Automatically detecting open academic review praise and criticism[1]

Mike Thelwall, University of Wolverhampton, UK. ORCID:0000-0001-6065-205X
Eleanor-Rose Papas, F1000Research, UK. ORCID: 0000-0003-4293-4488
Zena Nyakoojo, F1000Research, UK. ORCID: 0000-0001-6812-4120
Liz Allen, F1000Research, UK. ORCID: 0000-0000-0002-9298-3168
Verena Weigert, Jisc, UK. ORCID: 0000-0003-4887-7867

Purpose: Peer reviewer evaluations of academic papers are known to be variable in content and overall judgements but are important academic publishing safeguards. This article introduces a sentiment analysis program, PeerJudge, to detect praise and criticism in peer evaluations. It is designed to support editorial management decisions and reviewers in the scholarly publishing process and for grant funding decision workflows. The initial version of PeerJudge is tailored for reviews from F1000Research's open peer review publishing platform.

Design/methodology/approach: PeerJudge uses a lexical sentiment analysis approach with a human-coded initial sentiment lexicon and machine learning adjustments and additions. It was built with an F1000Research development corpus and evaluated on a different F1000Research test corpus using reviewer ratings.

Findings: PeerJudge can predict F1000Research judgements from negative evaluations in reviewers' comments more accurately than baseline approaches, although not from positive reviewer comments, which seem to be largely unrelated to reviewer decisions. Within the F1000Research mode of post-publication peer review, the *absence* of any detected negative comments is a reliable indicator that an article will be 'approved', but the presence of moderately negative comments could lead to either an approved or approved with reservations decision.

Originality/value: PeerJudge is the first transparent AI approach to peer review sentiment detection. It may be used to identify anomalous reviews with text potentially not matching judgements for individual checks or systematic bias assessments.

Keywords: Sentiment analysis; peer review; open peer review; F1000Research.

Introduction

Academic publishers manage millions of peer review reports for their journals, conferences, books, and publishing platforms, with usually at least two reviewers per document (Clarivate, 2018). These reports are typically accompanied by a judgement, such as accept, minor revisions, major revisions, or reject. These reports are typically read by journal editors and authors, but there is often inconsistency between reviewers' narrative descriptions and their overall judgement decisions (e.g., Langford and Guzdial, 2015). Assessing the quality of a peer review report is complex due to the variability in objectives and guidelines issued to reviewers across publishers (Jefferson *et al*., 2002). This variability can make it difficult for reviewers and for journal staff and editors digesting comments and decisions affecting the outcomes of submitted manuscripts. Recruiting experts for peer review can be difficult and publishers rely on subject-specialist editors to identify inconsistencies and take appropriate action. As a result, there is increasing interest in the potential of technological solutions, including

[1] Thelwall, M., Papas, E., Nyakoojo, Z., Allen, L. & Weigert, V. (in press). Automatically detecting open academic review praise and criticism. *Online Information Review*.

Artificial Intelligence (AI), to support editors and publishers managing the peer review process, including the consistency of peer review reports.

Sentiment analysis AI software works by detecting patterns in texts that associate with positive and/or negative sentences, including words and phrases. It has been primarily used to analyse consumer product reviews, such as for phones, restaurants, and cars, predicting the opinions of reviewers in the absence of explicit overall evaluations. Applying software to detect judgements in peer review documents is a broadly similar task, with overall review judgements (e.g. 'Accept' or 'Approve') providing evidence that can be used to train a sentiment analysis system. However, due to complex academic terminology and varied reasons for criticising a study, general sentiment analysis software is unlikely to be able to detect academic reviewer judgements. Thus, specialist software is needed for this task. There are two general approaches for sentiment analysis: machine learning and lexicon-based (Taboada, *et al.*, 2011). The machine learning approach tends to produce opaque solutions so that the reason for predictions is unknown. For example, support vector machines work by finding hyperplanes in high dimensional space (Steinwart and Christmann, 2008), giving humans no intuition about the solution. Whilst some are more transparent, such as Naïve Bayes and decision trees, the rules produced can be complex and these are not the best performing on most tasks. In contrast, lexical algorithms can trace decisions to the presence or absence of specific words or phrases, and linguistic rules (e.g., for negation), and are therefore preferable when transparency is needed.

This paper introduces a transparent lexical sentiment analysis program, *PeerJudge*, to estimate the judgement of an academic article reviewer based upon praise or criticism in the accompanying peer review report. It is transparent in the sense that each outcome is reported alongside a simple rationale, such as the presence of judgement words and the application of named linguistic rules, such as negation (see an example in the Methods). This has the same goal as a previous paper (Wang and Wan, 2018) except that the program has a human-understandable decision mechanism, is applied to a multidisciplinary collection of articles rather than conference papers from a narrow field and separates praise from criticism. The program was developed and tested on F1000Research open peer review reports since these are publicly available, with secondary tests on publicly available computer science conference reviews from the International Conference on Learning Representations (ICLR).

Background: Open peer review and sentiment analysis

This section gives some background on open peer review since differences between open and closed peer review might influence the extent to which PeerJudge can work on both systems.

Open peer review

Open peer review is an aspect of open science (Morey *et al.*, 2016), promoting research transparency, although most authors currently believe that double-blind reviews are preferable (Moylan *et al.*, 2014; Mulligan *et al.*, 2013; Ross-Hellauer *et al.*, 2017). There are many different types of open peer review, with not all including full public disclosure of the review (Ford, 2013; Ross-Hellauer, 2017). F1000Research, as used in the current article, has a maximal definition, making the authors, reviewers, manuscripts, reports, and comments publicly available throughout the review process. One of the goals of F1000Research is to help to overcome the publication bias towards more positive research outcomes (e.g., Emerson *et al.*, 2010).

Open peer review has differences from closed peer review: open reviews seem to take longer to complete, be more polite, and be more careful than comparable closed reviews for the same psychology journal (Walsh *et al*., 2000). For the Journal of Inflammation, review quality was similar for open peer review and single-blind reviews (Kowalczuk *et al*., 2015). For the BMJ, researchers were 12% more likely to decline open review invitations, but open reviews were not statistically significantly more likely to take longer or have a different quality (Van Rooyen *et al*., 1999); a later BMJ study agreed, but found that public reviews took longer to write (Van Rooyen *et al*., 2010). In general, studies of the quality of open peer review reports have produced mixed findings (Bravo *et al*., 2019; Jefferson *et al*., 2002), and so its effect may vary by field.

Sentiment analysis for peer review reports

There has been increasing interest in the potential of software and artificial intelligence (AI) to support various aspects of the scholarly publishing and discovery workflow. Software now exists to support reviews of aspects of the quality of papers (Frontiersin, 2018; Heaven, 2018; Mrowinski *et al*., 2017; Price and Flach, 2017) or peer review reports (Sizo *et al*., 2019), or to classify the content of peer review reports (Caciagli, 2019). Peer review evaluation and rating software for the Open Journal Systems (OJS) open-source journal publishing platform suggests a rating for the quality of a review, summarising reasons for the judgement (Priatna *et al*., 2017). Its aim is to help editors to score reviewer quality rather than to predict the review outcome, however. One important aspect of a review is whether the comments are consistent with the overall recommendation.

