Textual analysis of artificial intelligence manuscripts reveals features associated with peer review outcome.

Philippe Vincent-Lamarre 1,2,*  
Vincent Larivière 1

March 5, 2020

1 École de bibliothéconomie et des sciences de l’information, Université de Montréal, Montréal, Canada
2 École de psychologie, Université d’Ottawa, Ottawa, Canada

* Corresponding author

Abstract

We analysed a dataset of scientific manuscripts that were submitted to various conferences in artificial intelligence. We performed a combination of semantic, lexical and psycholinguistic analyses of the full text of the manuscripts and compared them with the outcome of the peer review process. We found that accepted manuscripts scored lower than rejected manuscripts on two indicators of readability, and that they also used more scientific and artificial intelligence jargon. We also found that accepted manuscripts were written with words that are less frequent, that are acquired at an older age, and that are more abstract than rejected manuscripts. The analysis of references included in the manuscripts revealed that the subset of accepted submissions were more likely to cite the same publications. This finding was echoed by pairwise comparisons of the word content of the manuscripts (i.e. an indicator or semantic similarity), which were more similar in the subset of accepted manuscripts. Finally, we predicted the peer review outcome of manuscripts with their word content, with words related to machine learning and neural networks positively related with acceptance, whereas words related to logic, symbolic processing and knowledge-based systems negatively related with acceptance.

1 Introduction

Peer review is a fundamental component of the scientific enterprise and acts as one of the main source of quality control of the scientific literature [1]. The primary form of peer review occurs before publication
and it is often considered as a stamp of approval from the scientific community [3, 4]. Peer-reviewed publications have a considerable weight in the attribution of research and academic resources [5, 6, 7].

One of the main concern about peer review is its lack of reliability [8, 9, 10]. Most studies on the topic find that agreement between reviewers is barely greater than chance [11, 12, 13], which highlights the considerable amount of subjectivity involved in the process. This leaves room for a lot of potential source of bias, which have been reported in several studies [14, 15, 16]. A potential silver lining is that it appears that the process has some validity. For instance, articles accepted at a general medicine journal [17] and journals in the domain of ecology [18] were more cited than the rejected articles published elsewhere, and the process appears to improve the quality of manuscripts, although marginally [19, 20, 21]. It is therefore surprising that a process that has little empirical support of its effectiveness, but a lot of evidence of its downsides [22] has so much importance.

The vast majority of studies on peer review have focused on the relationship between the socio-demographical attributes of the actors involved in the process and its outcome [23]. Comparatively, little research has focused on the association between the content of the manuscripts and the peer review process. This isn’t surprising given that there is little publicly available datasets of manuscripts annotated as rejected or accepted, and whenever they are made available to researchers it is usually through smaller samples designed to answer specific questions. Another factor contributing to this gap in the literature is that it is more time consuming to analyse textual data (either the referee’s report or the reviewed manuscript) than papers’ metadata. However, the increasing popularity of open access [24, 25] allows for a greater access to the full text of scientific manuscripts.

By scraping the content of arXiv, one of those repositories, [26] developed a new method to identify manuscripts that were accepted at conferences after the peer review process based on submissions around the time of major NLP, machine learning (ML) and artificial intelligence (AI) conferences. These preprints were then matched with manuscripts that were published at the target venues as a way to determine whether they were accepted or "probably-rejected". In addition, the manuscripts and peer-review outcomes were collected from conferences that agreed to share their data. [26] were able to achieve decent accuracy at predicting the acceptance of the manuscripts in their dataset. Other groups were able to obtain good performance at predicting paper acceptance with different machine learning models based on the text of the manuscripts [27], sentiment analysis of referee’s reports [28], or the evaluation score given by the reviewers [29].

In this manuscript, we take advantage of the full text access to those manuscripts and explore linguistic and semantic features that correlates with the peer review outcome. Such features are relevant to two types of biases that could be involved in the peer review process: language and content bias. In the language bias, authors who aren’t native English speakers could receive more negative evaluations due to the linguistic level of their manuscripts [30, 31, 32]. Our understanding of the extent to which such a bias could play
a role in research evaluation is still limited, which is worrying given the increasingly globalized scientific system that relies on one language: English [33]. In terms of content bias, innovative and unorthodox methods are less likely to be judged favourably [14]. This type of bias is also quite likely to play a role in fields that are dominated by a few mainstream approaches such as AI [34]. Conservatism in this field could impede the emergence of breakthrough or novel techniques that don’t fit with the current trends.

In this manuscript, we address both types of biases by comparing the textual data (title, abstract and introduction) of the manuscripts. We first used two readability metrics (the Flesch Reading Ease (FRE) and the New Dale-Chall Readability (NDC) Formula), as well as some indicators of scientific jargon content, and found that manuscripts that were less readable and used more jargon were more likely to get accepted. Accepted and rejected manuscripts were compared on their psycholinguistic and lexical attributes and we found that accepted manuscripts used words that were more abstract, less frequent and acquired at a later age compared to rejected manuscripts. We then compared manuscripts on their word content and their referencing patterns through bibliographic coupling, and found that the subset of accepted manuscripts were semantically closer than rejected manuscripts. Finally, we used the word content of the manuscripts to predict their acceptance, and found that specific topics were associated with greater odds of acceptance.

2 Methods

2.1 Manuscript data

We used the publicly available PeerRead dataset [26] to analyse the semantic and lexical differences between accepted and rejected submissions to some natural language processing, artificial intelligence and machine learning conferences. We therefore used content from six platforms archived in the PeerRead dataset: three arXiv sub-repositories tagged by subject including submissions from 2007 to 2017 (AI: artificial intelligence, CL: computation and language, LG: machine learning), as well as submissions to three other venues: (ACL 2017: Association for Computational Linguistics, CoNLL 2016: Conference on Computational Natural Language Learning, ICLR 2017: International Conference on Learning Representations). This resulted in a dataset with 12,364 submissions. Although the submissions to ACL 2017 and CoNLL 2016 had an acceptance rate in [26], the information for each submission was not available in the dataset at the time of the analysis.

