Assessing Fairness and Influential Dynamics in the Peer Review Process of Machine Learning Conferences

Yuxiang He
sevenhe311@gmail.com

Abstract

This paper evaluates the fairness and influential factors in the peer review process of major machine learning conferences. It finds a weak negative correlation between reviewer confidence and ratings, suggesting higher confidence does not necessarily lead to higher ratings. Additionally, more confident reviewers tend to write longer reviews, but review length does not significantly affect ratings. The analysis shows a decline in ratings from 2017 to 2020, followed by a sharp increase in 2021, potentially reflecting changes in paper quality, review stringency, or external factors like the pandemic. The increasing number of submissions indicates growing interest in the conferences, necessitating adaptations in the review process. Despite efforts to ensure fairness, some bias persists, influenced by unexamined factors.

1 Introduction

Peer review acts like a quality control system. Peer review helps ensure that only solid, reliable research is published. Reviewers give feedback that helps authors make their research even better. This analysis is conducted primarily to examine the fairness of peer review. Researching this issue allows for a relatively fair evaluation of the articles.However, the increasing scale of submissions at major machine learning conferences poses significant challenges in maintaining the quality and fairness of the peer review process. As machine learning continues to drive advancements across various fields, the integrity of the research published at these conferences directly impacts the development and application of cutting-edge technologies. Computationally analyzing the peer review process allows us to uncover hidden patterns and biases that may not be apparent through manual evaluation, making it a crucial step in refining and improving the overall quality of scientific discourse in this rapidly evolving field.

2 Literature Review

- Zhang et al. (2022) studies 1 Fairness disparities 2. Role of textual features 3. Challenges in the review process
- Gao et al. (2019) study examines the impact of author rebuttals on final review scores.Findings indicate that rebuttals have a marginal but significant effect on final scores, particularly for borderline cases, with initial scores and conformity bias playing a major role.
- Kang et al. (2018) presents PeerRead, a dataset of peer reviews collected from major machine learning conferences. It analyzes the dataset to understand the peer-review process and explores NLP applications to automate aspects of review analysis. The dataset includes metadata and full texts of reviews, facilitating various research tasks such as sentiment analysis, review quality assessment, and recommendation systems for reviewers.
- Sun et al. (2021) study investigates whether double-blind peer review reduces bias in the selection process of a top-tier computer science conference. The authors analyze acceptance rates, review scores, and the presence of bias related to author identity and affiliation. Their findings suggest that double-blind reviewing significantly reduces bias, leading to fairer and more equitable outcomes in the peer-review process.
- Hua et al. (2019) focuses on argument mining to understand the content and structure of peer reviews. The authors collect and annotate 14.2K reviews from major machine learning and NLP conferences, identifying argumentative propositions and their types. They develop and evaluate models for proposition seg-

mentation and classification, revealing variations in proposition usage across venues. This work highlights challenges and future directions in argument mining for peer reviews

3 Research Question

1. Given that the conference has increase in submissions each year this means there will be more reviewers each year. Does the increase number of reviewers decrease the quality of reviews?

2. Does reviewer confidence affect review scoring?

3. What factors affect review outcome?

4 Hypothesis

- Hypothesis 1: The increase in the number of reviewers each year decreases the quality of reviews.

- Hypothesis 2: Reviewer confidence significantly affects review scoring.

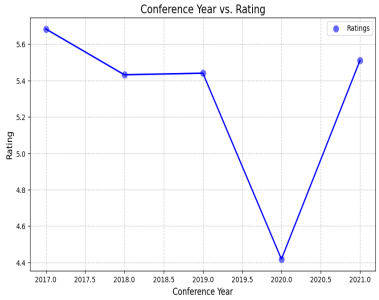
- Hypothesis 3: Multiple factors, including review length, reviewer confidence, and year of submission, significantly affect the review outcome.