Sentiment analysis software has been designed to estimate the sentiment of general texts (Leijen, 2014; Liu, 2012) and to assess the helpfulness of student peer feedback (Xiong and Litman, 2011), showing that it is a challenging task (Yadav and Gehringer, 2016). One study used off-the-shelf generic sentiment analysis software for polarity and subjectivity to compare open and closed academic reviews, finding no differences (Bravo *et al*., 2019). Generic software is not optimal for this task, however, because sentiment is likely to be expressed using specialist and indirect terminology within academic reviews.

A few sentiment analysis systems have been designed for peer review. One evaluated open reviews with anonymous reviewers from the International Conference on Learning Representations 2017 and 2018, building a machine learning neural network system to estimate the sentence sentiments and overall decisions based on grouping the 10-point scale reviewer ratings into Accept (6-10) or Reject (1-5); or into Accept (7-10), Borderline (5-6), or Reject (1-4) (Wang and Wan, 2018). Using 10-fold cross-validation (but not separating the development and evaluation data), the method was 78% correct on the two-class problem but 61% correct on the three-class problem. The positive and negative sentences extracted were able to give insights into the evaluative language used in reviews. Using part of the PeerRead dataset of information processing conference papers, reviews, and judgements (Kang *et al*., 2018), another approach uses neural network machine learning that harnesses paper text, review sentiment (separately for each sentence with the social media sentiment detection software VADER), and review text in order to build an abstract model to predict the overall decision for a paper, with good performance (Ghosal *et al*., 2019). The paper text and reviews are split into 512-dimensional sentence-level semantic vectors using an encoder. This hybrid approach risks learning the topics of accepted papers rather than the logic behind judgements, however, so may not perform well in the longer term when there is a different set of hot topics. The algorithm may also learn the affiliations of successful authors to aid its judgements because the paper does not state that author metadata was excluded.

Methods

PeerJudge was developed on the model of the lexical social media sentiment analysis software SentiStrength (Thelwall *et al*., 2012). SentiStrength uses a manually curated list of sentiment words and their positive or negative sentiment strengths together with a set of semantic rules, such as for negation, to estimate the strength of positive and negative sentiment in a short text. PeerJudge was built to detect review judgements by (1) constructing a new manually curated lexicon of praise and criticism terms for this task and (2) testing additional semantic rules. Both tasks relied on predominantly manual inspections of a development corpus of peer review reports. Whilst there are existing lexicons for related tasks, academic judgements are typically very formal, and this style does not seem to be covered by any current open lexical resource. For example, one public argumentation lexicon includes non-academic phrases, such as “hold on a sec” (Somasundaran, Ruppenhofer, and Wiebe, 2007).

Development and evaluation data sets

An increasing number of journals and publishers offer access to the peer review reports presented to accompany a decision to publish a submitted piece. Peer review reports are published by the BioMed Central family of journals, the F1000 family of journals, The Bmj, some computing conferences, and others. The journal-like publishing platform F1000Research was used as the source of peer review reports because they are open and easily accessible on the site in text format (rather than PDF), and there are enough to support a rigorous test of Peerjudge on reviews of scholarly published outputs. Its reviews provide a challenging sentiment analysis dataset because they are multidisciplinary, allowing variation in the nature of criticisms, and include standard judgement texts (which were removed before processing) so the reviewer does not need to write their own overall conclusions. Such conclusions might otherwise be expected to contain easily detectable summary judgements.

F1000Research is an open access, open peer review, general (non-specialist) platform that publishes articles after editorial checks but before peer review, then contacts reviewers for formal peer review reports. These reports are subsequently published alongside the article, usually within a few working hours of receipt, after a check from the editorial team to ensure the review is suitable and requires no clarifications or revisions before publication, for example if the reviewer recommendation does not seem to match the body of the report. When an article reaches an agreed threshold (“at least two *Approved* peer review reports, or one *Approved* plus two *Approved with Reservations* reviews”: F1000Research, 2019a), it is submitted for scholarly indexing across number of industry-standard bibliographic databases (e.g. PubMed, Scopus). Authors can revise their article and publish a new version regardless of the peer review status, and in these cases the reviewers are invited to re-review the article. F1000 platforms publish a range of research output types and the peer reviewer requirements are tailored somewhat according to the article types. Nevertheless, all reviewers are asked to provide (a) a narrative report, and (b) a judgement: ‘approved’, ‘approved with reservations’ or ‘not approved’. The post-publication peer review model means that the article is openly available regardless of the approval status it has been assigned. The main analysis here focuses on ‘Research Article’ submissions, which essentially conform to the standards and formats most typically used across the scholarly publishing industry and are likely to evoke the most detailed narrative assessments in reviewer reports. F1000Research is unusual for being equally willing to publish null results and replication studies, but Research Articles should still "present originality in findings and insights" (F1000Research, 2019b).

Peer review reports associated with the first version of all F1000Research articles available in July 2019 were downloaded. Additionally, F1000 operates identical publishing platforms in partnership with several research funders and so all peer review reports available on Gates Open Research, Wellcome Open Research, and HRB Open Research were collected for additional analyses. F1000Research publishes many types of article, including Case Reports, Antibody Validation Articles, Software Tools, Research Notes, and Opinion Articles. For consistency, only publications of type Research Article were retained. F1000Research publishes multiple versions of articles because authors can post updates in response to reviewer (or other) comments. Reports were discarded if they were associated with any version of an article other than the first. This is important because follow-up reviewer reports may simply confirm that previously requested changes have been completed.

The text and decision were extracted for each report, without the reference section, when present. Standard decision texts were also removed from reports, such as "I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard." and any mandatory peer review questions and answers, because this would make the sentiment analysis task misleadingly accurate and F1000Research-specific. After this, blank reviews were also removed. These were only found for articles with an *Approved* decision. Whilst the lack of anything to report about a manuscript is implicit approval from the submitting reviewer, the purpose of this article is to assess the text of reviews.

The version 1 reviewer reports on documents of the type Research Article from F1000Research were split into development and evaluation subsets using a random number generator, with stratified sampling to ensure that the number of reports corresponding to each of the three decisions was proportional to the number found in the overall corpus. The development set contained 500 reports and the evaluation set the rest (1778). To align with SentiStrength, which scores positive sentiment on the inter scale 1 to 5 and negative sentiment on the integer scale -1 to -5, the judgements were converted as follows.

- *Approved* (described in the rest of this article as ***Accept***): praise 4, criticism -1
- *Approved with Reservations* (described in the rest of this article as ***Revise***): praise 3, criticism -3
- *Not Approved* (described in the rest of this article as ***Reject***): praise 1, criticism -4

The evaluation set was not read or used for any purpose until computing the final statistics for this paper. Reviews on version 1 Research Articles from three sister sites were also analysed for secondary evaluation purposes only: Gates Open Research, Wellcome Open Research, and HRB Open Research. This data is shared online (10.6084/m9.figshare.9946748).

As a partial test of whether PeerJudge gives different results for other types of output, it was also tested on peer reports from the International Conference on Learning Representations.

PeerJudge Development

The first author read the reports in the development set and used it to construct a list of terms for the lexicon that seemed to associate with judgements. For example, "mistake" was added with score -4 (likely to be used in the context of strong criticism). Care was taken to take polysemy and differing contexts into account. For example, the term "error" could refer to author mistakes or author methods for ensuring statistical model error terms were correctly analysed. When distinctive phrases were found that had phrase-based meanings, they were added to a list of sentiment phrases instead of as individual terms (e.g., “did not reflect” scores -2: likely to be used in the context of mild criticism). During this process, common sense was used to make the terms as general as possible. For example, when “mistake” was found, other versions of this word were included with a wildcard entry: “mistake*” and “mistook” were both added.

