iARP: Identity-Agnostic Reviewer Profiling

Alon Schneider

Abstract

The peer review process is essential for maintaining scientific research quality, yet reviewer analysis remains limited due to anonymity. Understanding reviewer behavior, biases, and decision patterns is crucial for improving fairness and reliability in peer review systems.

To address this, we propose **iARP**, a method for profiling reviewers despite anonymity. Our approach defines five key features—**proxy impact factor, acceptance ratio, chutzpah, objectivity-confidence, and sentiment ratio**—to infer reviewer tendencies based on review characteristics. Leveraging over one million data points from multiple sources, we construct meaningful reviewer profiles without requiring identifiable information.

We present these profiles visually using a **radar graph**, offering an intuitive way to compare reviewers and analyze behavioral patterns. This work lays the foundation for future research in reviewer profiling and contributes to enhancing transparency and fairness in peer review.

1 Introduction

Peer review is central to academic publishing, ensuring research quality and credibility. It serves as a filtering process for evaluating manuscripts before publication. Traditionally, it follows a **single-blind** model, where reviewers remain anonymous, or a **double-blind** model, where both authors and reviewers are anonymous. While anonymity helps reduce bias, it also complicates the assessment of reviewer fairness, expertise, and reliability.

Recently, **open peer review** platforms have emerged to increase transparency by disclosing reviewer identities and comments. This model enhances accountability and allows the academic community to better evaluate reviewer contributions. Notable platforms embracing open review include:

- **Publons** – Allows reviewers to track and showcase their peer review and editorial contributions.
- **OpenReview** – Widely used in machine learning and artificial intelligence conferences (e.g., NeurIPS, ICLR) to facilitate open discussion and allow authors to respond to reviews.
- **F1000Research** – Supports post-publication peer review, where reviews and reviewer names are published alongside the article.
- **eLife** – Implements a more transparent review process, where reviews and editorial decision letters are openly shared.
- **PeerJ** – Encourages open peer review by giving authors the option to disclose reviewer identities.

Despite the growing adoption of open review models, the majority of academic reviews remain anonymous. This lack of identifiable information makes it difficult to analyze reviewer behavior, detect biases, and ensure accountability. As a result, understanding who the reviewers are and how they assess manuscripts remains a significant challenge.

2 Related Work

The study of peer review has gained significant attention in recent years, leading to the creation of multiple datasets and research efforts aimed at analyzing reviewer behavior and decision-making patterns.

2.1 PeerRead

PeerRead is the first public dataset of scientific peer reviews, offering insights into the review process. It contains 14.7K paper drafts with accept/reject decisions from top conferences like

ACL, NeurIPS, and ICLR, along with 10.7K expert-written reviews. PeerRead has facilitated NLP research on acceptance prediction and review aspect scoring (e.g., originality and impact).

2.2 When Reviewers Lock Horn: Identifying Disagreements in Peer Reviews

With the rise in submissions, especially at top AI conferences, reaching a consensus in peer review has become more complex. This work introduces **ContraSciView**, a dataset of 28K review pairs (8.5K papers) from ICLR and NeurIPS, designed to detect *contradictions among reviewers*. It also presents a baseline model for identifying contradictory statements, marking the first attempt to **automate disagreement detection** in peer reviews.

2.3 NLPeer

NLPeer is a dataset for studying peer review and developing NLP-based assistance tools. It aims to help junior reviewers improve accuracy and confidence. Notably, its creators state that **reviewer profiling is not an intended use case**, emphasizing ethical concerns.

2.4 PeerConf: A Dataset for Peer Review Aggregation

PeerConf aggregates peer review data from five scientific conferences, comprising 3,242 reviews for 1,236 papers. It includes recommendation scores, confidence levels, textual reviews, and final decisions, with identifiable information removed for privacy. The dataset supports research in **acceptance prediction, review summarization, and reviewer behavior analysis**.

3 Methods

In this section, we describe the methodology used to profile anonymous reviewers. Our approach involves multiple steps, from data exploration and feature engineering to modeling and evaluation. Given the lack of direct reviewer identification, we focus on extracting meaningful insights from review content and metadata. By leveraging proxy features, we aim to characterize reviewers based on their writing style, sentiment, confidence, and other behavioral indicators.

3.1 Data Collection

To build an effective reviewer profiling system, we gathered data from multiple sources, ensuring a diverse and comprehensive dataset. The data was collected using different methods, including **scraping open-source reviewing systems**, extracting information from **publicly available websites and datasets**, processing **private datasets from research repositories**, and obtaining structured data from **conference proceedings, PDFs, and JSON files**.