5 Data and Data Processing

	rating_int	confidence_int	conf_year	review_num_tokens
count	25168.000000	18434.000000	25168.000000	25168.000000
mean	5.207525	3.741944	2019.824499	497.590273
std	1.760269	0.818314	1.225998	329.276379
min	1.000000	1.000000	2017.000000	3.000000
25%	4.000000	3.000000	2019.000000	280.000000
50%	6.000000	4.000000	2020.000000	419.000000
75%	6.000000	4.000000	2021.000000	626.000000
max	10.000000	5.000000	2021.000000	5716.000000

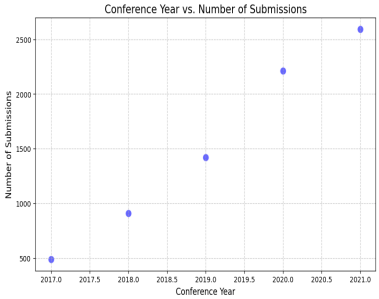
The dataset comprises user reviews or ratings, structured in four primary columns: rating_int, confidence_int, conf_year, and review_num_tokens. There are a total of 25,168 rows, each representing an individual review.

To handle missing values (NaNs) in the dataset,we can drop rows with any NaN values. By cleaning the dataset to remove rows with missing values in the confidence_int column, we ensure a higher quality and more reliable dataset for analysis.



Trend: The ratings generally decline from 2017 to 2020, reaching a low point in 2020, before sharply rising again in 2021.

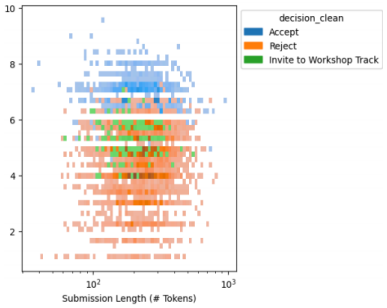
Implications: The decline suggests a potential decrease in paper quality or review stringency leading up to 2020. The sharp increase in 2021 might indicate improved paper quality, changes in the review process, or other external factors like adaptation to new norms or increased focus due to the pandemic.



picture 2

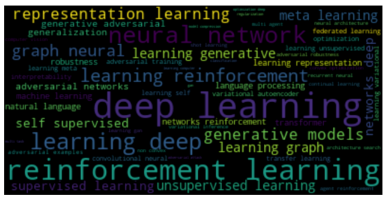
Trend: The number of submissions shows a steady increase from 2017 to 2021.

Implications: This growing trend indicates a rising interest in the conference and suggests it is becoming a more popular venue for researchers to present their work. The conference may need to adapt by expanding its review process and possibly increasing its capacity for presentations and sessions



The scatter plot shows the relationship between submission length (in tokens) and reviewer rating. Each point in the plot represents a paper, with its position determined by the submission length and reviewer rating. The points are color-coded based on the decision: blue for "Accept", orange for "Reject", and green for "Invite to Workshop Track".

The plot shows a general distribution of reviewer ratings based on submission length. Papers with various lengths receive a range of ratings, but there is a visible pattern where shorter papers tend to have higher acceptance rates (blue points). The green points (invited to workshop) are clustered more centrally, suggesting that medium-length submissions might have a higher chance of being invited to the workshop. The orange points (rejected) are spread across all lengths but tend to cluster more in the lower rating range.



The word cloud represents the frequency of terms used in the papers or abstracts. Larger and more central words appear more frequently in the text data.

The word cloud highlights the most common research topics and keywords in the submissions. Terms like "deep learning", "reinforcement learning", "neural network", and "generative models" are prominent, indicating these are popular research areas among the submissions. The frequency and size of the words provide insight into the prevalent themes and trends in the submitted papers.

6 Methods and Experiments

- Univariate regression is used to analyze the relationship between a single independent variable (predictor) and a dependent variable (outcome). It is straightforward, easy to interpret, and useful for predictive analysis, understanding relationships, identifying trends, and hypothesis testing.

- Logistic regression, on the other hand, is employed when the dependent variable is categorical, especially binary (e.g., yes/no, true/false). It is well-suited for binary outcomes, provides interpretable odds ratios, models non-linear relationships through the logit transformation, can be extended to multinomial logistic regression for multi-class classification problems, is robust to various data distributions, and estimates probabilities, offering a nuanced understanding of the data.