We limited our analysis to the title, abstract and introduction of the manuscripts, because the methods and results contained formulas, mathematical equations and variables, which made them unsuitable for textual analysis.
| Platform   | # Papers | # Accepted |
|------------|----------|------------|
| ICLR 2017  | 427      | 172        |
| ACL 2017   | 137      | -          |
| CoNLL 2016 | 22       | -          |
| arXiv:ai   | 4092     | 418        |
| arXiv:cl   | 2638     | 646        |
| arXiv:lg   | 5048     | 1827       |

Table 1: Number of papers per platform.

2.2 Semantic similarity

The textual data of each article, including the title, abstract, introduction were cleaned by making all words lowercase, eliminating punctuation, single character words and common stopwords. For all analyses except for the readability, scientific jargon and psycholinguistic matching, the stem of the word was extracted using the porter algorithm [35]. We used the Term Frequency Inverse Document Frequency (tf-idf) algorithm to create vectorial representations based on the field of interest (title, abstract or introduction). We then used those vectors to compute the cosine similarity between the pairs of documents.

2.3 Reference matching

In order to obtain manuscript’s bibliographic coupling, we developed a reference matching algorithm because their format was not standardized across manuscripts. We used four conditions to group references together: 1- They were published the same year 2- They had the same number of authors 3- They had a similarity score above 0.7 (empirically determined after manual inspection of matching results) with a fuzzy matching procedure (Token Set Ratio function from the FuzzyWuzzy python library, https://github.com/seatgeek/fuzzywuzzy) on the author’s names and 4- the article’s title.

2.4 Psycholinguistic and readability variables

For the word frequency estimation, we used the SUBTLEXUS corpus [36] from which we used the logarithm of the estimated word frequency + 1. For the concreteness, we used the [37] dataset providing concreteness rating for 40,000 commonly known English words. For the age of acquisition, we used the [38] age of acquisition ratings for 30,000 English words.

We used the readability functions as implemented in [39]. We used the Flesch Reading Ease (FRE; [40, 41]) and the New Dale-Chall Readability Formula (NDC; [42]). The FRE is calculated based on the number of syllables per words and the number of words per sentence. The NDC is based on the number of words per sentence and the proportion of difficult words that are not part of a list of “common words”. We also included two sources of jargon developed by [39]. The first one are science-specific common words, which are words used by scientist which are not in the NDC’s list of common words. The other is the general science jargon, which are words frequently used in science, but aren’t specific to
science (see [39] for methods). Finally, we compiled a list of AI jargon from three online glossaries (https://developers.google.com/machine-learning/glossary/, http://www.wildml.com/deep-learning-glossary/ and https://en.wikipedia.org/wiki/Glossary_of_artificial_intelligence).

2.5 Data analysis

Because of the exploratory nature of the study and the large size of the datasets, null hypothesis significance testing has many shortcomings [43]. In some cases, we performed statistical analysis of the results and reported the p-value, but those results should be interpreted carefully. Our analyses rely on the effect size, as well as the cross-validated effects on the independent subsets of the PeerRead dataset (manuscripts from different venues and online repositories). All error bars represent the standard error of the mean.

2.6 Identification of geographic location

We searched through the email addresses of the authors to identify research within the United-States (US) and outside the US. We considered a manuscript as US based if at least one author had an email address that ended with ".edu".

2.7 Code and data availability

The custom python scripts and the data used to generate the results of this manuscript can be found at https://github.com/lamvin/PeerReviewAI.git.

3 Results

3.1 Readability

The readability of scientific articles has been steadily declining in the last century [39]. One possible explanation for this is that writing more complex sentences and using more scientific jargon increase the likelihood that a manuscript will get accepted at peer review. To investigate this hypothesis, we used two measures of readability on our data: the Flesch Reading Ease (FRE) and the New Dale-Chall Readability Formula (NDC). FRE scores decrease as a function of a ratio of the number of syllables per word and the number of word per sentence. NDC scores increase as a function of the number of words in each sentence and as the proportion of difficult words increase (words that are not present in the NDC list of common words). We also included the proportion of words from a science-specific common words and general science jargon list (constructed by [39]). In order to control for potential demographic confounds, we divided our dataset in the two categories: manuscripts within and outside the United States (US).
We found that both indicators of readability were correlated with the peer review outcome. FRE (higher score = more readable) was lower for accepted manuscripts, while NDC (higher score = less readable) was higher for accepted manuscripts (Fig. 1). This was the case for both US and non-US manuscripts (with no sizeable differences), for every section of the manuscript (title, abstract and introduction) and the effect was replicated within most platforms, except for the introduction with the FRE indicator.

As [39] reported that the proportion of scientific jargon has increased over the last century, we also wondered if the peer review process would reflect this effect. Using a list of general and specific scientific jargon, we found that the manuscripts containing the higher ratios of jargon were associated with higher acceptance rates. However, our list of science jargon was biased towards content from the life sciences. To confirm the relevance of these results to our dataset, we generated an AI jargon list (see Methods). Using this new list, we found a robust effect across platforms and document section, where a larger proportion of AI jargon predicted greater odds of acceptance for the manuscripts.

The replication of our results independently for US and non-US based manuscripts suggest that the effect is not driven by geographic locations. Statistical analysis are summarised in Tables 6 and 7.

3.2 Lexical correlates of peer review outcome

We then investigated the differences between accepted and rejected submissions based on lexical and psycholinguistic attributes. Given that we haven’t found consistent differences between US and non-US based submissions, we pooled all manuscripts together for the rest of the analysis. We used the number of tokens (total number of words in a document) as well as two measures of lexical diversity: the number of types (unique words in a document) and the ratio between the types and token (Type-Token Ratio, TTR). We also used three psycholinguistic variables: the age of acquisition (AOA), concreteness and frequency (on a logarithmic scale). We computed the average values of those psycholinguistic variables on all types and all tokens.

We found consistent effects between the psycholinguistic variables and across platforms and section, with few exceptions. Words used in accepted manuscripts were less frequent, acquired later in life and more abstract than in rejected manuscripts on average (Fig. 2). The effects were consistent across all platforms except ICLR (which is smaller than the other ones).