The final dataset consists of **over 1.3 million samples**, covering a wide range of peer review data from various academic domains. Specifically, we utilized the following sources:

- **PeerRead** – A dataset of scientific peer reviews, including paper drafts and corresponding accept/reject decisions from top-tier conferences.
- **ReviewData** – A large-scale dataset containing peer review texts, scores, and metadata, facilitating the study of reviewer behavior.
- **OpenReview Scraped Data** – Reviews collected from OpenReview, a widely used open peer review system for AI conferences.
- **PeerConf** – A dataset aggregating peer review information from multiple scientific conferences.
- **Publons** – A dataset from Publons, a platform that tracks and showcases peer review contributions from researchers.

By leveraging these diverse sources, we ensure that our dataset captures **a broad spectrum of reviewing styles, decision-making tendencies, and reviewer behaviors**. The collected data serves as the foundation for extracting meaningful reviewer features, allowing us to profile reviewers even in anonymous settings.

3.2 Exploratory Data Analysis

Exploratory Data Analysis (EDA) is a crucial step in understanding the structure and characteristics of the dataset. Since reviewer identities are not available, our goal is to uncover patterns that can help in profiling reviewers using **proxy features**. EDA allows us to:

- Identify key variables that can contribute to reviewer profiling.
- Detect relationships between different features, such as confidence levels, sentiment scores, and review length.

- Examine the distribution of review attributes across different academic venues.
- Assess potential biases in the data that might affect the analysis.

By systematically analyzing the data, we gain insights into the key factors that distinguish reviewers, which in turn informs feature selection and modeling. The findings from this stage serve as the foundation for building meaningful reviewer representations.

3.3 Preprocessing

Once the data was collected from various sources, it required extensive preprocessing to ensure consistency, usability, and compatibility for feature extraction. Given the diverse nature of the datasets—originating from different formats such as **JSON, PDF, scraped text, and structured CSV files**—we performed multiple steps to clean, merge, and standardize the data.

The preprocessing phase involved the following key steps:

- **Data Cleaning:** Removing missing, incomplete, or duplicate records to ensure data integrity.
- **Standardization:** Unifying different formats, renaming columns for consistency, and converting textual data to a common encoding.
- **Merging Datasets:** Integrating multiple data sources into a unified structure while ensuring no loss of critical information.
- **Text Processing:** Tokenizing review text, lowercasing, removing stopwords, and handling special characters to prepare for sentiment and linguistic analysis.
- **Numerical Feature Scaling:** Normalizing numerical attributes (such as confidence scores and review lengths) to ensure comparability across datasets.

Through this preprocessing pipeline, we transformed the raw datasets into a structured and analyzable format, enabling effective **feature engineering** in the subsequent stages. This step was crucial for ensuring that all review data could be meaningfully compared and used in the construction of reviewer profiles.

3.4 Feature Engineering

In this step, we define the five key features used to profile reviewers. These features capture different aspects of reviewer behavior, including the nature of their reviews, decision tendencies, and sentiment analysis. Each feature was carefully designed to extract meaningful information from the available data.

3.4.1 Feature 1: Proxy Impact Factor Index

Previous research on Publons explored the relationship between **review length, country of the reviewer, and impact factor**. Inspired by these findings, we use the **mean review length** of a reviewer as a proxy to estimate their country of origin. The country with the closest statistical review length distribution is assigned to the reviewer. The reviewer is then assigned the **median impact factor** of that country.

To compute this feature, we extract the **mean**, **median**, and **standard deviation** of review lengths for each country and reviewer.

3.4.2 Feature 2: Acceptance/Reject Ratio

The acceptance ratio reflects a reviewer's decision tendencies. This feature is computed as the ratio of **accepted papers to the total reviewed papers**, where:

$$\text{Acceptance Ratio} = \frac{\text{Accepted Papers}}{\text{Accepted + Rejected}}$$

3.4.3 Feature 3: Chutzpah (Audacity)

This feature measures a reviewer's willingness to reject papers written by highly established authors. We identify **high-impact authors** as those whose **h-index** is above the **70th percentile** of the dataset. The feature is then computed as the **rejection ratio** of papers authored by these high h-index individuals:

$$\text{Chutzpah} = \frac{\text{Rejected by high h-index}}{\text{Total Papers by high h-index Authors}}$$

A high Chutzpah score suggests that the reviewer tends to reject even well-established authors, potentially indicating a more critical reviewing style.

