Analyzing sentiments in peer review reports: Evidence from two science funding agencies

Junwen Luo1, Thomas Feliciani2, Martin Reinhart3, Judith Hartstein4,5, Vineeth Das2, Olalere Alabi2, and Kalpana Shankar1

1School of Information and Communication Studies, University College Dublin, Dublin, Ireland
2School of Sociology and Geary Institute of Public Policy, University College Dublin, Dublin, Ireland
3Robert K. Merton Center for Science Studies, Humboldt-Universität zu Berlin, Berlin, Germany
4German Centre for Higher Education Research and Science Studies (D2ZH), Berlin, Germany
5Faculty of Humanities and Social Sciences, Humboldt-Universität zu Berlin, Berlin, Germany

Keywords: evaluation report, peer review, science funding, sentiment analysis

ABSTRACT

Using a novel combination of methods and data sets from two national funding agency contexts, this study explores whether review sentiment can be used as a reliable proxy for understanding peer reviewer opinions. We measure reviewer opinions via their review sentiments on both specific review subjects and proposals’ overall funding worthiness with three different methods: manual content analysis and two dictionary-based sentiment analysis algorithms (TextBlob and VADER). The reliability of review sentiment to detect reviewer opinions is addressed by its correlation with review scores and proposals’ rankings and funding decisions. We find in our samples that review sentiments correlate with review scores or rankings positively, and the correlation is stronger for manually coded than for algorithmic results; manual and algorithmic results are overall correlated across different funding programs, review sections, languages, and agencies, but the correlations are not strong; and manually coded review sentiments can quite accurately predict whether proposals are funded, whereas the two algorithms predict funding success with moderate accuracy. The results suggest that manual analysis of review sentiments can provide a reliable proxy of grant reviewer opinions, whereas the two SA algorithms can be useful only in some specific situations.

1. INTRODUCTION

Expert reviewers are central to peer review. Based on their recommendations, scientific journals select manuscripts to publish, hiring committees select faculty to hire, and funding agencies select grant proposals to fund. The latter case, grant peer review, is of special interest as reviewers are asked to assess scientific work that has not yet been performed. Even though grant review procedures can involve a large number of steps to ensure the objectivity and fairness of funding decisions, the literature still points to the lack of reliability and transparency in grant review mechanisms (Cicchetti, 1991; Pier, Brauer et al., 2018).

In addition to the many cognitive and social complexities of peer review, the difficulty of obtaining data about peer review processes makes it a difficult topic to study empirically (Squazzoni, Ahnveiler et al., 2020), further contributing to its opacity. The objective of this paper is to demonstrate how the analysis of review sentiment can supplement our understanding of reviewer opinions on both specific review subjects in the evaluation process and
proposals’ overall funding worthiness. To do so, we analyze a corpus of peer review texts and related documents from two national science funding agencies: Science Foundation Ireland (SFI) and the Swiss National Science Foundation (SNSF). We identify and extract reviews of individual subjects and interpret sentiments of the extracted reviews in two ways: manual coding based on content analyses and two dictionary-based algorithms, TextBlob and VADER (Hutto & Gilbert, 2014; Loria, Keen et al., 2014). We compare the coding results from the different methods and calibrate the results with other types of data in grant review contexts, such as review scores, rankings, and funding decisions of proposals.

We analyze review reports to address three empirical research questions as three dimensions of the reliability of review sentiments to be used in peer review research. First, we ask whether the sentiments of review texts correlate with the review scores. Second, we explore whether automated sentiment analysis (hereafter: SA) of review texts performs similarly to the manual analysis result, usually used as a benchmark for algorithmic tools (Turney, 2002). Third, we address whether review sentiments can accurately predict the proposals to be funded or not.

Our empirical results demonstrate that review sentiments correlate with scores or rankings positively, and the correlation is stronger for manually coded than for algorithmic results; manual and algorithmic results are overall correlated, but the correlations are not strong; and manually coded review sentiments can quite accurately predict proposals’ funding decisions, whereas the accuracy of the two SA algorithms is moderate. These findings suggest that review sentiments, especially the manually coded sentiments, can be used as a reliable proxy of reviewer opinions in peer review research whereas the two SA algorithms can be useful in specific situations. To begin with, the following literature review will guide us to understand what characterizes grant reviewer opinions and processes, and how reviewer opinions can be measured.

2. CHARACTERISTICS OF GRANT REVIEWER OPINIONS

2.1. Reviewer Opinions as a Processual Element of Communication

Funding agencies organize peer review very deliberately, with multiple review stages, different types of experts, and various evaluation criteria (Hartmann & Neidhardt, 1990; Langfeldt & Scordato, 2016; Reinhart, 2010; Schendzielorz & Reinhart, 2020). As a result, peer review decisions are part of a complex social process in which meaning making and negotiation between multiple actors play a central role (Hirschauer, 2010; Lamont, 2010; Mallard, Lamont, & Guetzkow, 2009).

Central to this process are peer reviewer opinions. A typical grant review process comprises a postal review stage, also called an independent, remote, or postal panel, followed by a collective panel review stage or a sitting panel. These two stages are adopted by many funders, including the two funders studied in this paper (SFI and SNSF), the US National Institutes of Health (NIH) (Pier et al., 2018), and the European Research Council (ERC) (Van den Besselaar, Sandström, & Schiffbaenker, 2018). Postal review is conducted first by peer experts who, independently of one another, review an assigned proposal. Reviewers communicate their opinions via (often semistructured) textual reviews to collective panels which synthesize the individual reviews and rank the proposals within competing pools to make their funding recommendations. Finally, agencies base their final funding decisions on the ranking provided by the panel, along with other considerations such as organizational goals, policy objectives, and available budget. Reviewers may even anticipate such considerations (e.g., by overly emphasizing
criticism or praise) (Reinhart, 2010). Therefore, although postal peer reviewer opinions on individual proposals are one (early) processual element of communication in the grant review process, they are often reserved for later uses in the process.