Interestingly, we found that shorter titles, abstract and introduction were all associated with higher acceptance rates. Unsurprisingly, this translated in a bias towards manuscript with lower total types (for every section) for manuscript acceptance. However, when taking the ratio between the two (TTR, an indicator of lexical richness), we found that this variable was positively associated with manuscript acceptance. Results from the statistical analysis are summarised in Table 8.
Figure 1: Average readability and jargon proportion of US and non-US manuscripts.
3.3 High-level semantic correlates of peer review outcome

3.3.1 Bibliographic coupling and semantic similarity

We then looked at how similar the accepted manuscripts were compared to the rejected ones based on their semantic content. First we looked at the similarity of their title, abstract or introduction based on a tf-idf representation of their word content. Secondly, we looked at their degree of bibliographic coupling. We only compared pairs of manuscripts that shared at least one common reference for the next analysis (see table 2).

| # of common cited references | Frequency |
|-----------------------------|-----------|
| 1                           | 1,619,173 |
| 2                           | 379,161   |
| 3                           | 124,574   |
| 4                           | 47,254    |
| 5-9                         | 36,753    |
| 10-19                       | 2,527     |
| 20-29                       | 230       |
| 30-39                       | 68        |
| 40-49                       | 31        |
| 50-59                       | 11        |
| 60-80                       | 4         |

Table 2: Citation intersection for all papers.

As the two approaches quantify the content similarity of the documents, we wanted to verify whether those two metrics measured different aspects of the document content. It was previously reported that there
is a moderate correlation between the two measures in the field of economics [44]. We correlated the semantic distance with the bibliographic coupling of the document submitted to each platform. We used a semantic distance metric based on the cosine similarity between the tf-idf representation of each document, as well as both the citation intersection (# common references) and the Jaccard similarity coefficient (#references in common/ # references in total) as a measure of bibliographic coupling. We found a moderate correlation (Pearson $r > 0.20$ and $< 0.40$) between both measures of bibliographic coupling and semantic distance when pooling all platforms together depending on what section of the manuscript were compared (Fig. 3). This suggests that those two measures aren’t redundant features of semantic content, and that they might capture different aspects of it. This also validates our algorithm for citation disambiguation as comparable correlations between the bibliographic coupling and textual similarity were reported in [44].

![Figure 3: Correlation between semantic similarity and bibliographic coupling.](image)

### 3.3.2 Bibliographic coupling and peer review outcome

We then looked at how accepted and rejected manuscripts differed based on the characteristics of their cited references (bibliographic coupling). We compared all pairs of manuscripts on the two indicators of bibliographic coupling (intersection and Jaccard index). Each pair of manuscripts was categorized as one of the following: "accepted": the two submissions were accepted, "rejected": the two submissions were rejected, and "mixed", one document was rejected and the other was accepted.

We found that accepted manuscripts had more references in common (Fig. 4) than the two other cat-
egories of manuscripts. The effect was slightly weaker for the Jaccard similarity (intersection over union of citations) and less consistent across platforms than the intersection. However, both metrics account for about 0.2% of the variance (All platforms, Jaccard: 0.228% and intersection: 0.21%).

![Bibliographic coupling between accepted and rejected manuscripts.](image)

**Figure 4:** Bibliographic coupling between accepted and rejected manuscripts.

### 3.3.3 Semantic similarity and peer review outcome

Having established that semantic similarity and bibliographic coupling capture different aspects of the relationship between documents, we also analysed the semantic similarity of the documents from the four platforms. Thus, for each platform we computed the tf-idf distance between all pairs of document based on their word stem.

Overall, we found that accepted manuscripts were more similar to each other than rejected manuscripts based on their abstracts and introduction (Fig. 5). We found a stronger effect for the similarity between the introduction ($R^2 = 0.01$) than for the abstract ($R^2 = 0.006$) when all platforms were pooled together. In other words, accepted pairs of manuscripts were more similar to each other compared to the other two pair types.

This analysis of the semantic similarity of documents (for both citations and text) showed some high levels trends based on whether or not the manuscripts were accepted after peer review. We therefore next examined the text content of the manuscripts with a more detailed approach to gain more insights on the patterns uncovered by the analysis on bibliographic coupling and textual similarity.
Finally, after having established high-level associations between the rejected and accepted manuscripts, we attempted to predict the peer review outcome with a logistic regression using a bag-of-words approach. Overall, the model was fairly successful at predicting the peer review outcome on a 10-fold cross-validated dataset (Tables 3, 4 & 5, random performance \( \sim 0.5 \) for all three metrics). The model was the most successful when the text of the introduction was used, followed by the text of the abstract and of the title.

After having established that we could predict to some extent the outcome of the peer review process with the word content of the manuscripts, we performed a more detailed analysis to try to get some insight about the key predictors of the outcome. We therefore computed the average tf-idf score of each stem for accepted and rejected manuscripts, and obtained measure of “importance” based on the difference between the two averages. This approach allowed us to identify the most important keywords predicting the
acceptance of a manuscript (Fig. 6).

Although some differences were noticeable across platforms regarding the predictors of acceptance (Tables 9, 10 & 11) and rejection (Tables 12, 13 & 14), some robust patterns emerged. Word stems related to sub fields of neural networks and machine learning (e.g., learn, neural, gradient, train) were increasing the odds of the manuscript to be accepted. However word stems related to the sub fields of logic, symbolic processing and knowledge representation (e.g, use, base, system, logic, fuzz, knowledge, rule) were decreasing the odds that a manuscript would get accepted.
4 Discussion and Conclusion

4.1 Summary of results

From a linguistic point of view, our results suggest that accepted manuscripts could be written in a more complex, less accessible English. Using two indices of readability, one of which is agnostic to the word content of the manuscript (FRE), we found that the accepted manuscripts obtained lower readability scores. Strikingly, we found the same effect for almost all our independent datasets. The same pattern was also observed for the title, the abstract and the introduction. Using a different type of readability indicator - the proportion of general, specific scientific and AI jargon words - we found that manuscripts that contained a greater proportion of jargon words were more likely to be accepted. This finding may partly explain that the readability of manuscripts has steadily declined during the last century [39]. In other words, it is possible that part of this decline could be attributed to a selection process taking place during peer review.
When considering the word content of the accepted manuscripts, we found that they had words that were acquired at a later age, that were more abstract and that were less common than the words from the rejected manuscripts. Additionally, these manuscripts were shorter and had increased lexical diversity. Once again, the effect size were small given the highly multivariate determination of the peer review outcome. The effects were replicated across multiples independent datasets from different fields in AI, which strengthens the conclusions of our analysis.