3.4.4 Feature 4: Objectivity Confidence

Each review contains a **confidence score** (ranging from 1 to 5) indicating how certain the reviewer is in their assessment. Additionally, we use **TextBlob** to extract the **subjectivity score** of the review text and define **objectivity** as:

$$\text{Objectivity} = 1 - \text{Subjectivity}$$

For each reviewer, we compute the **mean confidence** and **mean objectivity** over all their reviews. The final **Objectivity Confidence Score** is calculated as:

$$\text{ObjeConf} = \frac{\text{Normalized Conf} + \text{Normalized Obj}}{2}$$

where both confidence and objectivity are normalized using the global means obtained from the dataset.

3.4.5 Feature 5: Sentiment Ratio

To quantify sentiment tendencies, we use **TextBlob** to compute the **sentiment polarity** of each review. We first calculate the **mean and standard deviation of sentiment scores** over a dataset of **100K samples**. The **Sentiment Ratio** for a reviewer is then defined as:

$$\text{Sentiment Ratio} = \frac{\text{Mean Sentiment Score}}{\text{Global Mean Sentiment}}$$

This feature provides insight into whether a reviewer tends to write more positive or negative reviews relative to the overall dataset.

4 Results

In this section, we present key findings from our dataset analysis and demonstrate how the five extracted features were aggregated into a radar graph for reviewer profiling.

4.1 Data Analysis and Visualizations

To validate the effectiveness of the extracted features, we conducted an in-depth analysis of the dataset and examined various distributions and relationships between key variables.

4.1.1 Feature 1: Proxy Impact Factor Index

One of the most significant findings relates to the relationship between **median review length** and **impact factor** across different countries. Using a **3D scatter plot**, we observed that countries are **well-separated** when plotting the median or mean review length against the impact factor. Specifically, countries with a **higher median review length** also tend to have a **higher impact factor**, highlighting a correlation between review length and journal prestige. This insight confirms that **Proxy Impact Factor Index** is a valuable feature for reviewer profiling.

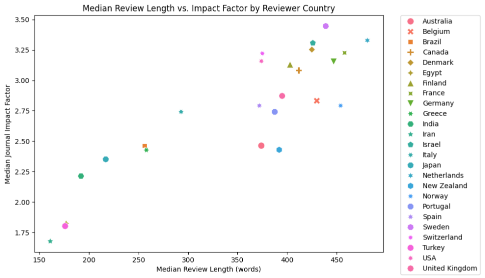

Figure 1: 3D scatter plot showing the relationship between median review length and impact factor across countries.

4.1.2 Confidence Distribution

Another interesting finding is the **distribution of reviewer confidence scores**. Confidence scores were analyzed across the dataset, revealing a strong human tendency in self-assessment. The mean confidence score was found to be **0.73**, with over **50% of reviewers reporting confidence levels above this mean**. This result highlights that while many reviewers are highly confident in their assessments, a considerable portion remains cautious in their evaluations.

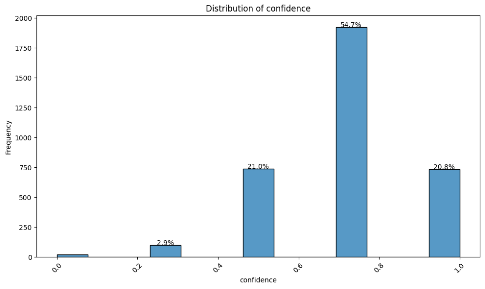

Figure 2: Distribution of confidence scores among reviewers.

4.1.3 Objectivity Distribution

We also analyzed the **objectivity distribution**, which was computed over **hundreds of thousands of reviews** using subjectivity scores extracted via TextBlob. Interestingly, the objectivity scores follow a **normal distribution**, suggesting that objectivity in reviews is balanced across the dataset. This finding reinforces the validity of **Objectivity Confidence** as a robust profiling feature.

4.2 Feature Aggregation into a Radar Graph

To visualize reviewer profiles, we aggregated five extracted features into a **radar graph**. The **Proxy Impact Factor Index** estimates impact factor based on review length, while the **Acceptance**

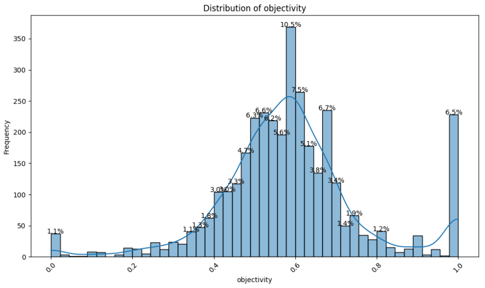

Figure 3: Normal distribution of objectivity scores computed from review texts.