In our study, we analyze the review reports from the postal review stage. In most cases including SFI and SNSF, postal reviewers cannot update or revise their reviews after submission, nor can they interact with other reviewers or applicants during their individual evaluation process. Thus, postal reviews serve as a one-way transformation of reviewer opinions into a standardized format comprising texts and scores, typically structured following the review guidelines set by the agency.

2.2. Two Carriers of Reviewer Opinions: Texts and Scores

Review texts can be a few words, sentences, or paragraphs. They can be written in a structured way in a form requested by agencies (for instance, via a checklist of evaluation criteria or a section-separated review report) or organized by reviewers themselves based on their interpretations of the guidelines and own (epistemological and writing) styles (Reinhart, 2010). Before giving their overall judgment of the proposals’ funding worthiness, reviewers are usually asked to comment on and grade the individual review subjects set by the agency, such as scientific excellence and nonscientific impact. Each subject can be assessed and graded separately from others via section-separated review reports (e.g., in SFI, some panels in SNSF, and EU H2020 programs (European Commission [2015]) or, assessed distinctly but without separate grades for the overall score of each proposal (e.g., some panels in SNSF and some programs in the US National Science Foundation, NSF). In the case of section-separated forms, reviewer evaluations on individual subjects can also be aggregated into one overall review, such as by taking the mean of the review scores, or by commenting on a proposal as a whole. Aggregation is usually done by the reviewers themselves when they are asked to provide one overall review for each proposal. For review scores specifically, this can also be done by the agency as a way to summarize a multifaceted review into one overall score.

With respect to scores, many agencies design their postal review reports to supplement the textual reviews with categorized grading scales and provide reviewers with labels and explanations of the scales for them to use (Langfeldt & Scordato, 2016). For instance, SFI and many EU programs use a five-point grading scales (1 = poor, 5 = excellent); SNSF uses six-point scales (1 = schlecht/bad, 6 = sehr gut/very good), and the NIH uses nine-point scales (1 = exceptional, 9 = poor).

Review scores are a very convenient source of data for research on peer review. For one, they are more amenable to statistical analysis, whereas review texts require more interpretative work. Much of the existing literature on peer review thus focuses on scores rather than texts. For instance, variation in scores is often used as a measurement of reviewers’ disagreement and interreviewer reliability (Cicchetti, 1991; Fogelholm, Leppinen et al., 2012; Obrecht, Tibelius, & D’Aloisio, 2007; Pier, Raclaw et al., 2017; Pina, Buljan et al., 2021). However, people with similar or even identical underlying views of proposals may express their opinions using a different grade due to reviewers’ different interpretations of the grading scales (Cole, Cole, & Simon, 1981). This issue is called “grading heterogeneity” (Morgan, 2014). Also, different reviewers might use the same grade despite having different underlying views of the same proposal. Both cases can color our understanding of reviewers’ consensus or disagreement if one looks at scores only.

Fewer studies on peer review focus exclusively on review texts (Kretzenbacher, 2017; Ma, Luo et al., 2020; Reinhart, 2010); even fewer explored both. Pier et al. (2018) compared
(simulated) NIH panel reviewers’ scores and texts and found the numbers of strengths and of weaknesses reported in the review texts did not generally agree with the scores. Van den Besselaar et al. (2018) studied the linguistic characteristics of ERC review texts (e.g., length and frequency of positive and negative emotions) in relation to the corresponding review scores and funding decisions. They found negative reviews had a stronger effect on the panel scores, confirming earlier work that suggested negative comments are significant in breaking the overall positive tone of review texts (Reinhart, 2010).

2.3. Sentiment as a Proxy of Reviewer Opinions

SA is the computational treatment of subjectivity in a given text to classify this textual opinion as positive, negative, or neutral (Turney, 2002). SA is a common method for the estimation of opinions and emotions in a variety of domains, such as online reviews of commercial products or services and in research areas (e.g., psychology, philosophy, sociology) (Liu, 2010). In this regard, peer review texts in research evaluation are similar to other types of reviews and can be analyzed by SA tools (Luo et al., 2020).

Scholars have examined review sentiments in linguistic studies of grant reviews (Buljan, Garcia-Costa et al., 2020; Kretzenbacher & Thurmair, 1992, 1995; Van den Besselaar et al. 2018), in bibliometric studies of literature review citation (Yan, Chen, & Li, 2020; Zhang, Ding, & Milojević, 2013), and in altmetrics of scientific articles (Hassan, Aljohani et al., 2020; Liu, & Fang, 2017). These studies either rely on purely qualitative linguistic methods or use algorithmic methods and manual annotations without assessing the reliability of the algorithmic methods. The generalizability of results from existing studies is limited in at least two ways: First, multimethod designs that use more than one type of data are rare; second, comparative studies that apply their research questions and methods across different organizational and policy contexts are nonexistent to our knowledge. In light of the variability and complexity of grant review procedures at different agencies (Schendzielorz & Reinhart, 2020), several issues remain overlooked (e.g., how review scores are contextualized with textual reviews, and how high and low funding rates may affect reviewers’ assessments as well as agencies’ use of reviewer opinions).

The ways in which peer reviews are written make them amenable to automated SA methods because dictionary-based SA algorithms are at less of a disadvantage when classifying texts that, like scientific reviews, are written in a formal and straightforward language, with relatively scarce use of harder-to-classify figures of speech such as metaphors, antiphrasis, and sarcasm. Scientific reviews tend to be longer and more detailed than other types of reviews and require professional experience to understand scientific and technical details. Yet, they are generally written without hidden meaning or nuance, as reviewers’ overall opinions on whether to recommend a paper be published or a proposal be funded often need to be written in lay language for sometimes nonspecialist decision-makers. Earlier studies on SNSF and SFI review data show that grant review texts are written in direct and factual language. Despite some use of technical jargon, the words chosen to describe the (de)merits of the evaluated proposals are mostly colloquial and refrain from concealing the evaluative meaning (Reinhart, 2010).