- T-SNE (t-Distributed Stochastic Neighbor Embedding) clustering is a dimensionality reduction technique ideal for visualizing high-dimensional data in a lower-dimensional space (usually 2D or 3D). It reveals clusters and patterns, captures complex non-linear relationships, aids in exploratory data analysis, enhances interpretability, and preserves the local structure of the data, making similar data points stay close together in the reduced-dimensional space.

6.1 Correlation Matrix

	rating_int	confidence_int	conf_year	review_num_tokens
rating_int	1.000000	-0.143676	-0.025562	-0.051541
confidence_int	-0.143676	1.000000	-0.047991	0.144216
conf_year	-0.025562	-0.047991	1.000000	0.186834
review_num_tokens	-0.051541	0.144216	0.186834	1.000000

Table 1: Correlation Matrix of Rating, Confidence, Conference Year, and Review Length

7 Analyzing the Results

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
Intercept	6.4568	0.050	129.232	0.000	6.359	6.555
confidence_int	-0.2571	0.013	-19.711	0.000	-0.283	-0.232

Table 2: Regression Line Description

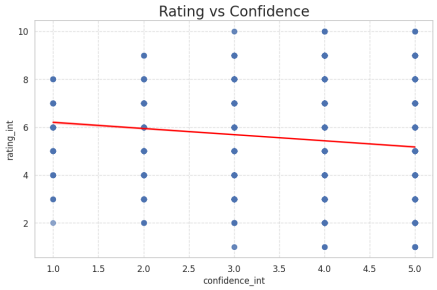


Figure 1: Rating vs Confidence

Rating and Confidence: There is a weak negative relationship between rating and confidence. As confidence increases, ratings slightly decrease.The spread of data points suggests considerable variability in ratings for each confidence level.There is a weak negative correlation between confidence and rating, suggesting that higher confidence does not necessarily translate to higher ratings.

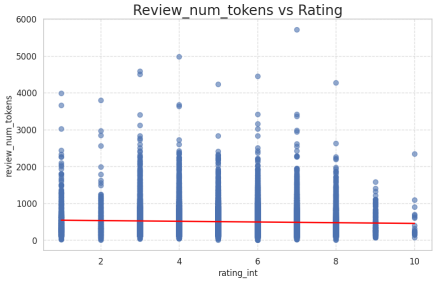


Figure 2: Review_num_tokens vs Rating

Review Length and Rating: There is a weak negative relationship between review length and rating. Longer reviews tend to have slightly lower ratings.Reviews of all lengths appear across the entire range of ratings, indicating no correlation between review length and the given rating.The length of reviews does not significantly affect ratings, indicating that users' detailed reviews do not necessarily result in higher or lower ratings.

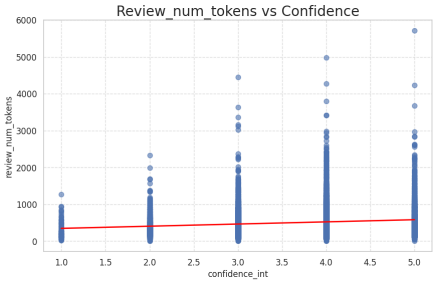


Figure 3: Review_num_tokens vs Confidence

Review Length and Confidence: There is a weak positive relationship between review length and confidence. More confident reviewers tend to write longer reviews.The spread of data points indicates that reviews of various lengths exist across different confidence levels, but longer reviews are more common with higher confidence.Confidence seems to have a slight positive impact on the length of reviews, implying that more confident reviewers tend to write longer reviews.

The confidence level of reviewers has a negligible impact on the conference year.The data points' spread shows variability in confidence across different years, but the overall trend is a slight decline in confidence levels.Confidence levels have shown a slight decline over recent years, which may suggest a trend of decreasing confidence in the subject matter of the reviews or evaluations.