From a content point of view, we compared manuscripts based on their referencing patterns and word content. We compared the coupling both based on the raw number of common references (intersection) and the fraction of overlap between the manuscripts’ references (Jaccard similarity). We found that accepted pairs had a larger intersection than other pairs, and found a similar, but less reliable effect for the similarity. We used a tf-idf vectorial representation of the text from all manuscripts in the database, compared all possible pairs of manuscripts, and we found that pairs of accepted manuscript had considerably more overlap between their word content. This high level analysis of the manuscript’s content revealed that some topics might be associated with different odds of acceptance. We performed a correlation between the bibliographic coupling and the semantic similarity to get an idea of the how independent was the information provided by these two semantic indicators. As reported previously [44], we found a weak to moderate correlation between the two, which suggest that they provide distinct sources of information about topic similarity in accepted manuscripts.

Finally, we built a logistic regression to predict the peer review outcome, which revealed that using the title, abstract or introduction words lead to robust predictions. Our results are compatible with the presence of content bias, where trending topics in AI such as machine learning and neural networks were linked with greater acceptance rate, whereas words related to symbolic processing and knowledge-based reasoning lead to lower acceptance rates.

4.2 Implications

Taken together, our analysis of the linguistic aspects of manuscripts are coherent with linguistic biases during peer review. It has been reported that writers using English as their main language (L1) use words that are more abstract and less frequent than writers with English as their second language (L2) [45]. Additionally, this effect is exacerbated by the L2 proficiency (where larger differences are observed for beginners than advanced L2 speakers) [46]. The complexity of L2 writing was also shown to correlate with proficiency [47, 48, 49]. Our results are therefore compatible with the hypothesis that L2 writers are less likely to get their manuscript accepted at peer review.

Our results are also compatible with a content bias where manuscripts on the topics of machine learning techniques and neural networks have greater odds to be accepted at peer review. Leading figures of the AI community have raised their voice against the overwhelming dominance of neural networks and deep
learning in the domain of AI [50, 51, 52]. Recent successes of deep learning and neural networks might explain their dominance in the field, but a bias against other techniques might impede developments similar to the ones that lead to the breakthroughs underlying the deep learning revolution [53]. Following this idea, several researchers have indicated that symbolic processing could hold the answer to shortcomings of deep learning [54, 50, 55].

Other than biases during peer review, our results have implications for the quality of scientific communications. A recent report on reproducibility in machine learning found a positive correlation between the readability of a manuscript and successful replication of the claims that it makes [56]. Selecting for less readable manuscripts during peer review may therefore increase the proportion of non-reproducible research findings.

4.3 Limitations

Although the main objective of our analysis was to investigate the presence of content or linguistic biases in peer review, all our analysis are correlational, and there are possible confounds that could explain our results. For instance, while our findings that some linguistic aspects of the manuscripts - the readability and psycholinguistic attributes - were correlated with the peer review outcome, we cannot infer that those variables directly lowered the odds of acceptance.

Similarly, we cannot infer that there is a bias against manuscripts on the topic of machine learning techniques and neural networks. For instance, reviewers favoring high benchmark performance might accept more manuscripts using the state of the art techniques. In the scenario, reviewers would not reject a manuscript using a non-mainstream technique because of bias against it, but simply because they value some aspects where it underperforms.

Another limitation to our findings is the methodology of the peer read dataset [26]. For most manuscripts included in the dataset, their status is inferred and the true outcome of the peer review process is unknown. Although [26] validated their method on a subset of their data, the accuracy is not perfect. However, we believe that the large size of the dataset is enough to counteract this source of noise. Only the minority of manuscripts included in their dataset had a true peer review outcome provided by the publishing venue. This highlight the need for publishers and conferences to open their peer review process in order to further advance our understanding of the strengths and limitations of the peer review process.

In sum, our results are compatible with the presence of a linguistic and a content bias in the peer review process of major conferences in AI. Although we were able to replicate our results across different dataset, similar studies have to be conducted both in the field of AI and in other disciplines to validate the conclusions of our study.
Acknowledgements

This research was funded by the Canada Research Chair program and the Social Sciences and Humanities Research Council of Canada.

References

[1] J. Ziman, Real science: What it is and what it means. Cambridge University Press, 2002.

[2] S. Wakeling, P. Willett, C. Creaser, J. Fry, S. Pinfield, V. Spezi, M. Bonne, C. Founti, and I. Medina Perea, “‘No comment’? A study of commenting on PLOS articles,” Journal of Information Science, p. 016555151881996, Jan. 2019.

[3] A. Mulligan, “Is peer review in crisis?,” Oral Oncology, vol. 41, no. 2, pp. 135–141, 2005.

[4] K. D. Mayden, “Peer Review: Publication’s Gold Standard,” J Adv Pract Oncol, vol. 3, no. 2, pp. 117–122, 2012.

[5] J. R. Tregellas, J. Smucny, D. C. Rojas, and K. T. Legget, “Predicting academic career outcomes by predoctoral publication record,” PeerJ, vol. 6, p. e5707, Oct. 2018.

[6] E. C. McKiernan, L. A. Schimanski, C. Muñoz Nieves, L. Matthias, M. T. Niles, and J. P. Alperin, “Use of the Journal Impact Factor in academic review, promotion, and tenure evaluations,” 2019.

[7] D. Moher, F. Naudet, I. A. Cristea, F. Miedema, J. P. A. Ioannidis, and S. N. Goodman, “Assessing scientists for hiring, promotion, and tenure,” PLOS Biology, vol. 16, p. e2004089, Mar. 2018.