In addition, the accuracy of algorithmic SA results is often measured by how well the results agree with human analyzers. The existing peer review literature that addressed review sentiment, to our knowledge, relies on either manual or algorithmic methods without testing or calibrating their accuracy and reliability against each other. Our study is novel in that we apply both manual and algorithmic coding approaches consistently across different funding
programs, review sections, languages, and agencies, as well as across different units of analysis, and then compare the results.

3. RESEARCH DESIGN, METHODS, AND RESEARCH QUESTIONS

Funding agencies apply their own review procedures and criteria (Langfeldt & Scordato, 2016), leading to very different characteristics and forms of review data that can be very difficult to compare explicitly. Our research is designed to incorporate the different forms of data that were shared by SFI and SNSF, especially the structure of review reports. Standardized section-separated review reports where each section represents one subject to be commented upon (such as impact) makes it easy to extract review texts on specific subjects. We call these “section-level reviews.” In the case of nonstandardized review reports, manual analyzers extract and code individual statements on particular subjects. We call these “statement-level reviews.” Both section-level and statement-level reviews extracted from review reports represent reviewer opinions on individual subjects and are considered as two different units of analysis in this study.

We apply manual and two algorithmic coding tools to measure sentiments of extracted reviews. Five human analyzers in the team read, interpreted, and coded textual reviews into one of five sentiment categories: very negative (−1), moderate negative (−0.5), neutral (0), moderate positive (0.5), and very positive (1). This five-point category was expanded from the three frequently used categories of sentiments (negative, neutral, and positive) to be more accurate by separating “very” or “moderate” degrees. This five-point scale is also compatible with the actual grading scales applied by SFI, where 1 indicates “very bad” to 5 for “outstanding.” We conducted manual coding of section-level reviews from standardized review reports at SFI and some SNSF cases. All manually coded reviews were independently coded by two team members. Internal training and a pilot exercise ensured a satisfactory level of intercoder reliability. All instances where the two coders disagreed were resolved through discussion. We also reutilized the manual coding results of statement-level reviews from the same SNSF pools conducted 10 years ago (Reinhart, 2010).

Two open-source SA algorithms were applied: TextBlob (Loria et al., 2014) and VADER (Valence Aware Dictionary and sEntiment Reasoner—Hutto & Gilbert, 2014), both from the NLTK.org (Natural Language Toolkit) library for Python 3.x (Loper & Bird, 2002). Both algorithms rely on pretrained dictionaries where each word is assigned a sentiment weight based on the sign and intensity of its emotion. The sentiment of a text (such as a review) is then the sum of the weights of its words. Sentiment scores range between −1 (the most negative) and +1 (the most positive). Although more sophisticated tools for SA have emerged in the literature (e.g., BERT: see Devlin, Chang et al. 2019), here we focus on intuitive dictionary-based algorithms to make the point that even simple and accessible tools for SA might have sufficient validity to be useful additions to the peer review research.

In Table 1, we summarize how we use these approaches with our empirical data to explore three research questions.

**RQ1: Do review sentiment scores correlate with review scores?**

When review scores are available, such as in the SNSF case, section-level review sentiments can be compared with section-level scores to explore the similarities and differences between the two carriers of reviewer opinions: texts and scores. Such similarities/differences can also be used to measure the consistency between reviewing and grading behaviors of either one single reviewer or multiple reviewers evaluating the same proposals. A correlation analysis between
review scores and sentiment scores is a straightforward way to measure such consistency: The stronger the correlations, the higher the consistency between the variables. We use Spearman’s rank correlation coefficient because the review score data is ordinal (1 = bad, 6 = very good). We explore further the correlation between review sentiments and scores under groups of different disciplines, review sections, languages, etc.

When review scores themselves are not available for study but the proposals’ rankings by review scores are (which is the case for SFI data), we instead compare the similarity between two types of ranking positions of the proposals: actual ranking based on review scores and sentiment ranking based on sentiment scores. We use the Spearman rank correlation coefficient to measure the similarity between the two types of rankings. A stronger correlation indicates a higher similarity between the two types of rankings.

RQ2: Do algorithmic SA results correlate with manually coded sentiment scores?

Human coders do not act like algorithms that count words and attribute valency based on a predefined lexicon; instead, they interpret reviewer opinions using knowledge and awareness of context. As a rule of thumb, an achievement of 70% similarity in replicating manual classifications is often considered an acceptable performance (Turney, 2002). Therefore, if our algorithmic results suggest a satisfactory correlation to manual analysis across different subjects and units of analysis, the reliability of some quick and low-cost SA algorithms for peer review studies is supported.

RQ3: Can review sentiment scores predict funding decisions?

Focusing next on proposals as the unit of analysis, we ask how strong the links are between proposals’ review sentiments, measured by the review texts, and their funding decisions (awarded or rejected) made by agencies. If the links are strong, we infer that review sentiment works as a reliable proxy of reviewer opinions to predict proposals’ funding worthiness. To explore this question, we label each proposal as either “awarded” or “rejected” and as either “relatively positive” or “relatively negative.” By “relatively positive/negative,” we refer to the

---

1 VADER dictionary is only supported for the English language. As explained in the next section, one of the two data sets (SNSF) contains review texts in several languages, so not all reviews could be assigned a VADER sentiment score.

2 In a few instances, SNSF reviewers only commented but did not grade, and the officer graded instead based on the review texts. In this study we assume all scores came from reviewers.
sentiment score for each proposal (assigned by aggregating all the review sentiments it received). As funding decisions are usually made upon the relative rather than absolute (perceived) merits of proposals, we rank proposals based on their review sentiment scores from the highest (the most positive) to the lowest (the most negative). We classify the top-ranking proposals as the “relatively positive” group, and the bottom-ranking as “relatively negative”\(^3\), and the threshold line between the two groups is the funding rate of the program.