The term lexicon and set of phrases were used to build the initial version of PeerJudge from the development set of reviews, which was then applied to the development set. Anomalies in PeerJudge scores for the development set were then examined (articles where PeerJudge disagreed with the reviewer) to find additional words or phrases and to modify the sentiment strengths of words already in the lexicon. This step was repeated several times.

Two extra rules were developed to capture additional methods of expressing judgements, both of which used questions. Criticism opinions about a paper were often expressed in the form of a question, such as “Why did you …?” rather than through a judgment term or phrase. The first extra rule was to ignore praise in questions, for this reason. The second rule was to give all question sentences a minimum criticism score of -2 (mild criticism) due to the implicit criticism of a question in an academic reviewing context. Whilst questions may also suggest future work, their normal role was to express implicit criticism. These rules both intuitively aligned with the data but lowered the accuracy of PeerJudge and so were discarded (but retained as software options for users). This suggests that questions were not always critical. For example, reviewers might copy a paper’s research questions (if any) to praise them, or might like a paper so much that they phrase criticism as a question to weaken it.

Two machine learning stages were used to refine the human-generated version of PeerJudge. For the first stage, the sentiment weights for each term were individually increased and decreased by 1, retaining changes that increased overall accuracy. This was then repeated until no changes were made. The changes were manually checked, and anomalous changes were rejected (e.g., making “but” praise, +3). The second machine learning stage examined all words in all reviews and calculated the average difference between the PeerJudge predicted sentiment scores and the review sentiment. Thus, words occurring more often in reviews with underestimated praise would have a high average praise difference and the opposite for criticism. The most common words with high average differences were then manually checked and those that seemed intuitively valid were added to the lexicon. For example, “and” was rejected as not meaningful but “clearly” was added.

The development dataset is unbalanced in the sense that there are unequal numbers of review outcomes. In this situation, the two machine learning stages can give misleading results because they will tend to push terms towards the scores of the dominant classes. This was considered during the human judgement stage when transferring terms and term strengths to the main dictionary.

PeerJudge reports both the dual scores and the score rationale. For example, the output for the (short) review, “This is a well written paper with flawed methods,” is as follows:

> Praise strength 2 and criticism strength -4. Classification rationale: [term weight(s) changed by idiom “**well written”**] This is a **well**[2] **written** paper with **flawed**[-4] methods.

Evaluation

The accuracy of PeerJudge was evaluated in two different ways: correlation with reviewer assessment and counting the number of correct predictions.

1. For correlation tests, Pearson correlations were calculated between the PeerJudge scores for a report and the reviewer decision associated with the report, separately for praise and criticism. Positive correlations suggest that PeerJudge is at least to some extent identifying the opinions of reviewers, on average. This method is useful for unbalanced data sets because high accuracy can be difficult to achieve on these. The baseline correlation is 0: the value expected if guesses were made at random or if all judgements were assigned the majority case.
2. For exact match counts, the praise and criticism PeerJudge scores were used separately to predict the review outcome and the number of correct predictions was counted. The baseline count is the majority class (Accept). The simplest rule is to predict that all reviewers recommend Accept (i.e., *Approved* in F1000Research).

Results

Overall predictions

The correlation results for the F1000Research evaluation data set show that PeerJudge praise scores for the evaluation data set bear no relation to reviewers' decisions (correlations close to 0) but PeerJudge criticism scores correlate moderately strongly with reviewers' decisions (the two correlation columns in Table 1). The same trend occurs across the different F1000 publishing platforms (F1000Research, Gates Open Research, Wellcome Open Research, and HRB Open Research), with low praise prediction correlations for peer reviewer reports and moderate criticism prediction correlations (the two correlation columns in Table 1). As a result, praise predictions are ignored for the remainder of this analysis due to their lack of predictive power.

Table 1. Data size, accuracy, and Pearson correlations between praise and criticism PeerJudge scores and reviewer decisions for non-empty reviews on version 1 Research Articles, with standard sentences removed.

Dataset**	**Reject**	**Revise**	**Accept**	**Total**	**Corr. Pos.**	**Corr. Neg.**	**Acc %***	**Base %***	**Bin acc %^**	**Bin base %^**
F1R eval	124	706	948	1778	0.045	0.410	60.6	53.3	67.5	53.3
Wellcome	14	210	339	563	0.094	0.357	60.7	60.2	65.4	60.2
Gates	7	69	93	169	-0.025	0.383	59.5	57.1	65.6	57.1
HRB	0	12	16	28	0.035	0.417	60.7	57.1	67.9	57.1
F1R dev+	*34*	*199*	*267*	*500*	*0.201*	*0.479*	*63.0*	*53.4*	*69.6*	*53.4*

+The F1000Research development set is included for context only.

* Accuracy is the percent of correct predictions from criticism, counting -1, -2 as Accept, -3 as Revise and -4, -5 as Reject; the baseline accuracy is always predicting Accept.

^ Binary accuracy is the percent of correct predictions from criticism, counting -1, -2 as Accept and -3, -4, -5 as Reject/Revise; the baseline accuracy is always predicting Accept.

** Accept is mapped to (-1,4), revise is mapped to (-2,2) and Reject is mapped to (-4,1) for correlations.

Since the negative correlation is much higher than the positive correlation and explains 5.7 times more of the data on the development data set (comparing squared correlations; 83 times more on the evaluation data set), only the negative sentiment was used to predict peer review outcomes. The intuitive rationale for this is that reviewers often try to report something positive about a rejected paper to soften the blow of a negative outcome. It is not known whether this would be more prevalent for open reviews.

PeerJudge criticism scores can be exploited to make reviewer predictions. From the design of the experiment, the natural mapping of PeerJudge scores to review predictions is, -5, -4: Reject; -3: Revise; -1: Accept. The mapping for -2 is unclear since no decision maps to this. From the development data set, the best mapping of -2 is to an Accept decision, since this gives the highest precision (55.5%, as shown in Table 2). The following procedure works best for this on the development dataset.

- PeerJudge: -1, -2; Prediction: Accept.
- PeerJudge: -3; Prediction: Revise.
- PeerJudge: -4, -5; Prediction: Reject.

Based on this procedure (mapping shown in Table 3), PeerJudge makes the correct recommendation a clear majority of the time (60.6%) on the F1000Reseach evaluation data set. This is only 7.3% better than the baseline strategy of always predicting Accept, which is correct 53.3% of the time (Table 1). Thus, PeerJudge offers low performance at the task of predicting reviewer judgements from their reviews. Because of the scarcity of Reject decisions, PeerJudge predictions could be made more accurate by only predicting Reject for -5 scores and predicting Revise for a score of -4. With this adjustment, the overall accuracy would increase from 60.6% to 63.7%, which is 10.4% higher than the baseline (always predicting Accept).

Table 2. PeerJudge scores against reviewer judgements on the **development data** set of F1000Research reviews (included for context only).

