of papers. However, this dataset does not include citations. To solve that we benefit from SemanticScholar API. After completing the data acquisition, we cleaned the

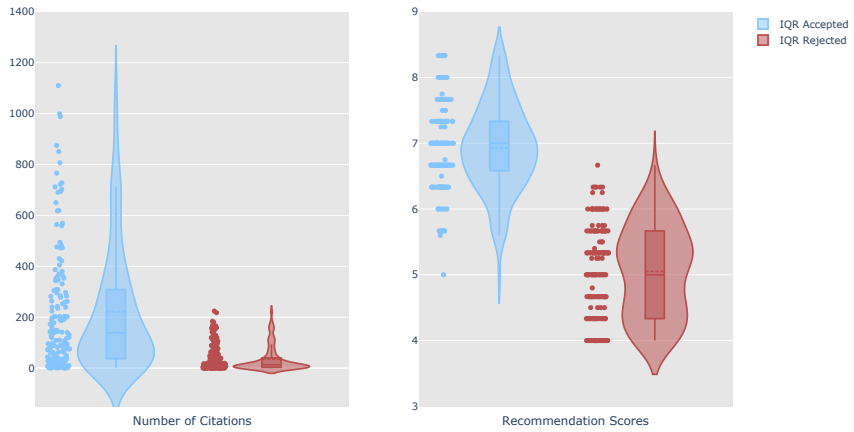

Figure 1: Distributions of Number of citations and recommendations scores for IQR accepted/rejected papers. The dashed line represents the mean and the line represents the median in box plots

raw data by removing the duplicate reviews and scores. We noticed that SemanticScholar API can not find 32 articles in its database hence we had to provide citation information for those papers manually. In total there are 427 submitted papers which 172 of them were accepted and 255 of them were rejected. Recommendation scores take place between 0 and 10, we discard papers that have lower recommendation points than 4 because they are mostly rejected due to simple quality measures and should not be mixed with other papers. Finally, we exclude outlier papers in terms of citation numbers for the soundness of analyses by using the 1.5 interquartile range (IQR) method for accepted and rejected papers separately. [4].

2.2 First Insights

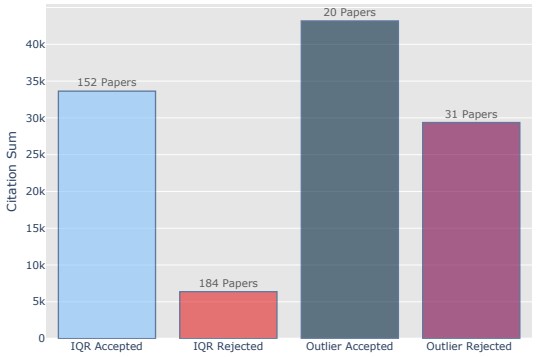

Figure 2: Total citation comparison of outliers

Distribution of citation and scores. Figure 2 clearly shows that there are significant differences between the distributions of accepted and rejected papers in terms of both the number of citations and recommendation scores. To start with, IQR-accepted papers have a median citation of 139, on the other hand, rejected papers only have 13 citations. The difference is nearly 10 folds and the recommendation scores are also far from each other with medians 7 and 5. Another difference that we realize while comparing citations and recommendation scores as in Figure 3, IQR accepted papers are highly populated between 6.5 and 7.5 scores and this density reaches up to 200 citations. In contrast, IQR-rejected papers are more uniformly distributed between 4.0 and 6 while reaching a much lower citation number.

The effect of outliers. As can be seen from Figure **??**, discarded outlier papers have significant total citations. The rejected 31 papers have almost the same citation as the IQR-accepted 152 papers. A natural question that can occur with this information is whether these 31 outlier papers might be borderline rejections due to the randomness of the peer-reviewing process. According to the data,

Table 1: Nearest Neighbors of "Dialogue Learning With Human-in-the-Loop"

Rank	Title
1	Learning through Dialogue Interactions by Asking Questions
2	Third Person Imitation Learning
3	Batch Policy Gradient Methods for Improving Neural Conversation Models
4	Learning to Perform Physics Experiments via Deep Reinforcement Learning
5	Multi-Agent Cooperation and the Emergence of (Natural) Language

rejected outliers have a median of 5.66, moreover, all rejected papers except those that are lower than point 4, have a median of 5.25 and there are 9 accepted papers between 5.0 and 5.66. Therefore, We can articulate that the upper quantile of the rejected outliers were actually borderline rejections. On the other hand, both all accepted papers and accepted outliers have a median of 7. Hence, accepted outliers do not follow a pattern in terms of recommendation scores. There are 40 rejected papers that have scores lower than 4 and they have a median citation of 3 and all rejected papers that have a higher score than 4 have a median citation of 17. The amount of difference justifies dropping these papers in regard to citation impact.

2.3 Clustering Papers

To measure the impact of a paper, we use the citation count in the following sections. However, we believe a direct comparison of this metric would be unfair because different topics of Machine Learning have different popularity. We aim to solve this problem by separating papers into several groups based on their fields.