[8] J. C. Bailar, “Reliability, fairness, objectivity and other inappropriate goals in peer review,” Behavioral and Brain Sciences, vol. 14, no. 1, pp. 137–138, 1991.

[9] C. J. Lee, “A Kuhnian critique of psychometric research on peer review,” Philosophy of Science, vol. 79, no. 5, pp. 859–870, 2012.

[10] D. V. Cicchetti, “The reliability of peer review for manuscript and grant submissions: A cross-disciplinary investigation,” Behavioral and Brain Sciences, vol. 14, no. 1, pp. 119–135, 1991.

[11] L. Bornmann, R. Mutz, and H.-D. Daniel, “A reliability-generalization study of journal peer reviews: A multilevel meta-analysis of inter-rater reliability and its determinants,” PloS one, vol. 5, no. 12, p. e14331, 2010.

[12] E. Price, “The NIPS experiment,” 2014.

[13] P. S. Forscher, M. Brauer, W. T. Cox, and P. G. Devine, “How many reviewers are required to obtain reliable evaluations of NIH R01 grant proposals?,” 2019.
14] C. J. Lee, C. R. Sugimoto, G. Zhang, and B. Cronin, “Bias in peer review,” *Journal of the American Society for Information Science and Technology*, vol. 64, pp. 2–17, Jan. 2013.

15] P. U. K. De Silva and C. K. Vance, “Preserving the Quality of Scientific Research: Peer Review of Research Articles,” in *Scientific Scholarly Communication: The Changing Landscape* (P. U. K. De Silva and C. K. Vance, eds.), Fascinating Life Sciences, pp. 73–99, Cham: Springer International Publishing, 2017.

16] D. Murray, K. Siler, V. Lariviére, W. M. Chan, A. M. Collings, J. Raymond, and C. R. Sugimoto, “Gender and international diversity improves equity in peer review,” *bioRxiv*, Jan. 2018.

17] J. L. Jackson, M. Srinivasan, J. Rea, K. E. Fletcher, and R. L. Kravitz, “The Validity of Peer Review in a General Medicine Journal,” *PLOS ONE*, vol. 6, p. e22475, July 2011.

18] C. E. T. Paine and C. W. Fox, “The effectiveness of journals as arbiters of scientific impact,” *Ecology and Evolution*, vol. 8, no. 19, pp. 9566–9585, 2018.

19] S. Goodman, J. Berlin, S. Fletcher, and R. Fletcher, “Manuscript quality before and after peer review and editing at annals of internal medicine,” *Annals of Internal Medicine*, vol. 121, pp. 11–21, July 1994.

20] J.-P. E. Pierie, H. C. Walvoort, and A. J. P. Overbeke, “Readers’ evaluation of effect of peer review and editing on quality of articles in the Nederlands Tijdschrift voor Geneeskunde,” *The Lancet*, vol. 348, no. 9040, pp. 1480–1483, 1996.

21] V. Calcagno, E. Demoinet, K. Gollner, L. Guidi, D. Ruths, and C. d. Mazancourt, “Flows of Research Manuscripts Among Scientific Journals Reveal Hidden Submission Patterns,” *Science*, vol. 338, pp. 1065–1069, Nov. 2012.

22] R. Smith, “Classical peer review: an empty gun,” *Breast Cancer Research*, vol. 12, no. 4, p. S13, 2010.

23] O. Sabaj Meruane, C. González Vergara, and Á. Pina-Stranger, “What We Still Don’t Know About Peer Review,” *Journal of Scholarly Publishing*, vol. 47, pp. 180–212, Jan. 2016.

24] H. Piwowar, J. Priem, V. Larivièrè, J. P. Alperin, L. Matthias, B. Norlander, A. Farley, J. West, and S. Haustein, “The state of OA: a large-scale analysis of the prevalence and impact of Open Access articles,” *PeerJ*, vol. 6, p. e4375, Feb. 2018.

25] C. Sutton and L. Gong, “Popularity of arXiv.org within Computer Science,” *arXiv:1710.05225 [cs]*, Oct. 2017. arXiv: 1710.05225.
[26] D. Kang, W. Ammar, B. Dalvi, M. van Zuylen, S. Kohlmeier, E. Hovy, and R. Schwartz, “A Dataset of Peer Reviews (PeerRead): Collection, Insights and NLP Applications,” arXiv preprint arXiv:1804.09635, 2018.

[27] W. Jen, S. Zhang, and M. Chen, “Predicting Conference Paper Acceptance,” p. 7, 2018.

[28] T. Ghosal, R. Verma, A. Ekbal, and P. Bhattacharyya, “A Sentiment Augmented Deep Architecture to Predict Peer Review Outcomes,” in 2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL), pp. 414–415, June 2019.

[29] F. Qiao, L. Xu, and X. Han, “Modularized and Attention-Based Recurrent Convolutional Neural Network for Automatic Academic Paper Aspect Scoring,” in Web Information Systems and Applications (X. Meng, R. Li, K. Wang, B. Niu, X. Wang, and G. Zhao, eds.), Lecture Notes in Computer Science, pp. 68–76, Springer International Publishing, 2018.

[30] J. S. Ross, C. P. Gross, M. M. Desai, Y. Hong, A. O. Grant, S. R. Daniels, V. C. Hachinski, R. J. Gibbons, T. J. Gardner, and H. M. Krumholz, “Effect of Blinded Peer Review on Abstract Acceptance,” JAMA, vol. 295, pp. 1675–1680, Apr. 2006.

[31] T. Tregenza, “Gender bias in the refereeing process?,” Trends in Ecology & Evolution, vol. 17, pp. 349–350, Aug. 2002.

[32] A. J. Herrera, “Language bias discredits the peer-review system,” Nature, vol. 397, p. 467, Feb. 1999.

[33] V. Larivière and J.-P. Warren, “Introduction: The Dissemination of National Knowledge in an Internationalized Scientific Community,” Canadian Journal of Sociology, vol. 44, pp. 1–8, Mar. 2019.

[34] K. Hao, “We analyzed 16,625 papers to figure out where AI is headed next,” Jan. 2019.

[35] M. F. Porter, “An algorithm for suffix stripping,” Program, vol. 14, no. 3, pp. 130–137, 1980.