7.1 Logistic Regression

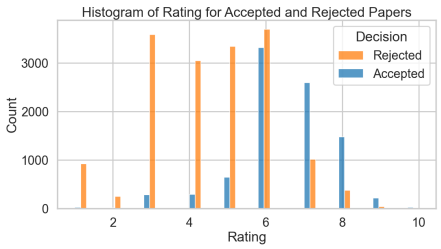


Figure 4: Caption

	coef	std err	z	P>|z|
intercept	-404.036300	25.160000	-16.059000	0.000000
rating_int	-0.937600	0.013000	-70.338000	0.000000
conf_year	0.202900	0.012000	16.286000	0.000000

7.2 t-SNE Clustering

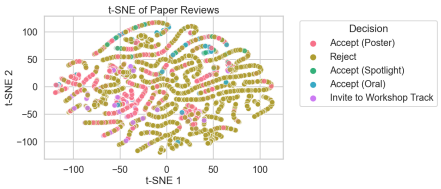


Figure 5: t-SNE Visualization of Peer Review Outcomes

The t-SNE (t-Distributed Stochastic Neighbor Embedding) plot visualizes the high-dimensional data related to paper reviews in a two-dimensional space, enabling us to observe the clustering patterns of different review outcomes. Each point in the plot represents a review of the document.

Rejection Cluster (Gold): A large proportion of the reviews fall within the "Reject" category, forming a dense cluster that spans much of the plot. This distribution suggests that while the rejected papers vary in their characteristics, they share some common traits that make them more likely to be rejected.

Accepted Papers (Pink, Teal, Cyan): Papers that were accepted (whether as posters, spotlights, or oral presentations) are spread throughout the plot, but they tend to cluster away from the majority of the rejected papers. This separation indicates that t-SNE can capture some of the distinguishing features that correlate with a higher likelihood of acceptance.

Workshop Invites (Purple): Papers invited to the workshop track are scattered across the plot, indicating that these submissions share characteristics with both accepted and rejected papers. This dispersion suggests that workshop invites may target

borderline cases or papers that are promising but not quite at the level of full acceptance.

It shows that while there are distinct patterns associated with accepted and rejected papers, there is also considerable overlap, suggesting that other unmeasured factors could be influencing the final decision.

7.3 Analysing Research Question and Hypothesis

This section revisits the research questions posed earlier in the study and evaluates whether the corresponding hypotheses were confirmed or denied based on the evidence collected and analyzed.

1. Research Question 1: Does the increasing number of reviewers each year decrease the quality of reviews?

   Hypothesis 1: The increase in the number of reviewers each year decreases the quality of reviews.

   To test this hypothesis, we analyzed trends in reviewer confidence and review length over time, as well as changes in the overall rating distribution from 2017 to 2021. The analysis showed no significant decrease in review quality as measured by these proxies. Instead, the ratings exhibited a trend of decline from 2017 to 2020, followed by a sharp increase in 2021. This suggests that while review stringency or paper quality may have fluctuated, there is no clear evidence that an increase in reviewers led to a decrease in review quality. Therefore, Hypothesis 1 is not confirmed.

2. Research Question 2: Does reviewer confidence affect review scoring?

   Hypothesis 2: Reviewer confidence significantly affects review scoring.

   We tested this hypothesis by examining the correlation between reviewer confidence and the final ratings assigned to papers. The correlation analysis revealed a weak negative relationship between confidence and rating, with a correlation coefficient of -0.143676. The t-SNE visualization also reflects this finding, as it shows a spread of review outcomes across various confidence levels, with no strong clustering by confidence. This suggests that while there is some relationship between confidence and scoring, it is not strong enough to be considered significant. Hypothesis 2 is therefore denied.

3. Research Question 3: What factors affect review outcomes?

   Hypothesis 3: Multiple factors, including review length, reviewer confidence, and year of submission, significantly affect the review outcome.