PeerJudge	**Reject**	**Revise**	**Accept**	**Reviews**	**Predict**	**Correct**	**Wrong**	**Precision**
-4	7	10	4	21	Reject	7	14	33.3%
-3	23	100	55	178	Revise	100	78	56.2%
-2	4	73	96	173	Accept	96	77	55.5%
-1		16	112	128	Accept	112	16	87.5%
Reviews	**34**	**199**	**267**	**500**	**Baseline**			53.4%

Table 3. Confusion matrix comparing PeerJudge and reviewer scores on the **evaluation data** set of F1000Research reviews on version 1 Research Articles.

PeerJudge	**Reject**	**Revise**	**Accept**	**Reviews**	**Predict**	**Correct**	**Wrong**	**Precision**
-5	2	0	0	2	Reject	2	0	100.0%
-4	20	51	25	96	Reject	20	76	20.8%
-3	73	402	270	745	Revise	402	343	54.0%
-2	21	215	321	557	Accept	321	236	57.6%
-1	8	38	332	378	Accept	332	46	87.8%
Reviews	124	706	948	1778	**Baseline**			53.3%

PeerJudge criticism scores can also be mapped to **binary** reviewer predictions about whether an article will be accepted, by merging the Reject and Revise categories as follows.

- PeerJudge: -1, -2; Prediction: Accept.
- PeerJudge: -3, -4, -5; Prediction: Reject or Revise.

Based on this procedure, PeerJudge makes the correct binary (Accept vs. Reject or Revise) recommendation just over two thirds of the time (67.5%) on the F1000Reseach evaluation data set. This is 14.2% better than the baseline strategy of always predicting Accept (Table 5, see also Table 4). PeerJudge therefore arguably offers moderate performance at the task of assessing whether a review text is associated with an immediate Accept decision.

Table 4. PeerJudge scores against reviewer judgements on the **development data** set of F1000Research reviews on version 1 Research Articles (included for context only).

PeerJudge	**Reject**	**Revise**	**Accept**	**Reviews**	**Predict**	**Correct**	**Wrong**	**Precision**
-4	7	10	4	21	Rej/rev	17	4	80.9%
-3	23	100	55	178	Rej/rev	123	55	69.1%
-2	4	73	96	173	Accept	96	77	55.5%
-1		16	112	128	Accept	112	16	87.5%
Reviews	**34**	**199**	**267**	**500**	**Baseline**			53.4%

Table 5. Confusion matrix comparing PeerJudge and reviewer scores on the **evaluation data** set of F1000Research reviews on version 1 Research Articles.

PeerJudge	**Reject**	**Revise**	**Accept**	**Reviews**	**Predict**	**Correct**	**Wrong**	**Precision**
-5	2	0	0	2	Rej/rev	2	0	100.0%
-4	20	51	25	96	Rej/rev	71	25	74.0%
-3	73	402	270	745	Rej/rev	475	343	58.1%
-2	21	215	321	557	Accept	321	236	57.6%
-1	8	38	332	378	Accept	332	46	87.8%
Reviews	124	706	948	1778	**Baseline**			53.3%

More detailed predictions

From the development (Table 2, 4) and (checked with the) evaluation (Table 3, 5) data sets of F1000Research reviews on version 1 Research Articles, the following can be deduced by analysing the Precision columns.

- -5 reliably predicts a Reject decision (but based on almost no data; it is a rare score)
- -4 unreliably predicts Revise or Reject but has moderate power as a Revise/Reject prediction
- -3 unreliably predicts Revise or Revise/Reject
- -2 unreliably predicts Accept
- -1 reliably predicts Accept

Thus, given that -5 is a rare score, the only reliable score is -1, which predicts Accept. This suggests that i**f a review contains no criticism then the decision is very likely to be Accept, otherwise the judgement is difficult to predict**.

The above conclusions are consistent with the results from the other three publishing platforms for standard three class (Appendix, Tables A1-A3) and binary (Appendix, Tables A4-A6) predictions.

Other document article types in F1000Research

Although this paper focuses on Research Articles, all other types were also investigated to get insights into whether the prediction accuracy using PeerJudge might be similar (Table 6). The F1000Research reviewer guidelines and criteria vary by type to ensure that reviews are relevant. For example, F1000Research Software Tool Articles ask reviewers to evaluate whether sufficient code details have been provided to allow replication.

Ignoring the results for article types with less than 100 reviews, which are less reliable, the results confirm that praise scores are largely irrelevant to the overall judgement (low correlations). The criticism correlations are weaker than for F1000Research Research Articles, suggesting that some alternative terminology may be used for non-standard article types and where peer reviewers may be guided to check specific things. For example, software reviews included evaluative phrases that seem tailored to the output like, “is very useful for”, “miss a step [] on how to do this”, “potential of being highly utilized”, “work is still very much in the 'prototype' phase”.

Table 6. Data size, accuracy, and Pearson correlations between praise and criticism PeerJudge scores and reviewer decisions for non-empty reviews on version 1 documents from F1000Research, with standard sentences removed. Types are in decreasing order of sample size.

Article type**	Reject	Revise	Accept	Total	Corr. Pos.	Corr. Neg.	Acc %*	Base %*	Bin acc %^	Bin base %^
Case Report	36	179	334	549	0.143	0.353	63.4	60.8	67.2	60.8
Software Tool	26	227	293	546	-0.022	0.369	61.5	53.7	67.0	53.7
Opinion Article	21	153	346	520	-0.127	0.372	63.7	66.5	68.5	66.5
Research Note	45	209	252	506	0.013	0.356	57.5	49.8	64.8	51.2
Method Article	16	126	199	341	0.120	0.388	63.0	58.4	69.2	58.4
Review	6	67	203	276	0.044	0.404	71.4	73.6	73.2	73.6
Data Note	4	28	62	94	-0.081	0.371	69.1	66.0	72.3	66.0
Systematic Review	4	36	51	91	-0.066	0.394	61.5	56.0	65.9	56.0
Short Research Article	11	25	52	88	0.257	0.371	54.5	59.1	62.5	59.1
Correspondence	2	11	70	83	0.202	0.191	54.2	84.3	57.8	84.3
Study Protocol	1	20	51	72	-0.011	0.428	75.0	70.8	75.0	70.8
Commentary	2	8	43	53	-0.094	0.305	73.6	81.1	77.4	81.1
Clinical Practice Article	3	15	34	52	0.188	0.434	71.2	65.4	76.9	65.4
Web Tools	2	15	34	51	-0.051	0.206	49.0	66.7	54.9	66.7
Observation Articles	2	16	19	37	0.007	0.570	73.0	51.4	75.7	51.4
Antibody Validation Article	0	10	22	32	0.271	0.323	65.6	68.8	100.0	68.8
Data Article	2	3	20	25	0.048	0.264	64.0	80.0	72.0	80.0

* Accuracy is the percent of correct predictions from criticism, counting -1, -2 as Accept, -3 as Revise and -4, -5 as Reject; the baseline accuracy is always predicting Accept.

^ Binary accuracy is the percent of correct predictions from criticism, counting -1, -2 as Accept and -3, -4, -5 as Reject/Revise; the baseline accuracy is always predicting Accept.

** Accept is mapped to (-1,4), revise is mapped to (-2,2) and Reject is mapped to (-4,1) for correlations.

Detailed results for the six document types with at least 100 articles (Appendix, Tables A7-A11) confirm the universal high precision of the -1 score as an Accept prediction. The partial exception is for Research Notes (Appendix, Table A9). Its precision is still high at 74% but is substantially lower than for the other document types.