Ideally, awarded proposals should have received “relatively positive” reviews, whereas rejected proposals should have received “relatively negative” reviews. A standard way to investigate the agreement between two dichotomous variables is a confusion matrix like the one depicted in Table 2 (Kohavi & Provost, 1998). Thus, we construct a confusion matrix for each set of proposals in the same competition pool.

The literature offers several standard metrics calculated on the confusion matrix to measure the predictive power of the model (Kohavi & Provost, 1998). The metrics we borrow include accuracy, precision (i.e., positive predictive value), recall (also known as sensitivity), F1 score (harmonic mean of precision and recall), and negative predictive value, as shown below. Here we use these metrics for each of the three sentiment estimation methods to measure their reliability in predicting proposals’ funding worthiness based on review sentiments. High values on these metrics indicate that review sentiments can provide a reliable proxy of grant reviewer opinions on proposals’ funding worthiness.

\[
\text{Accuracy} = \frac{TP + TN}{TP + TN + FN + FP} \\
\text{Precision} = \frac{TP}{TP + FP} \\
\text{Recall} = \frac{TP}{TP + FN} \\
\text{F1 Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \\
\text{Negative Predictive Value} = \frac{TN}{TN + FN}
\]

Table 2. A confusion matrix captures the relationship between (binary) review sentiment and proposals funding decisions

| Review sentiment        | Relatively positive | Relatively negative |
|-------------------------|---------------------|---------------------|
| Proposal awarded        | TP: True positive scenario | FN: False negative scenario |
| Proposal rejected       | FP: False positive scenario | TN: True negative scenario |

To summarize our design, we use manual content analysis and two different SA algorithms on textual reviews to interpret reviewer opinions on individual subjects and overall proposals’ funding worthiness. We test the reliability of review sentiment as a relatively new proxy of reviewer opinions by addressing three empirical questions using peer review reports and other data from two agencies: SFI and SNSF. The next section will explain the characteristics of our empirical data sets.

4. EMPIRICAL DATA CHARACTERISTICS

Our empirical data set contained peer review reports from three funding programs: two from SFI (Industry Fellowship [SFI:IF] and Investigators Programme [SFI:IvP]), and one from SNSF.

---

\(^3\) Note that there were ties in the ranking (e.g., when two reviews have the same sentiment scores). All proposals that tie for the ranking position that sets the threshold between “relatively positive” and “relatively negative” are treated as “relatively positive.”
(abbreviated as SNSF) that was also investigator-driven. The analyzed documents for the three programs included publicly accessible call documents and evaluation guidelines as well as confidential peer review reports, proposals’ rankings and funding decisions. The three programs had a similar review process where proposals were first evaluated by postal peer reviewers, the reviews synthesized to rank competing proposals, and then discussed in a panel. This panel makes funding recommendations for each proposal as awarded or rejected, which the agency might follow or not. As a reminder, our study focuses on postal reviews to explore independent reviewer opinions.

All SFI and SNSF reviews were redacted by the agency (SFI) or during data curation (SNSF) to deidentify individuals, organizations, or specific research areas. Such redaction was irrelevant for the sentiment estimation by both humans and the SA algorithms, because it does not affect the key parts of a sentence that carry reviewers’ judgments. Both agencies averaged the review scores from individual review sections and from multiple reviewers to obtain an overall score for each proposal. SFI and SNSF applied slightly different grading scales: 1–5 for SFI, 1–6 for SNSF. In both cases, higher scores indicate better quality. All SFI reviews were written in English. For SNSF reviews, we examined those written in three languages: English (29% of the total), German (38%), and French (34%)4.

The three programs (SNSF, SFI:IF, SFI:IvP) differed in their organizational mandates, target applicants, and disciplinary areas. Furthermore, SFI applied standardized three-section structured review reports where each section represented one review subject (applicants, proposed research, and potential for impact). This allowed us to estimate sentiments at the level of individual review sections. By contrast, SNSF review reports covered seven subjects (scientific impact, originality, suitability of methods, feasibility, experience and past performance, specific abilities, and other comments) but did not always apply structured review reports. For the relatively small pool of SFI:IF proposals, only one comprehensive panel was organized by SFI each year to review proposals from different disciplines. We were provided with SFI:IF reviews from the years 2014–2016 as three independent pools. For SFI:IvP and SNSF, different disciplinary panels were always organized in 1 year as independent competing pools. We were provided with the review reports from four different disciplinary panels of SFI:IvP in 2016 (all in STEM fields) and three panels of SNSF including Humanities and Social Sciences (see Table 3). SNSF review data comprised the funding decision for each proposal, their review texts, and scores from which we reconstructed the ranking by review scores (Table 1). By contrast, SFI data comprised funding decisions, review texts, and ranking by review scores, but not the scores themselves.

The sizes and funding rates of the competing pools were consistent within the same programs but varied across programs (see Table 3). Overall, 527 section-separated review reports from SFI and 125 from SNSF were shared with us and analyzed, totaling 2,456 section-level reviews (527 × 3 + 125 × 7) coded by both manual analyzers and algorithmic SA tools. The scale, proportion of our data set in the pool, and composition of the reviews by program, year, language, and disciplinary panel are presented in Table 3. The data sets shared by SFI and SNSF were randomly selected whereas the SFI sample was stratified by funding decisions: approximately 50% awarded and 50% declined.

Lastly, we conducted the algorithmic SA of the 9,532 SNSF statement-level reviews and compared the result with the manual content analysis of the same data set from a previous

---

4 We excluded from our data set two SNSF reviews that were written in other languages not supported in the SA algorithms.
study (Reinhart, 2010). These statement-level reviews were manually coded segments of review texts from unstandardized SNSF review reports, so they did not overlap with section-level reviews. Coding was guided by a qualitative coding scheme containing 22 typical research evaluation subjects, such as “originality” or “methods” (Reinhart, 2010). On average, each review has 14.2 statements, and each statement represents a reviewer opinion on one specific subject. Most of the statements are short review segments (e.g., “This is a small but competent team with a substantial record of achievement”). We excluded those very short segments with fewer than three words, such as “his previous work,” and fed the relatively complete 9,532 reviews to the two SA algorithms.