[36] M. Brysbaert and B. New, “Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English,” Behavior research methods, vol. 41, no. 4, pp. 977–990, 2009.

[37] M. Brysbaert, A. B. Warriner, and V. Kuperman, “Concreteness ratings for 40 thousand generally known English word lemmas,” Behavior research methods, vol. 46, no. 3, pp. 904–911, 2014.

[38] V. Kuperman, H. Stadthagen-Gonzalez, and M. Brysbaert, “Age-of-acquisition ratings for 30,000 English words,” Behavior Research Methods, vol. 44, no. 4, pp. 978–990, 2012.

[39] P. Plavén-Sigray, G. J. Matheson, B. C. Schiffler, and W. H. Thompson, “The readability of scientific texts is decreasing over time,” eLife, vol. 6, p. e27725, Sept. 2017.
[40] R. Flesch, “A new readability yardstick,” *Journal of Applied Psychology*, vol. 32, no. 3, pp. 221–233, 1948.

[41] J. Kincaid, R. Fishburne, R. Rogers, and B. Chissom, “Derivation Of New Readability Formulas (Automated Readability Index, Fog Count And Flesch Reading Ease Formula) For Navy Enlisted Personnel,” *Institute for Simulation and Training*, Jan. 1975.

[42] J. S. Chall and E. Dale, *Readability revisited: the new Dale-Chall readability formula*. Brookline Books, May 1995. Google-Books-ID: 2nbuAAAAMAAJ.

[43] D. Szucs and J. P. A. Ioannidis, “When Null Hypothesis Significance Testing Is Unsuitable for Research: A Reassessment,” *Front. Hum. Neurosci.*, vol. 11, 2017.

[44] M. Sainte-Marie, P. Mongeon, and V. Larivière, “Do you cite what I mean? Assessing the semantic scope of bibliographic coupling in economics,” p. 10, 2018.

[45] S. A. Crossley and D. S. McNamara, “Computational assessment of lexical differences in L1 and L2 writing,” *Journal of Second Language Writing*, vol. 18, pp. 119–135, June 2009.

[46] S. A. Crossley, T. Salsbury, D. S. McNamara, and S. Jarvis, “Predicting lexical proficiency in language learner texts using computational indices,” *Language Testing*, vol. 28, pp. 561–580, Oct. 2011.

[47] J.-y. Kim, “Predicting L2 Writing Proficiency Using Linguistic Complexity Measures: A Corpus-Based Study,” *english teaching*, vol. 69, pp. 27–51, Dec. 2014.

[48] A. C. Lahuerta Martínez, “Analysis of syntactic complexity in secondary education EFL writers at different proficiency levels,” *Assessing Writing*, vol. 35, pp. 1–11, Jan. 2018.

[49] N. N. Radhiah and Z. Abidin, “A longitudinal corpus study of syntactic complexity development in L2 writing,” 2018.

[50] G. Marcus, “Deep learning: A critical appraisal,” *arXiv preprint arXiv:1801.00631*, 2018.

[51] M. Jordan, “Artificial Intelligence — The Revolution Hasn’t Happened Yet,” Apr. 2018.

[52] W. Knight, “One of the fathers of AI is worried about its future,” 2018.

[53] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” pp. 1097–1105, 2012.

[54] H. Geffner, “Model-free, Model-based, and General Intelligence,” *arXiv:1806.02308 [cs]*, June 2018. arXiv: 1806.02308.
[55] M. Garnelo and M. Shanahan, “Reconciling deep learning with symbolic artificial intelligence: representing objects and relations,” Current Opinion in Behavioral Sciences, vol. 29, pp. 17–23, Oct. 2019.

[56] E. Raff, “A Step Toward Quantifying Independently Reproducible Machine Learning Research,” arXiv:1909.06674 [cs, stat], Sept. 2019. arXiv: 1909.06674.
Table 6: Summary of statistics for readability and jargon usage of US based accepted and rejected manuscripts.

| Section   | Variable | p     | Acc-Rej | Mean rejected | S.d. rejected | N rejected | Mean accepted | S.d. accepted | N accepted | Cohen's d |
|-----------|----------|-------|---------|---------------|---------------|------------|---------------|---------------|------------|------------|
| Title     | FRE      | 2.96E-02 | .56    | -1.89         | 40.84         | 1714       | -11.45        | 43.08         | 1113       | 0.085      |
| Abstract  | FRE      | 1.7E-07  | 2.35   | 18.47         | 12.69         | 1882       | 16.12         | 11.94         | 1239       | 0.190      |
| Introduction | FRE   | 3.0E-01  | .38    | 27.14         | 10.04         | 1545       | 26.75         | 9.22          | 1215       | 0.040      |
| Title     | NDC     | 9.8E-09  | .53    | 15.19         | 2.38          | 1714       | 15.72         | 2.38          | 1113       | 0.222      |
| Abstract  | NDC     | 5.7E-06  | .15    | 12.50         | 0.90          | 1882       | 12.65         | 0.87          | 1239       | 0.165      |
| Introduction | NDC | 5.7E-11  | .18    | 12.12         | 0.77          | 1545       | 12.30         | 0.69          | 1215       | 0.249      |
| Title     | General | 1.2E-02  | .02    | 0.38          | 0.22          | 1715       | 0.41          | 0.23          | 1113       | 0.099      |
| Abstract  | General | 5.6E-06  | .01    | 0.30          | 0.07          | 1882       | 0.31          | 0.06          | 1239       | 0.163      |
| Introduction | General | 1.7E-01  | .00    | 0.26          | 0.06          | 1545       | 0.26          | 0.05          | 1215       | 0.052      |
| Title     | Specific  | 1.2E-02 | .02    | 0.41         | 0.23          | 1715       | 0.43          | 0.24          | 1113       | 0.098      |
| Abstract  | Specific  | 1.1E-05 | .01    | 0.32          | 0.07          | 1882       | 0.33          | 0.07          | 1239       | 0.158      |
| Introduction | Specific | 1.1E-04 | .01    | 0.28          | 0.06          | 1545       | 0.29          | 0.05          | 1215       | 0.147      |
| Title     | AI      | 2.9E-02  | .01    | 0.09         | 0.13          | 1715       | 0.10          | 0.14          | 1113       | 0.085      |
| Abstract  | AI      | 1.1E-05  | .01    | 0.05          | 0.03          | 1882       | 0.05          | 0.03          | 1239       | 0.159      |
| Introduction | AI     | 1.0E-06 | .00    | 0.04          | 0.02          | 1545       | 0.04          | 0.02          | 1215       | 0.186      |