Error analysis

This section examines the least reliable predictions on the evaluation data set of Research Articles and summarises their common causes.

Reviewer Reject decision but Peer Review criticism score of -1 (Accept): This occurred eight times in the F1000Research evaluation data set. The list below shows relevant extracts from the reviews, with opinion terms or phrases highlighted. The common themes are (a) rare or specialist ways of expressing criticism and (b) contextual polysemy in the sense of key terms being not criticism in other contexts (e.g., “problem”). Some of these terms might be added to make an improved version of PeerJudge (e.g., lack, lacks, insufficient, not justified, disappointing, could not find) but others could not be added (e.g., issue, problem, artificial, limitations). Testing would be needed to check whether additions would improve accuracy, however. For example, “problem” would almost certainly reduce accuracy if added to the dictionary as a criticism term because it can be used as a neutral term. Similarly, “pity” or “disappointing” could be used as a comment on an aspect of the paper (criticism) or the results (neutral). Other phrases are rare enough to make little difference overall, such as “would be more pertinent”.

- "**failed** to show a clear break in tolerance"
- "it **lacks** the basics of experimental design", "The **major issue** was no replicates", "All results, discussion, and conclusion were from one sample which does **not make much sense** in a scientific publication"
- "The conclusion is **not justified**", "**insufficient** detail", "**No detail** provided on level of care"
- "this report does **not contain sufficient evidence** for the conclusion"
- "found some **issues** with the data", "**would be more pertinent** to look at", " the differences in the two groups is **artificial** because the general management differs"
- "the final discussion section is **extremely disappointing**", "**no attempt** by the authors to add much value to the rather **fragmented results** found through the review", "Part of the **problem** is that the characteristics listed are treated as equally weighted"
- "mainly described the results rather than mechanisms", "there was **lack** of direct evidence", "**did not show** why they did this work"
- "I **could not find** the information concerning the type of the specimen", "Certain **limitations** of the study were visible", "Significant proportion of EBV infected patient may be asymptomatic", "it is a **pity** that data concerning clinical symptoms are **not mentioned**"

Reviewer Accept decision but Peer Review score of -4 (Revise): This occurred 25 times in the F1000Research evaluation data set. The list below shows relevant extracts from the reviews, with opinion terms or phrases highlighted, and relevant criticism scores annotated. The causes include (a) the criticism being expressed about prior research or the current state of knowledge rather than the reviewed paper (b) PeerJudge not identifying a negation because of intervening terms, (c) the criticism being arguably minor, often about a minor aspect of the clarity of the work or small flaws, (d) complex sentence construction, and (e) polysemy. These are difficult issues to develop rules to deal with. Despite the repeated occurrence of the words meaning *flawed* or *clear* in these examples, their strengths were validated on the development data set so there are many more examples where their values support correct predictions.

Criticism not about the reviewed Research Article

- "The article reviews a previous publication that was based on two **flawed**[-4] experiments"
- "The authors argue the imputation method used in the original article is **flawed**[-4]"
- "attempting to address the **flaws**[-4] that the deficit model in communication may pose"
- Quote, 'This new student/technician is really **terrible**, he certainly confused the labels on the cages/test tubes!'. [quote to illustrate the reviewer's point]
- "whose diversity is **poorly**[-4] known"
- "established cell lines are likely to **poorly**[-4] represent the heterogenous nature of the disease"
- "in software and processing options **poorly**[-4] described in Lin et al"
- "on the other hand, those which are **poorly**[-4] managed and run"
- "which has been **poorly**[-4] addressed from a public health research perspective"

Negation not identified

- "there are no major **flaws**[-4] or concerns"
- "there is no big **mistake**[-3] or major **flaw**[-4] in the introduction"

Apparently minor criticism about clarity, novelty or flaws

- "the justification for screening just amines was **not really clear [-4]** [*negated]"
- "result section of the abstract is **not very clear [-4]** [*negated]"
- "It is furthermore **not very clear**[-4] [*negated] to me what is meant exactly by"
- "it is **not completely clear**[-4] [*negated] why"
- "but it isn't **clear**[-4] [*negated] to which variable this test"
- "in Figure 3 it **isn't clear**[=4] [*negated] to me why"
- "it is **not completely clear**[-4] [*negated] whether cells are filtered out only if they have"
- "but this is **not novel**[-4] [*negated] and the technique does not add further evidence"
- "Construction of simple cellular models to capture biological characteristics of individual neurons is **not novel**[=4] [*negated]"
- "Despite the article having a few **flaws**[-4]"
- "The **flaws**[-4] of the 'in -vitro' study of the retention of a thermoplastic oral device should, in my opinion, be also discussed"

Complex sentence construction

- "The focus of the manuscript is **not novel**[-4] [*negated] aspects of the membrane organisation, it is rather the methodology"
- "Whilst the individual elements of the work are **not novel**[-4] [*negated] […], the combination"

Polysemy (sentiment term in a non-criticism context)

- "**opaque**[-4] syringes and lines"

International Conference on Learning Representations (ICLR)

AI software developed on text from one source may not perform well on the same type of text from a different source due to differences in style or terminology. It is therefore important to evaluate PeerJudge on other types of reviews to test the extent to which it will need to be customised for different publishing systems or contexts. PeerJudge was therefore applied to reviews from ICLR 2017 (n=1511 reviews) and ICLR 2018 (n=2748 reviews) to test it in a different open peer review setting, and as previously assessed (Wang and Wan, 2018). ICLR is a computer science conference related to deep learning that uses an open review system (openreview.net) publishing reviewer scores, review text and author responses. It is similar to F1000Research except that reviewers are anonymous, scores are given on a numerical scale of 1 ("Trivial or wrong") to 10 ("Top 5% of accepted papers, seminal paper"), there is a topic focus on computer science, and the objective is to select the best papers for oral presentation at the conference, so it is a competitive process.

PeerJudge has weak results on reviews from both years of ICLR. The praise correlation is stronger and the criticism correlation is weaker than for F1000Research (Table 7, rows 2 and 3). The criticism correlation is again stronger than the praise correlation, but the difference is small for ICLR. The highest correlation, although only weak to moderate, is the sum of the praise and criticism scores to give an overall score that is positive if the praise score is higher than the criticism score, or negative if it is the other way round. Overall, however, PeerJudge gives only weak predictive power on these reviews.

Table 7. Data size, accuracy, and Pearson correlations between praise and criticism PeerJudge scores and reviewer decisions for non-empty reviews on ICLR reviews.

Dataset	**Min**	**Max**	**Average**	**Reviews**	**Corr. Pos.**	**Corr. Neg.**	**Corr. Pos. + Neg.**
ICLR 2017	1	9	5.67	1511	0.148	0.240	0.287
ICLR 2018	1	9	5.43	2748	0.113	0.196	0.237
*ICLR 2017**	*1*	*9*	*5.67*	*1511*	*0.169*	*0.256*	*0.316*
*ICLR 2018**	*1*	*9*	*5.43*	*2748*	*0.080*	*0.185*	*0.215*

* Scores from the version of PeerJudge optimised on ICLR 2017.