5. RESULTS

First, we examine the descriptive statistics of the sentiment scores obtained from three different methods (manual coding, TextBlob, and VADER) for each of the three programs (SFI:IvP, SFI:IF, and SNSF). Figure 1 shows the distribution of sentiment scores on section-level reviews—the box plots partition each distribution into quartiles. Review scores (dark red boxplots) were known for SNSF only. Overall, some general differences in the distribution of sentiment scores across the three methods can be observed. Manual coding results show greater variability from “most negative” all the way up to “most positive”: This can be seen in the vertical span of the distribution (blue violins) and of their boxplots (i.e., the interquartile range). By contrast, the SA algorithmic results are always above the “neutral” mark, and TextBlob scores (green) show the least variability.

Review sentiments of different sections (within each program and produced by the same method) can be compared to address reviewer opinions and disagreement with respect to

---

Table 3. Overview of section-separated review samples analyzed in this study

| Program | Year       | No. of reviews | Proportion of data set | Funding rate | Review Language | Disciplinary panel |
|---------|------------|----------------|------------------------|--------------|-----------------|-------------------|
| SFI:IvP | 2016       | 261            | 35%                    | 49%          | English: 100%    | A: Communications, Engineering, Computer Science, Mathematics, and Geoscience |
|         |            |                |                        |              |                 | B: Materials and Chemistry |
|         |            |                |                        |              |                 | C: Life Sciences-Human Health |
|         |            |                |                        |              |                 | D: Life Sciences-Microbiology, Agriculture, Ecology, Biomarkers of Disease |
| SFI:IF  | 2014–2016  | 266            | 46%                    | 50%          | English: 100%    | One comprehensive panel in each year |
| SNSF    | 1998 or 2004\(^5\) | 125\(^6\) | 18%                    | 71%          | English: 28.8%   | E: Humanities and Social Sciences |
|         |            |                |                        |              | French: 33.6%    | F: Mathematics, Natural and Engineering Sciences |
|         |            |                |                        |              | German: 37.6%    | G: Biology and Medicine |

---

\(^5\) 1998: panels E and F; 2004: panel G.

\(^6\) Unstandardized SNSF reviews (i.e., not section-separated) were excluded for some of our analyses.
individual subjects. For instance, reviewer opinions on “proposed research” are overall less positive and more variable than their opinions on “applicant” in the two SFI programs. This is consistent with previous work, where it was found that SNSF reviewers generally made more positive remarks about applicants (experiences, abilities) and more negative remarks about the properties of the projects (originality, methods) (Reinhart, 2010).

Furthermore, some SNSF review sections received on average a lower review score (dark red) than other sections (e.g., “originality” and “suitability of methods”). This indicates that reviewers might have a higher bar for grading on those subjects. Such differences across subjects are better captured by manually coded sentiments than the algorithmic results. In the appendix we also show that, similarly, disciplinary and language differences are captured better by manual coding than by algorithms (Appendix Figures A1, A2).

In the next sections (5.1–5.3) we examine the three research questions sequentially. Note that our results are reported at three levels of analysis: the individual review sections; the reviews as a whole aggregated from separated sections; and the individual statements from

Figure 1. Violin plots and boxplots showing the distribution and quartiles of section-level review scores and review sentiments obtained from the three methods (manual, TextBlob, VADER) for the three programs (SFI:IF, SFI:IfP, SNSF).
5.1. Do Review Sentiment Scores Correlate with Review Scores?

We started by examining the correlation between SNSF review scores and sentiment scores in each of the six review sections7. Figure 2 shows regression lines between the scores assigned by reviewers (x-axis) and the sentiment scores obtained from manual coding, TextBlob, and VADER (y-axis). Overall, manually coded sentiments (blue) correlate positively with review scores for all the sections, and some sections (e.g., feasibility) show stronger correlation than others (e.g., specific abilities). Both algorithmic SA results correlate poorly with review scores.

To examine these correlations more closely, we then calculated the Spearman rank correlation coefficients between the SNSF review scores and sentiment scores (RQ1), and between the sentiment scores produced by the different methods (RQ2). The correlation coefficients were calculated for section-level reviews and whole reviews (sections aggregated) respectively. The correlation results on the two levels are largely consistent, and Table 4 shows the whole review level results.

In Table 4, there is a positive and significant correlation for manual coding to review scores. The correlation is also positive for TextBlob, although the effect is only significant for SNSF Panel F. There is no significant correlation for VADER. These results indicate that manually coded sentiment scores were generally similar to the actual review scores; the two algorithmic SA results were less so.

---

7 We excluded the seventh section “other comments” because many reviewers did not comment. In such cases, the overall review score was averaged from the six sections.
For SFI data the review scores were not available for our study, but the resulting rankings of proposals were. Thus, we compared the similarity between the two types of the proposals’ rankings: actual ranking by review scores and the ranking by sentiment scores produced by different methods. Table 5 shows the Spearman rank correlation coefficients and significance levels for SFI’s seven pools. The results generally show the positive correlation across all the pools and all the methods. Manual coding can produce more similar rankings to the proposals’ actual rankings than the two algorithms, which is consistent with our finding from the SNSF data.

5.2. Do Algorithmic SA Results Correlate with Manually Coded Sentiment Scores?

We turn to comparing the performance of the two algorithmic SAs to manual coding as a benchmark. Figure 3 plots the regression lines and confidence interval between the sentiment scores from algorithmic SA results (y-axis) and the manually coded sentiment scores (x-axis). VADER correlates with the benchmark more strongly than TextBlob. Furthermore, both VADER and TextBlob values sit above the “neutral” mark on the y-axis, which indicates that neither SA algorithm could detect negative reviews as identified by manual coding. This suggests that the SA algorithms are not as reliable as the benchmark to measure individual absolute review sentiments. One reason could be that even the most negative reviews were often written in rather restrained or neutral language. In such cases, algorithms are far less reliable than human coders in detecting the key negative messages out of long, more complicated comments. A fully negative review is very rare in our data set. Appendix Figure A3 shows the correlation between the two SA algorithms and manual benchmark for groups of different programs and different review sections. The results are largely consistent with the overall result, as shown in Figure 3.