Table 7: Summary of statistics for readability and jargon usage of non-US based accepted and rejected manuscripts.

| Section   | Variable | p     | Acc-Rej | Mean rejected | S.d. rejected | N rejected | Mean accepted | S.d. accepted | N accepted | Cohen's d |
|-----------|----------|-------|---------|---------------|---------------|------------|---------------|---------------|------------|------------|
| Title     | FRE      | 6.3E-06 | .11    | -12.42        | 38.52         | 6257       | -16.02        | 41.16         | 1615       | 0.131      |
| Abstract  | FRE      | 2.4E-10 | 1.16   | 21.00         | 13.59         | 7100       | 18.06         | 12.21         | 1817       | 0.157      |
| Introduction | FRE   | 1.7E-01 | .36    | 28.02         | 10.52         | 5514       | 28.38         | 9.29          | 1769       | 0.036      |
| Title     | NDC     | 4.6E-20 | .61    | 15.01         | 2.37          | 6257       | 15.61         | 2.36          | 1635       | 0.256      |
| Abstract  | NDC     | 2.0E-05 | .11    | 12.44         | 1.09          | 7100       | 12.54         | 0.91          | 1817       | 0.101      |
| Introduction | NDC | 2.0E-16 | .05    | 12.05         | 1.03          | 5514       | 12.20         | 0.71          | 1769       | 0.187      |
| Title     | General | 8.3E-04 | .02    | 0.38          | 0.22          | 6262       | 0.40          | 0.24          | 1635       | 0.098      |
| Abstract  | General | 4.1E-26 | .02    | 0.29          | 0.07          | 7104       | 0.31          | 0.07          | 1817       | 0.262      |
| Introduction | General | 8.0E-10 | .01    | 0.25          | 0.06          | 5514       | 0.26          | 0.05          | 1769       | 0.150      |
| Title     | Specific  | 8.1E-03 | .01    | 0.41          | 0.23          | 6262       | 0.44          | 0.25          | 1635       | 0.116      |
| Abstract  | Specific  | 1.4E-26 | .02    | 0.31          | 0.08          | 7104       | 0.33          | 0.07          | 1817       | 0.263      |
| Introduction | Specific | 1.3E-35 | .02    | 0.27          | 0.06          | 5514       | 0.29          | 0.05          | 1769       | 0.310      |
| Title     | AI      | 1.5E-03 | .01    | 0.09         | 0.13          | 6262       | 0.10          | 0.15          | 1635       | 0.095      |
| Abstract  | AI      | 8.0E-37  | .01    | 0.04          | 0.03          | 7104       | 0.05          | 0.03          | 1817       | 0.330      |
| Introduction | AI     | 3.0E-54 | .01    | 0.03          | 0.02          | 5514       | 0.04          | 0.02          | 1769       | 0.400      |

Table 8: Summary of statistics for lexical features for accepted and rejected manuscripts.
Table 9: Most important predictors for acceptance in decreasing order of importance: title
| AI Attr | AI Score accepted | AI Score rejected | L2 Attr | L2 Score accepted | L2 Score rejected | L2 Attr | L2 Score accepted | L2 Score rejected | L2 Attr | L2 Score accepted | L2 Score rejected | All Attr | All Score accepted | All Score rejected |
|---------|------------------|------------------|---------|------------------|------------------|---------|------------------|------------------|---------|------------------|------------------|---------|------------------|------------------|
| recall  | 0.013 ± 0.001    | 0.012 ± 0.001    | matrix  | 0.33 ± 0.007     | 0.32 ± 0.007     | 0.33 ± 0.007 | 0.32 ± 0.007 | 0.33 ± 0.007 | 0.32 ± 0.007 | 0.33 ± 0.007 | 0.32 ± 0.007 | 0.33 ± 0.007 | 0.32 ± 0.007 | 0.33 ± 0.007 | 0.32 ± 0.007 |
| precision | 0.017 ± 0.005   | 0.015 ± 0.005   | algorithm | 0.75 ± 0.002   | 0.75 ± 0.002   | 0.75 ± 0.002 | 0.75 ± 0.002 | 0.75 ± 0.002 | 0.75 ± 0.002 | 0.75 ± 0.002 | 0.75 ± 0.002 | 0.75 ± 0.002 | 0.75 ± 0.002 | 0.75 ± 0.002 | 0.75 ± 0.002 |
| F1 score | 0.017 ± 0.005   | 0.015 ± 0.005   | sentiment | 0.43 ± 0.002   | 0.43 ± 0.002   | 0.43 ± 0.002 | 0.43 ± 0.002 | 0.43 ± 0.002 | 0.43 ± 0.002 | 0.43 ± 0.002 | 0.43 ± 0.002 | 0.43 ± 0.002 | 0.43 ± 0.002 | 0.43 ± 0.002 | 0.43 ± 0.002 |
| AUC     | 0.012 ± 0.001   | 0.011 ± 0.001   | text     | 0.38 ± 0.002   | 0.37 ± 0.002   | 0.38 ± 0.002 | 0.37 ± 0.002 | 0.38 ± 0.002 | 0.37 ± 0.002 | 0.38 ± 0.002 | 0.37 ± 0.002 | 0.38 ± 0.002 | 0.37 ± 0.002 | 0.38 ± 0.002 | 0.37 ± 0.002 |
| accuracy| 0.013 ± 0.001   | 0.012 ± 0.001   | user     | 0.34 ± 0.002   | 0.33 ± 0.002   | 0.34 ± 0.002 | 0.33 ± 0.002 | 0.34 ± 0.002 | 0.33 ± 0.002 | 0.34 ± 0.002 | 0.33 ± 0.002 | 0.34 ± 0.002 | 0.33 ± 0.002 | 0.34 ± 0.002 | 0.33 ± 0.002 |