An error analysis of articles with the lowest score (1,2) but no criticism detected by PeerJudge (-1) found that criticism was expressed in complex or indirect language. Examples include, "There is absolutely nothing of interest to ICLR except []", "The paper has little to no content", "Such systems have been [] investigated", and "If this was the case then paragraph vectors wouldn't work when representing a whole document ,which it already does as can be seen in table 2".

An error analysis of articles with high scores (8,9,10) but strong criticism detected by PeerJudge (-4,-5) found similar problems as for F1000Research, such as, "It was not very clear[-4] to me if the baseline decoders [] are fairly compared here", and "performs surprisingly poorly[-4] in most examples".

A new version of PeerJudge was made by the optimisation process described above for F1000Research (optimising term weights and adding extra term) using ICLR 2017 as the development set and ICLR 2018 as the test set, but this did not improve its performance (Table 7, rows 4 and 5). The root cause of the difficulty seems to be the use of relatively technical or complex phrases, even when passing overall judgements. For example, reports on papers scoring 9 out of 10 had a clear final sentence judgement, but expressed in complex language, “Overall, this is a very well-written paper that creatively combines a number of interesting ideas to address an important problem” and “As adversarial training is an important topic for deep learning, I feel this work may lead to promising principled ways for adversarial training”. Another review with a 9 score had a main judgement sentence that is more moderate, “it's nice to see it applied in the context of modern deep learning architectures, and the analysis of the results is very interesting.” The root cause of the problem may be the complexity of the highly technical topic area of the conference so that simple judgement statements in reviews are rare.

Limitations and discussion

The performance scores for PeerJudge should not be interpreted as being the maximum possible for AI on open peer reviews. This is because an alternative algorithm may have performed better, so its performance is a lower bound for the potential for lexical analyses of open non-anonymous reviews. It is not reasonable to try a generic machine learning algorithm on this multidisciplinary dataset, however, since this may learn *topics* that tend to attract good or bad outcomes, rather than focusing on the judgement language. As an extreme example to illustrate, if a generic machine learning algorithm was fed with 100 reviews of cancer research that were rejected and 100 reviews of politics research that were accepted then it would learn that cancer research should always be rejected, and politics research should always be accepted (100% accurate; 0% insights). In a multidisciplinary collection like that of F1000, a generic machine learning algorithm risks detecting topics that tend to attract favourable or negative reviews, even if the outcomes are not as clear cut as the above simplistic cancer/politics example. This issue can be resolved by ensuring that the input texts are balanced – if the input was 50 accept and 50 accept for both cancer and politics then the research topic could no longer be used to predict the outcome. For this to work, the reviews would need to be categorised by topic/field and then equal numbers of each category selected for each field. The F1000Research articles have keywords rather than subjects so this is not practical. This issue would not apply to subject-specific convergences or narrow journals.

The PeerJudge vocabulary contains 78 criticism terms and 45 paradise terms. These are relatively small numbers, possibly balancing the output towards criticism (although positive comments seem to be expressed using less varied terminology). Its performance may increase with additional corpora of open peer reviews to allow additional common evaluative terms to be identified.

PeerJudge has been developed on open peer review reports with the F1000Research three outcome model and may perform less well on closed peer review, single- or double-blinded open peer review and with the more standard four outcomes (accept, minor revisions, major revisions, reject) or conference outcomes (accept, reject). It seems likely that it would perform reasonably well on some topics because the reports tend to be structured like closed peer review reports, few (under 5%) take advantage of open peer review to allude to other reviewers’ comments, and most reviewers have probably written closed peer review reports in the past and may therefore largely harness their prior experience and working patterns. It seems reasonably likely that reviewers would be more generous in open peer review than in closed peer review, however, given the potential damage to personal relationships from public criticism. On the other hand, failure to identify problems in an article might damage the reputation of a reviewer, and so the tendency towards generosity in open peer review is not clear. This aligns with previous research, as reviewed above, which has not found consistent differences in the quality of open and closed reviews. The ICLR example illustrates that PeerJudge may perform substantially worse in some fields, however.

A practical limitation for detecting praise is that standard review judgement sentences were removed from each review, so a reviewer might see this text as enough to explain an Accept decision, and that explicit praise or a summary of how the article is valuable would be unnecessary. Similarly, answers to mandatory peer review questions were removed from the reviews, which may also have been used to justify an Accept, Revise, or Reject decision without expanding upon these in the body of the review text. Praise estimation may therefore be more accurate on other types of reviews without the requirement to include standard texts, or separate reviewer questions.

Conclusions

The PeerJudge software can predict peer reviewer decisions from the narrative included in their reports with a moderate degree of accuracy. Its accuracy was highest for traditional Research Articles on the F1000Research publishing platforms compared to others (e.g. software tool, data paper), although this may have been due to the development set only including Research Articles. Its prediction accuracy is high for reviews lacking criticism but lower for reviews with a moderate amount of criticism. The difficulty in predicting judgements for reviews with a moderate amount of criticism is a preliminary finding (and related to a similar difficulty for machine learning: Wang and Wan, 2018) because future research may find alternative approaches to tackle this issue. This data shared online may be useful for this (10.6084/m9.figshare.9946748).

The following conclusions about praise and criticism found by PeerJudge can be drawn from the analysis.

- **Praise** in F1000Research, Wellcome Open Research, Gates Open Research and HRB Open Research reviews of all submission types are largely irrelevant to reviewer judgements. This raises the possibility that an open peer review system promotes polite or constructive reviews.
- The **absence of criticism** is a strong indicator that the reviewer judgement will be 'Approve' (possibly with limited suggested changes) for a document published on F1000Research, Wellcome Open Research, Gates Open Research and HRB Open Research.
- In the **presence of some criticism**, all outcomes are reasonably likely (although Reject is rare overall): 'Approve', 'Approve with reservations', or 'Not approved'. It is possible that the difficulty found by PeerJudge in making this decision is a *technical problem of language complexity* and a deeper linguistic analysis would be required to consider how and where expressions of praise and criticism occur within a peer review report. It is also possible that these boundaries are social rather than technical, meaning that the reviewer chooses a *decision based on human factors* (e.g., avoiding publicly offending the authors; wishing to encourage new researchers) rather than purely on the merits of the paper.

Practical applications and implications

PeerJudge is available online at http://sentistrength.wlv.ac.uk/PeerJudge.html and can be supplied as a (fast: 1,000 reviews per second) Java application for research.

Anomalous judgement detection: PeerJudge's predictions might be used to prompt reviewers submitting apparently anomalous reviews. For example, if PeerJudge scores a review +1,-4 but the reviewer registers Accept, they might be asked to double-check if their criticisms warrant this outcome.

Politeness warnings: If a reviewer posts a Reject or Revise decision and PeerJudge identifies no praise, the reviewer might be prompted to check that they have said something positive to encourage the author.

Journal reviewing consistency checks: PeerJudge's predictions might be used to calibrate the relationship between reviewer comments and judgements between journals or publishing platforms, to check if the peer review process is consistent between them. This could also be used to compare different subsets, such as between reviewers from different countries (although language problems would complicate this). For this, the lack of accuracy is not important because errors will tend to cancel out over large datasets.

Concern about reviewer judgements: The inability to predict the reviewer judgement in the presence of any amount of detected criticism raises the *possibility* that judgements are not made solely on the weight of evidence but also based on human factors. This possibility seems important to be aware of, and consider potential solutions for, because it could fundamentally undermine the open reviewing model. This does not imply that closed peer review is less biased since reviewers in the closed review system might hide behind anonymity to sabotage rivals, support friends, or exhibit other prejudices. In the open system, any biases are at least transparent and may be evaluated by article readers and other reviewers.