We also tested the Spearman correlation coefficients between each of the two algorithms and the manual coding, for each program and their different units of analysis: SNSF review statements (N = 9,532), SNSF section-separated reviews (N = 125), and SFI section-level reviews (N = 527). Overall, the two SA algorithmic results correlate positively and significantly with the manual coding. However, the correlation coefficients are not high, especially for SNSF review statements. Interestingly, VADER results are more similar to manual coding than TextBlob (Table 6), but we found earlier (Tables 4 and 5) that TextBlob results are more similar to actual review scores. The reason for this variability in the relative performance of TextBlob

Table 4. Spearman rank correlation coefficient between SNSF (section-separated) review scores and sentiment scores produced by three methods (*p ≤ 0.05; **p ≤ 0.01; ***p ≤ 0.001)

|                      | SNSF Panel E | SNSF Panel F |
|----------------------|--------------|--------------|
| Manual               | 0.654***     | 0.333**      |
| TextBlob             | 0.22         | 0.319*       |
| VADER n.a.           | 0.128        | 0.128        |

For SFI data the review scores were not available for our study, but the resulting rankings of proposals were. Thus, we compared the similarity between the two types of the proposals’ rankings: actual ranking by review scores and the ranking by sentiment scores produced by different methods. Table 5 shows the Spearman rank correlation coefficients and significance levels for SFI’s seven pools. The results generally show the positive correlation across all the pools and all the methods. Manual coding can produce more similar rankings to the proposals’ actual rankings than the two algorithms, which is consistent with our finding from the SNSF data.

8 SNSF reviews are in different languages, and only English is supported by VADER. This particularly affects SNSF Panel E (Humanities and Social Sciences) where only three out of 61 reviews were written in English. For this reason, we omitted VADER results for SNSF Panel E.
and VADER can be attributed to measurement noise, because VADER results are from English reviews only.

To sum up, the correlations between manual coding as a benchmark and the two SA algorithms respectively are generally positive and significant, but the correlations are not strong. Furthermore, SA algorithms show an important shortcoming: Although they usually can tell a relatively negative review from a relatively positive one, they systematically fail to detect absolute negative reviews—instead, they interpret even the most negative reviews as being at least “neutral.” This is a first hint that the usefulness of dictionary-based SA algorithms for measuring scientific reviews is limited to very specific tasks. For example, TextBlob and VADER are not useful to accurately estimate the absolute sentiment of specific reviews, but they can be of

|                  | SFI:IvP Panel A | SFI:IvP Panel B | SFI:IvP Panel C | SFI:IvP Panel D | SFI:IF Panel 2014 | SFI:IF Panel 2015 | SFI:IF Panel 2016 |
|------------------|-----------------|-----------------|-----------------|-----------------|-------------------|-------------------|-------------------|
| Manual           | 0.753***        | 0.877***        | 0.902***        | 0.94***         | 0.358             | 0.844***          | 0.888***          |
| TextBlob         | 0.427           | 0.691**         | 0.518*          | 0.631**         | 0.197             | 0.666**           | 0.318             |
| VADER            | 0.392           | 0.686**         | 0.546*          | 0.516*          | 0.447*            | 0.57**            | 0.637***          |

Figure 3. Linear regression between manually coded sentiments and algorithmic SA for three programs. Confidence interval: 95%.
help when comparing the relative sentiments between groups of reviews. We will test this for the
two funding decision groups in the next section.

5.3. Can Review Sentiments Predict Funding Decisions?

We explore the link between review sentiments and funding decisions of proposals by first
looking at the distribution of sentiment scores for awarded and rejected proposals for the three
different programs. Overall, the manual coding results in Figure 4 agree with our expectation:
Review sentiment is more positive for awarded proposals. However, only small differences
between the two groups are shown by VADER; almost no noticeable difference is captured
by TextBlob.

To explore this research question further, we examine the four prediction scenarios from the
confusion matrix (TP: true positive, TN: true negative, FP: false positive, and FN: false negative)

|                | SNSF review statements | SNSF section-separated reviews | SFI section-separated reviews |
|----------------|------------------------|-------------------------------|-------------------------------|
| VADER          | 0.3***                 | 0.586***                      | 0.558***                      |
| TextBlob       | 0.195***               | 0.209*                        | 0.63***                       |

**Table 6.** Spearman correlation coefficient between manually coded sentiments and two algorithmic
SA results of two agencies’ reviews (*p ≤ 0.05; ***p ≤ 0.001)

![Figure 4. Review sentiments under different groups of funding decisions for the three programs.](image)
introduced in the research design. For each proposal, we averaged the overall sentiment score of all the reviews it received. Table 7 tallies the four scenarios in each of the three programs (aggregated from multiple pools in different years or disciplines) as well as the five metrics to measure the performance of the prediction model.

Table 7 shows the confusion matrix for each of the three programs as well as the five standard metrics that measure the performance of our prediction model. Overall, the accuracy of the three methods in predicting proposals’ funding decisions ranges from moderate to high. Manual coding still performs the best in all scenarios on all the metrics, and the two algorithms are not far behind. The values of the five metrics are often close to each other, which flags the overall reliability of our prediction model. Even though the two algorithms did not perform well in detecting the individual absolute review sentiments compared with manual coding, they still provide valuable insights for decision-making processes.