Ratio measures accepted papers among reviewed submissions. **Chutzpah (Audacity)** captures the tendency to reject high-profile authors, **Objectivity Confidence** combines confidence scores with text-based objectivity, and the **Sentiment Ratio** quantifies sentiment relative to a global benchmark. These features provide a structured and interpretable representation of reviewer behavior.

By normalizing these features and plotting them on a radar graph, we enable a direct comparison of reviewers, allowing us to identify **reviewer tendencies and behavioral patterns** at a glance. This method provides a structured way to analyze anonymous reviewers by capturing essential characteristics in a visually interpretable manner.

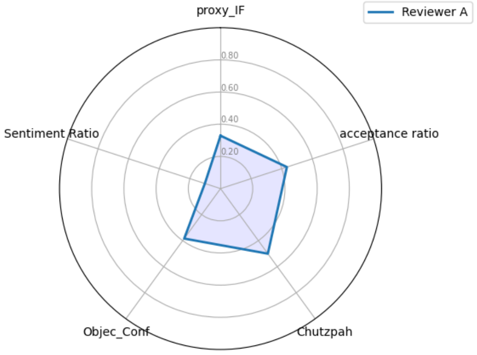

Figure 4: Radar graph representation of a reviewer profile based on the five extracted features.

Through this visualization, we can compare different reviewers, detect outliers, and analyze trends in reviewer behavior. This approach serves as a crucial step toward the goal of profiling reviewers even in **anonymous settings**.

5 Conclusions and Future Work

This study demonstrates that it is possible to analyze and profile anonymous reviewers without relying on identifiable information. By leveraging key review characteristics—such as confidence, objectivity, sentiment, and decision tendencies—we developed a framework that provides meaningful reviewer representations. Our results show that profiling can be achieved using proxy features while maintaining anonymity, offering insights into reviewer behavior, biases, and reviewing styles.

Proxy measurements, such as review length as a proxy for impact factor, have proven useful in inferring reviewer characteristics. Future work should explore additional **proxy-based approaches** to enhance reviewer profiling using available data. Identifying and validating new proxies could significantly improve the robustness and accuracy of profiling techniques, especially when dealing with limited or anonymized information.

Additionally, many aspects of human behavior, such as confidence and objectivity distributions, tend to follow **normal distributions**. Future studies could leverage these behavioral patterns to refine feature extraction methods and develop more sophisticated reviewer representations. Expanding the dataset, incorporating **more advanced features**, and eventually integrating **non-anonymous data** could further validate and enhance this approach.

As peer review continues to evolve, developing advanced analytical methods will be crucial for ensuring transparency, fairness, and accountability in the review process.

References

Acknowledgments

The code and data used in this study are available on GitHub: https://github.com/AlonSchneider27/iARP.

We acknowledge the assistance of AI tools, including **ChatGPT**, **Claude**, and **Perplexity**, in supporting various aspects of this work, such as text generation, data analysis, and literature review. These tools contributed to improving efficiency and refining the presentation of our findings.

A References

References

[1] Kang, D., Ammar, W., Dalvi, B., Van Zuylen, M., Kohlmeier, S., Hovy, E., and Schwartz, R., 2018. *A dataset of peer reviews (PeerRead): Collection, insights and NLP applications*. arXiv preprint arXiv:1804.09635.

[2] Kumar, S., Ghosal, T., and Ekbal, A., 2023. *When Reviewers Lock Horn: Finding Disagreement in Scientific Peer Reviews*. arXiv preprint arXiv:2310.18685.

[3] Dycke, N., Kuznetsov, I., and Gurevych, I., 2022. *NLPeer: A unified resource for the computational study of peer review*. arXiv preprint arXiv:2211.06651.

[4] Hasan, Md Tarek; Shamael, Mohammad Nazmush; Billah, Mutasim; Akter, Arifa; Hossain, Md Al Emran; Islam, Sumayra; Shatabda, Swakkhar; Islam, Salekul, 2022. *PeerConf: A dataset for peer review aggregation*. Mendeley Data, V1, doi: 10.17632/wfsspy2gx8.1.

[5] Salimi, B., Parikh, H., Kayali, M., Getoor, L., Roy, S., and Suciu, D., 2020, June. *Causal relational learning*. In Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data (pp. 241-256).

[6] *It's Not the Size That Matters*. Clarivate Analytics. Clarivate Blog.