Acknowledgements

This project was funded by Jisc (www.jisc.ac.uk). Jisc suggested the idea for automatic peer review of open reviews, suggested F1000Research as the test dataset and helped to write this article. **Competing interests**: Three authors are currently employed by F1000Research Ltd. One author is employed by Jisc, which funded the project, enlisted advice from F1000Research and contributed to the write-up. The data analysis was conducted independently by a fifth author.

References

Bravo, G., Grimaldo, F., López-Iñesta, E., Mehmani, B., and Squazzoni, F. (2019), "The effect of publishing peer review reports on referee behavior in five scholarly journals", *Nature Communications*, Vol. 10 No. 1, pp. 322.

Caciagli, A. (2019), "PeerTax: Investigating the taxonomy of peer reviews", *eLife* , available at: https://elifesciences.org/labs/abb00264/peertax-investigating-the-taxonomy-of-peer-reviews (accessed 5 November 2019).

Clarivate (2018), "2018 global state of peer review", available at: https://publons.com/static/Publons-Global-State-Of-Peer-Review-2018.pdf (accessed 5 November 2019).

Emerson, G. B., Warme, W. J., Wolf, F. M., Heckman, J. D., Brand, R. A., and Leopold, S. S. (2010), "Testing for the presence of positive-outcome bias in peer review: a randomized controlled trial", *Archives of Internal Medicine*, Vol. 170 No. 21, pp. 1934-1939.

F1000Research (2019a), "FAQs", available at: https://f1000research.com/faqs (accessed 5 November 2019).

F1000Research (2019b), "Reviewer guidelines", available at: https://f1000research.com/for-referees/guidelines (accessed 5 November 2019).

Ford, E. (2013), "Defining and characterizing open peer review: A review of the literature", *Journal of Scholarly Publishing*, Vol. 44 No. 4, pp. 311-326.

Frontiersin (2018), "AI-enhanced peer review: Frontiers launches next generation of efficient, high-quality peer review, available at: https://blog.frontiersin.org/2018/12/14/artificial-intelligence-peer-review-assistant-aira/ (accessed 5 November 2019).

Ghosal, T., Verma, R., Ekbal, A., and Bhattacharyya, P. (2019), "DeepSentiPeer: Harnessing Sentiment in Review Texts to Recommend Peer Review Decisions", In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, New York, NY: ACL Press, pp. 1120-1130.

Heaven, D. (2018), "AI peer reviewers unleashed to ease publishing grind", *Nature News*, available at: https://www.nature.com/articles/d41586-018-07245-9 (accessed 5 November 2019).

Jefferson, T., Alderson, P., Wager, E., and Davidoff, F. (2002), "Effects of editorial peer review: a systematic review", *Jama*, Vol. 287 No. 21, pp. 2784-2786.

Jefferson, T., Wager, E., and Davidoff, F. (2002), "Measuring the quality of editorial peer review", *Jama*, Vol. 287 No. 21, pp. 2786-2790.

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", In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Vol. 1 (pp. 1647-1661).

Kowalczuk, M. K., Dudbridge, F., Nanda, S., Harriman, S. L., Patel, J., and Moylan, E. C. (2015), "Retrospective analysis of the quality of reports by author-suggested and non-author-suggested reviewers in journals operating on open or single-blind peer review models", *BMJ Open*, Vol. 5 No. 9, pp. e008707.

Langford, J., and Guzdial, M. (2015), "The arbitrariness of reviews, and advice for school administrators", Communications of the ACM, Vol. 58 No. 4, pp. 12-13.

Leijen, D. A. (2014), "Applying machine learning techniques to investigate the influence of peer feedback on the writing process", in *Methods in Writing Process Research*, Frankfurt Am Main: Peter Lang, pp. 167-183.

Liu, B. (2012), "Sentiment analysis and opinion mining", *Synthesis Lectures on Human Language Technologies*, Vol. 5 No. 1, pp. 1-167.

Morey, R. D., Chambers, C. D., Etchells, P. J., Harris, C. R., Hoekstra, R., Lakens, D., and Vanpaemel, W. (2016), "The peer reviewers' openness initiative: Incentivizing open research practices through peer review", *Royal Society Open Science*, Vol. 3 No. 1, pp. 150547. https://doi.org/10.1098/rsos.150547

Moylan, E. C., Harold, S., O'Neill, C., and Kowalczuk, M. K. (2014), "Open, single-blind, double-blind: which peer review process do you prefer?", *BMC Pharmacology and Toxicology*, 15, pp. 55. doi:10.1186/2050-6511-15-55

Mrowinski, M. J., Fronczak, P., Fronczak, A., Ausloos, M., and Nedic, O. (2017), "Artificial intelligence in peer review: How can evolutionary computation support journal editors?", *PloS ONE*, Vol. 12 No. 9, e0184711.

Mulligan, A., Hall, L., and Raphael, E. (2013), "Peer review in a changing world: An international study measuring the attitudes of researchers", *Journal of the American Society for Information Science and Technology*, Vol. 64 No. 1, pp. 132-161.

Priatna, W. S., Manalu, S. R., and Sundjaja, A. M. (2017), "Development of review rating and reporting in open journal system", *Procedia Computer Science*, Vol. 116, pp. 645-651.

Price, S., and Flach, P. A. (2017), "Computational support for academic peer review: a perspective from artificial intelligence", *Communications of the ACM*, Vol. 60 No. 3, pp. 70-79.

Ross-Hellauer, T., Deppe, A., and Schmidt, B. (2017), "Survey on open peer review: Attitudes and experience amongst editors, authors and reviewers", *PloS One*, Vol. 12 No. 12, e0189311.

Ross-Hellauer, T. (2017), "What is open peer review? A systematic review", *F1000Research*, Vol. 6, 588. doi:10.12688/f1000research.11369.2

Sizo, A., Lino, A., Reis, L. P., and Rocha, Á. (2019), "An overview of assessing the quality of peer review reports of scientific articles", *International Journal of Information Management*, Vol. 46 No. 1, pp. 286-293.

Somasundaran, S., Ruppenhofer, J., and Wiebe, J. (2007), "Detecting arguing and sentiment in meetings". In Proceedings of the SIGdial Workshop on Discourse and Dialogue (Vol. 6).

Steinwart, I., and Christmann, A. (2008). "Support vector machines". Berlin, Germany: Springer Science & Business Media.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., and Stede, M. (2011), "Lexicon-based methods for sentiment analysis", *Computational Linguistics*, Vol. 37 No. 2, pp. 267-307.

Thelwall, M., Buckley, K., and Paltoglou, G. (2012), "Sentiment strength detection for the social web", *Journal of the American Society for Information Science and Technology*, Vol. 63 No. 1, pp. 163-173.

Xiong, W., and Litman, D. (2011), "Automatically predicting peer-review helpfulness", In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2 (pp. 502-507). New York, NY: Association for Computational Linguistics.

Van Rooyen, S., Godlee, F., Evans, S., Black, N., and Smith, R. (1999), "Effect of open peer review on quality of reviews and on reviewers' recommendations: a randomised trial", *Bmj*, Vol. 318 No. 7175, pp. 23-27.