### Table 7. Confusion matrices for each of the three programs (aggregated from pools). The observed funding rate in the samples of each program is used as a threshold to set binary classifications (relatively positive or negative sentiment)

#### SFI: IF, \( N = 22 + 20 + 30 = 72 \) (sum of three pools)

|       | Positive sentiment | Negative sentiment | TextBlob | Positive sentiment | Negative sentiment | VADER | Positive sentiment | Negative sentiment |
|-------|--------------------|--------------------|----------|--------------------|--------------------|-------|--------------------|--------------------|
| Funded | TP 30              | FN 6               | Funded   | TP 25              | FN 11              | Funded| TP 26              | FN 10              |
| Rejected | FP 9              | TN 27              | Rejected | FP 11              | TN 25              | Rejected| FP 10              | TN 26              |

Accuracy = \( \frac{(TP + TN)}{(TP + TN + FN + FP)} \)

Precision = \( \frac{TP}{(TP + FP)} \)

Recall = \( \frac{TP}{(TP + FN)} \)

F1 Score = \( \frac{2 \times \text{Precision} \times \text{Recall}}{(\text{Precision} + \text{Recall})} \)

Negative Predictive Value = \( \frac{TN}{(TN + FN)} \)

|       | Positive sentiment | Negative sentiment | TextBlob | Positive sentiment | Negative sentiment | VADER | Positive sentiment | Negative sentiment |
|-------|--------------------|--------------------|----------|--------------------|--------------------|-------|--------------------|--------------------|
| Funded | TP 29              | FN 4               | Funded   | TP 27              | FN 6               | Funded| TP 25              | FN 8               |
| Rejected | FP 6              | TN 27              | Rejected | FP 7               | TN 26              | Rejected| FP 8               | TN 25              |

Accuracy = \( \frac{(TP + TN)}{(TP + TN + FN + FP)} \)

Precision = \( \frac{TP}{(TP + FP)} \)

Recall = \( \frac{TP}{(TP + FN)} \)

F1 Score = \( \frac{2 \times \text{Precision} \times \text{Recall}}{(\text{Precision} + \text{Recall})} \)

Negative Predictive Value = \( \frac{TN}{(TN + FN)} \)
coding, they are fairly reliable in predicting the proposals’ relative positions either above or below the funding lines, especially for the two SFI programs\(^\text{10}\).

Across different programs, predictions were most accurate for SFI:IvP, then SFI:IF, and the least for SNSF. A significant difference that may affect the predictions among the programs is their funding rates; the smallest is SFI:IvP (17% on average in original pools) with much bigger rates for SFI:IF (73%) and SNSF (72%). The funding rate is exogenously imposed, as it depends on the organizational objectives and budget availability and is thus beyond the consideration of peer reviewers. However, the funding rate plays an important role for the agency in setting the threshold between awardable and rejectable proposals, and for us to classify the two groups of proposals as either “relatively positive” or “relatively negative.” For instance, in a large pool with a low success rate (which is often the case), grants will only be awarded to a small number of proposals even if many of them are outstanding.

In our case, due to the generally high quality of proposals as perceived from the reviewers’ comments skewing towards “most positive” in Figure 1, (absolute) negative reviews were rare across the three programs and could explicitly lead to bottom ranking of the proposals. However, the negative reviews in SFI:IvP have higher reliability to predict rejection decisions than the other two programs, as seen by its 87% negative predictive value (by manual coding), the highest indicator among all. Because SFI:IvP’s small funding rate signals that the majority of the proposals would get rejected, negatively reviewed proposals could be more easily excluded from the pool to reduce the workload in the decision-making process. But with a high funding rate, the majority would be funded, so the rejection decision became the minority, and it was uncertain how the rejection decision was made. As in our SNSF case, several rejected proposals actually received quite high grades and positive comments (what we call false positive, or FP cases), which resulted in the lowest negative predictive value among all the programs. Here, the rejection decision made by funders could be based upon some other considerations than reviewer opinions.

\(^9\) VADER was only applied to English written reviews, so the SNSF pools do not have complete VADER score-based ranking.

\(^{10}\) TextBlob predictions appear to be relatively poorer for SNSF reviews than for SFI reviews, especially with respect to the negative predictive value. For SNSF reviews, TextBlob is only moderately better than randomly correct predictions.

---

### Table 7. (continued)

| Program | Positive sentiment | Negative sentiment | TextBlob | Positive sentiment | Negative sentiment | VADER | Positive sentiment | Negative sentiment |
|---------|--------------------|--------------------|----------|--------------------|--------------------|-------|--------------------|--------------------|
| Manual  |                    |                    |          |                    |                    |       |                    |                    |
| Funded  | TP 74              | FN 15              |          | Funded             | TP 67              | FN 22 | Funded             | n.a.\(^9\)        | n.a.               |
| Rejected| FP 17              | TN 19              |          | Rejected           | FP 23              | TN 13 | Rejected           | n.a.               | n.a.               |

\[\text{Accuracy} = (TP + TN)/(TP + TN + FN + FP)\]

\[\text{Precision} = TP/(TP + FP)\]

\[\text{Recall} = TP/(TP + FN)\]

\[\text{F1 Score} = (2 \times \text{Precision} \times \text{Recall})/(\text{Precision} + \text{Recall})\]

\[\text{Negative Predictive Value} = TN/(TN + FN)\]

Accuracy = 74.40% 64.00% n.a.\(^9\) 74.40% 64.00% n.a.\(^9\)

Precision = 81.32% 74.44% n.a. 81.32% 74.44% n.a.

Recall = 83.15% 75.28% n.a. 83.15% 75.28% n.a.

F1 Score = 82.22% 74.86% n.a. 82.22% 74.86% n.a.