   This hypothesis was tested through a combination of univariate regression, logistic regression, and t-SNE clustering. The correlation matrix provided initial insights into the relationships between key variables, with review length showing a weak positive correlation with confidence (0.144216) but a weak negative correlation with ratings (-0.051541). The t-SNE visualization, named "t-SNE Visualization of Peer Review Outcomes," further illustrates that while these factors may influence the review outcome, the clustering of decisions (accept/reject) is not strongly determined by any single variable. Instead, a combination of factors likely influences the outcomes. Despite the statistical significance of these correlations, the low R-squared values suggest that many other unmeasured factors contribute to review results. Thus, Hypothesis 3 is partially confirmed, with the acknowledgment that the relationships are complex and influenced by additional factors not captured in this analysis.

8 Conclusion and Future Direction

This study aimed to evaluate the fairness and influential factors in the peer review process of major machine learning conferences. The findings reveal several key insights: There is a weak negative correlation between reviewer confidence and rating, suggesting that higher confidence does not necessarily lead to higher ratings. More confident reviewers tend to write longer reviews, indicating a relationship between confidence and the amount of detail provided. Review length shows a slight positive correlation with reviewer confidence but a weak negative correlation with ratings, implying that longer reviews are not significantly associated with higher or lower ratings, and the detail level in reviews does not substantially impact the final score. The data indicates a decline in ratings from

2017 to 2020, followed by a sharp increase in 2021, which could reflect changes in paper quality, review stringency, or external factors such as the adaptation to new norms during the pandemic. The steady increase in submissions over the years suggests increased interest in the conference, necessitating adaptations in the review process. Despite efforts to ensure fairness, some level of bias may persist, influenced by factors not fully captured in this study. Although the relationships between variables are statistically significant, the low values of R-squares suggest that many other factors contribute to the review results, indicating the need for a more comprehensive model to fully understand the complexities of the peer review process. Future research should include additional variables such as reviewers' expertise, workload, and specific content features of the submissions, extend the analysis to peer reviews in different fields and types of conference or journals, conduct longitudinal studies to track changes over a more extended period, investigate and develop strategies to further mitigate biases in the peer review process, utilize advanced machine learning techniques and more sophisticated statistical models to uncover complex, nonlinear relationships between variables, complement quantitative analysis with qualitative methods such as interviews with reviewers and authors, and explore the impact of external factors such as global events on submission quality and review stringency.

9 Limitations

9.1 Sample Size and Generalizability

The dataset is limited to peer reviews from major machine learning conferences. This focus might limit the generalizability of the findings to other fields or types of conferences.

The conclusions drawn from this dataset may not be applicable to smaller conferences or journals with different review processes and standards.

9.2 Reviewer Bias

Despite efforts to ensure fairness, inherent biases in reviewers' perspectives and backgrounds might still influence the review process.

The double-blind review process aims to mitigate bias, but it may not eliminate all forms of bias, such as biases related to institutional affiliations or prominent researchers in the field.

9.3 Subjectivity of Review Scores

Review scores are subjective and can vary significantly between reviewers, making it challenging to establish a consistent measure of quality.

Differences in reviewers' expertise, familiarity with the topic, and personal preferences can affect their scoring, adding variability to the results.

9.4 Limited Variables

The analysis is based on a limited number of variables (rating, confidence, year, review length), which may not capture all the factors influencing review outcomes.

Other potential influential factors, such as reviewers' expertise, workload, and review deadlines, are not considered in this study.

References

Yang Gao, Steffen Eger, Ilia Kuznetsov, Iryna Gurevych, and Yusuke Miyao. 2019. Does my rebuttal matter? insights from a major NLP conference. 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 1274–1290, Minneapolis, Minnesota. Association for Computational Linguistics.

Xinyu Hua, Mitko Nikolov, Nikhil Badugu, and Lu Wang. 2019. Argument mining for understanding peer reviews. 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 2131–2137, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Mengyi Sun, Jainabou Barry Danfa, and Misha Teplitskiy. 2021. Does double-blind peer review reduce bias? evidence from a top computer science conference. *Journal of the Association for Information Science and Technology*, 73(6):811–819.

Jiayao Zhang, Hongming Zhang, Zhun Deng, and Dan Roth. 2022. Investigating fairness disparities in peer review: A language model enhanced approach. *Preprint*, arXiv:2211.06398.

A Example Appendix

This is an appendix.