Due to lack of previously defined categories, we benefit from some NLP methods. For each paper, we generate an embedding using its abstract and then create clusters using the k-means algorithm. We use various methods for computing the embeddings. The first one is counting words that exist in the abstract. Before counting, we remove stop-words and stem the remaining words. We also apply Principal component analysis to decrease 3848-dimensional embeddings to 300 dimensions. Another method we use is FastText [2]. We compute a 300-dimensional vector of every word in the abstract and then average them. Finally, we use BERT [1] embeddings to compute vectors of 768 dimensions and average them for each token. We have divided every vector mentioned above by its dimension before clustering. We present results for 1, 2, 5 and 10 clusters. We have chosen cluster counts after visualizing loss values for different k.

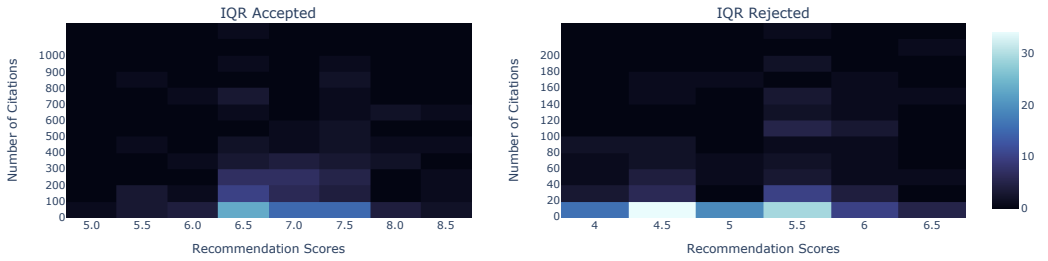

Figure 3: 2D histogram of IQR Accepted/Rejected Papers with citation and recommendation scores. For both histograms X axis is evenly spaced in terms of recommendation scores and the same bin size is used in Y-Axis.

In Table 1, we present the nearest neighbors of a paper in our dataset, calculated using BERT embeddings. The main paper, "Dialogue Learning With Human-in-the-Loop", develops a natural language model using reinforcement learning. We see that our method can retrieve papers with similar topics.

Table 2: Pearson Correlation Results. Numbers show correlation coefficients. Corresponding p-values are in parentheses.

Num Clusters	Cluster Index									
	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10
1	0.43 (0.00)									
2	0.41 (0.00)	0.46 (0.00)								
5	0.66 (0.00)	0.34 (0.00)	0.33 (0.01)	0.33 (0.05)	0.47 (0.00)					
10	0.27 (0.25)	0.76 (0.00)	0.27 (0.13)	0.43 (0.02)	0.45 (0.05)	0.51 (0.00)	0.72 (0.00)	0.40 (0.00)	0.53 (0.00)	0.46 (0.00)

3 Acceptance Analysis

In this and the following section, we give results only for the clusters computed using BERT embeddings. We did not observe a significant difference between the included and excluded results. The remaining results can easily be obtained using our code.

We performed a t-test with the null hypothesis claiming no difference between means of citations of accepted and rejected papers. We repeated the experiment for each cluster separately. The computed p-values are very small (0.00 when we round to two decimals) for almost all of the clusters, with few exceptions when cluster count is larger.

4 Review Score Analysis

Similar to the experiment above, for each cluster we computed Pearson correlation coefficient to analyse the relationship between scores given by reviewers and number of citations. We have also calculated the p-value for testing non-correlation. The results are given in Table 2.

5 Discussion

By outcomes of the above experiments, we conclude that peer review process is a good way to measure potential impact of a paper. Accepted papers have significantly higher citation count compared to rejected ones. Review scores are also correlated with the academic impact. However, our findings do not imply causality. Results given above only measure statistical correlation.

We observed when the number of clusters increase, even though the averages of correlations and p-values do not change much, we observe clusters having varying results. We believe that this is a result of having smaller sample sets. We have also conducted the same experiments without excluding outlier papers. Outlier removal increases Pearson correlation from 0.24 to 0.43 for a single cluster, and has similar results for other cluster counts.

We have also computed Spearman's rank correlation coefficient. With 1 cluster, the correlation is calculated as 0.53 and p-value for non-correlation is 0.00. Results are similar for higher number of clusters. On the contrary to Pearson correlation, Spearman's correlation does not decrease when we include outliers. Therefore we conclude that outlier papers disrupt the linear relationship between review score and citation count, but they do not affect their rankings. Spearman's correlation experiments are included in our code.

Our research is limited by several topics. Some of these are institutional bias of the authors, online publication of the paper without the results of peer review, and gender bias.

References

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[2] Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*, 2016.

[3] Dongyeop Kang, Waleed Ammar, Bhavana Dalvi, Madeleine van Zuylen, Sebastian Kohlmeier, Eduard H. Hovy, and Roy Schwartz. A dataset of peer reviews (peerread): Collection, insights and NLP applications. *CoRR*, abs/1804.09635, 2018.

[4] John W. Tukey. *Exploratory Data Analysis*. Addison-Wesley, 1977.