Van Rooyen, S., Delamothe, T., and Evans, S. J. (2010), "Effect on peer review of telling reviewers that their signed reviews might be posted on the web: randomised controlled trial", *Bmj*, Vol. 341, c5729.

Walsh, E., Rooney, M., Appleby, L., and Wilkinson, G. (2000), "Open peer review: a randomised controlled trial", *The British Journal of Psychiatry*, Vol. 176 No. 1, pp. 47-51.

Wang, K., and Wan, X. (2018). Sentiment analysis of peer review texts for scholarly papers", In *41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, New York, NY: ACM Press, pp. 175-184

Yadav, R.K., and Gehringer, E.F. (2016), "Metrics for automated review classification: What review data show", In *State-of-the-Art and Future Directions of Smart Learning,* Springer, Singapore, pp. 333-340.

Appendix

Table A1. Confusion matrix comparing PeerJudge and reviewer scores on the **Wellcome Open Research** reviews on version 1 Research Articles. The baseline overall precision is 60.2% (predicting that all decisions are Accept).

PeerJudge	Reject	Revise	Accept	Reviews	Predict	Correct	Wrong	Precision
-4	3	16	11	30	Reject	3	27	10.0%
-3	10	131	120	261	Revise	131	130	50.2%
-2	1	52	117	170	Accept	117	53	68.8%
-1		11	91	102	Accept	91	11	89.2%
Reviews	14	210	339	563	**Baseline**			

Table A2. Confusion matrix comparing PeerJudge and reviewer scores on the **Gates Open Research** reviews on version 1 Research Articles. The baseline overall precision is 57.1% (predicting that all decisions are Accept).

PeerJudge	Reject	Revise	Accept	Reviews	Predict	Correct	Wrong	Precision
-4	1	6	3	10	Reject	1	9	10.0%
-3	4	36	30	70	Revise	36	34	51.4%
-2	2	21	38	61	Accept	38	23	62.3%
-1			22	22	Accept	22	0	100.0%
Reviews	7	63	93	163	**Baseline**			

Table A3. Confusion matrix comparing PeerJudge and reviewer scores on the **HRB Open Research** reviews on version 1 Research Articles. The baseline overall precision is 57.1% (predicting that all decisions are Accept).

PeerJudge	Reject	Revise	Accept	Reviews	Predict	Correct	Wrong	Precision
-4	0	2	1	3	Reject	0	3	0.0%
-3	0	7	5	12	Revise	7	5	58.3%
-2	0	3	6	9	Accept	6	3	66.7%
-1			4	4	Accept	4	0	100.0%
Reviews	**0**	12	16	28	**Baseline**			

Table A4. Confusion matrix comparing PeerJudge and reviewer scores on the **Wellcome Open Research** reviews on version 1 Research Articles for a **binary** decision. The baseline overall precision is 60.2% (predicting that all decisions are Accept).

PeerJudge	Reject	Revise	Accept	Reviews	Predict	Correct	Wrong	Precision
-4	3	16	11	30	Rej/rev	19	11	63.3%
-3	10	131	120	261	Rej/rev	141	120	54.0%
-2	1	52	117	170	Accept	117	53	68.8%
-1		11	91	102	Accept	91	11	89.2%
Reviews	14	210	339	563	**Baseline**			

Table A5. Confusion matrix comparing PeerJudge and reviewer scores on the **Gates Open Research** reviews on version 1 Research Articles for a **binary** decision. The baseline overall precision is 57.1% (predicting that all decisions are Accept).

PeerJudge	Reject	Revise	Accept	Reviews	Predict	Correct	Wrong	Precision
-4	1	6	3	10	Rej/rev	7	3	70.0%
-3	4	36	30	70	Rej/rev	40	30	57.1%
-2	2	21	38	61	Accept	38	23	62.3%
-1			22	22	Accept	22	0	100.0%
Reviews	7	63	93	163	**Baseline**			

Table A6. Confusion matrix comparing PeerJudge and reviewer scores on the **HRB Open Research** reviews on version 1 Research Articles for a **binary** decision. The baseline overall precision is 57.1% (predicting that all decisions are Accept).

PeerJudge	Reject	Revise	Accept	Reviews	Predict	Correct	Wrong	Precision
-4	0	2	1	3	Rej/rev	2	1	66.6%
-3	0	7	5	12	Rej/rev	7	5	58.3%
-2	0	3	6	9	Accept	6	3	66.7%
-1			4	4	Accept	4	0	100.0%
Reviews	**0**	12	16	28	**Baseline**			

Table A7. Confusion matrix comparing PeerJudge and reviewer scores on the F1000Research **Software Tool** version 1 reviews.

PeerJudge	Reject	Revise	Accept	Reviews	Predict	Correct	Wrong	Precision
-4	3	15	13	31	Reject	3	28	10%
-3	15	142	89	246	Revise	142	104	58%
-2	5	55	80	140	Accept	80	60	57%
-1	3	15	111	129	Accept	111	18	86%
Reviews	**26**	**227**	**293**	**546**	**Baseline**			

Table A8. Confusion matrix comparing PeerJudge and reviewer scores on the F1000Research **Review** version 1 reviews.

PeerJudge	**Reject**	**Revise**	**Accept**	**Reviews**	**Predict**	**Correct**	**Wrong**	**Precision**
-4	0	1	7	8	Reject	0	8	0%
-3	4	45	44	93	Revise	45	48	48%
-2	2	18	60	80	Accept	60	20	75%
-1	0	3	92	95	Accept	92	3	97%
Reviews	**6**	**67**	**203**	**276**	**Baseline**			

Table A9. Confusion matrix comparing PeerJudge and reviewer scores on the F1000Research **Research Note** version 1 reviews.

PeerJudge	**Reject**	**Revise**	**Accept**	**Reviews**	**Predict**	**Correct**	**Wrong**	**Precision**
-5	0	1	0	1	Reject	0	1	0%
-4	4	13	3	20	Reject	4	16	20%
-3	23	94	56	173	Revise	94	79	54%
-2	13	71	92	176	Accept	92	84	52%
-1	5	30	101	136	Accept	101	35	74%
Reviews	**45**	**209**	**252**	**506**	**Baseline**			

Table A10. Confusion matrix comparing PeerJudge and reviewer scores on the F1000Research **Opinion Article** version 1 reviews.

PeerJudge	**Reject**	**Revise**	**Accept**	**Reviews**	**Predict**	**Correct**	**Wrong**	**Precision**
-4	4	13	7	24	Reject	4	20	17%
-3	12	86	98	196	Revise	86	110	44%
-2	3	41	124	168	Accept	124	44	74%
-1	2	13	117	132	Accept	117	15	89%
Reviews	**21**	**153**	**346**	**520**	**Baseline**			

Table A11. Confusion matrix comparing PeerJudge and reviewer scores on the F1000Research **Method Article** version 1 reviews.

PeerJudge	**Reject**	**Revise**	**Accept**	**Reviews**	**Predict**	**Correct**	**Wrong**	**Precision**
-4	2	12	7	21	Reject	2	19	10%
-3	9	77	56	142	Revise	77	65	54%
-2	4	26	65	95	Accept	65	30	68%
-1	1	11	71	83	Accept	71	12	86%
Reviews	**16**	**126**	**199**	**341**	**Baseline**			