Table 10: Most important predictors for acceptance in decreasing order of importance: abstract
Table 13: Most important predictors for rejection in decreasing order of importance: abstract
| AI STem | CL STem | LG STem | ICML Score accepted | All STem |
|--------|--------|--------|---------------------|----------|
| detect | 0.0052 | 0.0013 | 0.0060 ± 0.0007    | 0.0054 ± 0.0002 |
| graph  | 0.0093 | 0.0029 | 0.0112 ± 0.0006    | 0.0089 ± 0.0007 |
| symp   | 0.0066 | 0.0014 | 0.0135 ± 0.0008    | 0.0080 ± 0.0007 |
| state  | 0.0082 | 0.0011 | 0.0041 ± 0.0001    | 0.0081 ± 0.0006 |
| matrix | 0.0015 | 0.0002 | 0.0041 ± 0.0001    | 0.0001 ± 0.0001 |
| cost   | 0.0005 | 0.0001 | 0.0005 ± 0.0007    | 0.0001 ± 0.0001 |
| model  | 0.0433 | 0.0027 | 0.0034 ± 0.0008    | 0.0004 ± 0.0001 |
| recomm | 0.0311 | 0.0008 | 0.0046 ± 0.0011    | 0.0011 ± 0.0005 |
| event  | 0.0003 | 0.0001 | 0.0057 ± 0.0008    | 0.0019 ± 0.0003 |
| factor | 0.0429 | 0.0020 | 0.0027 ± 0.0012    | 0.0031 ± 0.0001 |
| decics | 0.0056 | 0.0012 | 0.0079 ± 0.0006    | 0.0023 ± 0.0005 |
| entity | 0.0033 | 0.0011 | 0.0067 ± 0.0007    | 0.0010 ± 0.0002 |
| reason | 0.0228 | 0.0006 | 0.0032 ± 0.0005    | 0.0030 ± 0.0005 |
| anomal | 0.00 ± 0.0 | 0.00 ± 0.0 | 0.00 ± 0.0 | 0.00 ± 0.0 |
| design | 0.0066 | 0.0007 | 0.0068 ± 0.0005    | 0.0004 ± 0.0011 |
| action | 0.0079 | 0.0016 | 0.0001 ± 0.0009    | 0.0001 ± 0.0009 |
| patient | 0.0104 | 0.0006 | 0.0036 ± 0.0006    | 0.0018 ± 0.0006 |
| anomali | 0.00 ± 0.0 | 0.00 ± 0.0 | 0.00 ± 0.0 | 0.00 ± 0.0 |
| base   | 0.1278 | 0.0008 | 0.0047 ± 0.0004    | 0.0009 ± 0.0002 |
| reward | 0.0067 | 0.0018 | 0.0037 ± 0.0008    | 0.0028 ± 0.0004 |
| algorithm | 0.0047 | 0.0011 | 0.0054 ± 0.0008    | 0.0028 ± 0.0005 |
| category | 0.0033 | 0.0008 | 0.0052 ± 0.0004    | 0.0031 ± 0.0006 |
| entail | 0.0018 | 0.0007 | 0.0038 ± 0.0005    | 0.0031 ± 0.0006 |
| gaussian | 0.0051 | 0.0013 | 0.0071 ± 0.0009    | 0.0062 ± 0.0004 |
| causal  | 0.0011 | 0.0007 | 0.0033 ± 0.0007    | 0.0026 ± 0.0004 |
| similar | 0.0096 | 0.0012 | 0.0116 ± 0.0006    | 0.0076 ± 0.0004 |
| game   | 0.0048 | 0.0010 | 0.0066 ± 0.0008    | 0.0064 ± 0.0004 |
| rank   | 0.0024 | 0.0005 | 0.0044 ± 0.0005    | 0.0015 ± 0.0004 |
| manifol | 0.0013 | 0.0002 | 0.0016 ± 0.0004    | 0.0010 ± 0.0003 |
| dynam   | 0.0047 | 0.0011 | 0.0047 ± 0.0006    | 0.0040 ± 0.0004 |
| norm   | 0.0056 | 0.0012 | 0.0055 ± 0.0008    | 0.0052 ± 0.0004 |
| practic | 0.0061 | 0.0006 | 0.0080 ± 0.0003    | 0.0074 ± 0.0003 |
| manifold | 0.0019 | 0.0012 | 0.0044 ± 0.0005    | 0.0040 ± 0.0003 |
| recover | 0.0009 | 0.0004 | 0.0027 ± 0.0004    | 0.0024 ± 0.0003 |
| tangent | 0.0007 | 0.0003 | 0.0018 ± 0.0002    | 0.0016 ± 0.0002 |
| inliner | 0.0081 | 0.0017 | 0.0015 ± 0.0005    | 0.0014 ± 0.0003 |
| data   | 0.0033 | 0.0002 | 0.0039 ± 0.0004    | 0.0032 ± 0.0003 |
| th     | 0.0043 | 0.0012 | 0.0041 ± 0.0005    | 0.0035 ± 0.0003 |
| hypothe | 0.0024 | 0.0006 | 0.0042 ± 0.0003    | 0.0038 ± 0.0003 |
| simili | 0.0023 | 0.0005 | 0.0042 ± 0.0003    | 0.0038 ± 0.0003 |
| real   | 0.0037 | 0.0005 | 0.0005 ± 0.0003    | 0.0004 ± 0.0002 |
| mean   | 0.0007 | 0.0003 | 0.0006 ± 0.0002    | 0.0005 ± 0.0002 |
| known  | 0.0087 | 0.0017 | 0.0083 ± 0.0003    | 0.0080 ± 0.0003 |
| miler  | 0.0102 | 0.0012 | 0.0018 ± 0.0007    | 0.0016 ± 0.0007 |
| control | 0.0053 | 0.0001 | 0.0069 ± 0.0006    | 0.0058 ± 0.0006 |

Table 14: Most important predictors for rejection in decreasing order of importance: introduction