Negative Predictive Value = 55.88% 37.14% n.a. 55.88% 37.14% n.a.
We took a closer look at some specific false prediction cases and found different underlying reasons. For instance, in SFI:IvP Panel A, where both TextBlob and VADER had low accurate prediction, we found that the false predictions of the two SA algorithms could be attributed to outliers: Their estimated sentiments were not as positive as manual coders assigned for some awarded proposals (resulting in false negatives) and not as negative for some rejected proposals (resulting in false positives). In both situations, the algorithms performed less accurately than manual coders when dealing with very long or very short reviews. For very short reviews, where one or a few key words represent reviewer opinions, misinterpretation of these words would skew the overall message. For example, for this comment “The PI is a leader in the field of (redacted),” both TextBlob and VADER interpret it as “neutral” (sentiment score = 0) because the sum of the weights of these words is zero. By contrast, manual coders interpreted the comment as “moderate positive,” because being recognized as a “leader” in an academic field is a compliment. For very long comments where reviewers wrote rich technical details and diverse considerations, manual coders attempted to identify the most summative signals such as the sentences after a term of “in summary” or “to conclude” and then to interpret reviewers’ key judgmental opinions from the summative sentences. The SA algorithms still worked by valency attribution of individual words, which would likely misinterpret reviewers’ overall opinions expressed in the long and complex texts.

We notice that another factor for the false prediction cases could be the incomplete pool that this study examined, as shown by the proportion of the data set in Table 3. The unshared or excluded reviews might have had variant opinions from the included reviews and thus have led to a different funding result.

Furthermore, we analyzed the link between sentiments, obtained from manual coding and TextBlob, and the funding decisions of the 9,532 SNSF statement-level reviews. Figure 5 shows that the manual results detect more positive sentiments for awarded proposals than rejected ones, whereas no difference is noticeable for TextBlob. This further supports the higher reliability of the manual coding in predicting funding decisions than TextBlob at the review statement level.

Lastly, we analyzed the funding prediction of the reviews on the 21 specific topics Reinhart (2010) identified and found that for almost all the topics (except “previous affiliation”) the average sentiment scores were significantly higher for awarded proposals than rejected ones. That said, the statement-level reviews are less reliable in this funding prediction analysis than section-level reviews, because the number and weight of statements in each review depend on reviewers’ epistemic and writing styles. In this sense, section-level reviews act as the result of reviewers having been guided to organize their otherwise freeform statements into summative signals for the next stage actors (panelists and decision-makers).

The three research questions, examined as whole, present an interesting puzzle. The manually coded sentiments are shown to correlate positively with both review scores or rankings (RQ1) and funding decisions (RQ3), as we expected. However, the results from the SA algorithms paint a somewhat different picture: TextBlob and VADER sentiments correlate poorly with review scores or rankings, but they do correlate with the funding decisions.

We believe there are two plausible reasons. First, manual coders are better at detecting sentiment: They understand review nuances and can easily pick up cues from very long or

---

11 Note that statements are nested by review texts. The classification of statements in “relatively positive” and “relatively negative” requires further assumptions that make it impossible to directly compare our results for RQ3 between section and statement levels. For these reasons, our replication of RQ3 at the statement level relies only on a plot to show the distribution of sentiment scores for awarded and rejected proposals (Figure 5).
very short reviews; whereas SA algorithms such as TextBlob and VADER analyze individual
words, ignoring their contexts and with several constraints. Thus, it is not surprising that man-
ually coded sentiment correlates more strongly with review scores or rankings than the SA
algorithms: TextBlob and VADER are just less accurate.

Second, the result might be due to the different measurement scales we used for the differ-
ent research questions. When correlating sentiment with scores or rankings, we considered
sentiment as an ordinal variable. However, we then dichotomized the ordinal scale to treat
sentiment as a binary classifier to predict funding decisions (awarded or rejected). Thus, the
lack of correlation between algorithmically measured sentiments and scores or rankings might
signal that SA can produce binary classifications well but cannot produce finer-grained rank-
ings equally well. Take for example a set of reviews that fall into two groups: those that are
relatively positive and those that are relatively negative. If there are large differences in the
sentiment between the two groups but only subtle differences within each group, then senti-
ments would allow for good binary classifications but relatively poorer rankings. Our results
from the SA algorithms suggest that this might be the situation.

6. DISCUSSION

This study contributes to the literature in several ways. Methodologically, it contributes
empirical evidence on the calibration and triangulation of different proxies of reviewer opinions
by using different analytic methods and data sets. First, we compared two SA algorithms with
similar working principles on the same data to check their reliability. Second, we benchmarked
SA algorithms with the manual coding as an independently established form of measurement.
Third, we tested and calibrated the sentiment analysis/interpretation results with related and
supplementary data, such as review scores (when available), rankings and funding results of
proposals. All these data constitute different measurements by different actors (peer reviewers,
panelists, and decision-makers) of the same subject, proposals’ (perceived) funding worthiness.

Our study is the first, to our knowledge, to compare algorithmic and manual coding of peer
review data. Reviews from different disciplines, agencies, and national contexts in our data set
strengthened the generalizability of our methods as well as the findings. We find that the two
SA algorithms (TextBlob and VADER) perform better in detecting proposals’ relative worthiness
than individual absolute review sentiments. The two simple algorithmic tools that we applied
worked quickly and almost as accurately (69–82% accuracy in most of our cases) as manual coders (74–87% accuracy) to detect proposals positioned either above or below the funding line, especially for programs with low funding rates. This informs a cost-benefit analysis to decide whether to apply algorithmic SA tools. The cost of lower accuracy compared to manual analysis can be weighed against the benefit of using algorithms to process large data sets quickly. If the cost, 5–10% loss in accuracy, for instance, is tolerable by users, then SA algorithms can be of use in some specific situations, (e.g., to quickly classify binary groups of “relative positives” and “relative negatives”). Although more advanced SA methods have been developed (such as BERT: see Devlin et al., 2019), it is not clear yet how these would perform with peer review reports, especially for data with heavy information redaction (for anonymity in our case). The exploration of other SA methods suggests some potential venues for future research.

Our findings support the reliability of using review sentiment to predict proposals’ funding decisions, especially for the manual analysis method and for programs with low funding rates. The funding rate is shown to play an important role in agencies’ decision-making (Roebber, & Schultz, 2011; Van den Besselaar et al., 2018) and are often beyond peer reviewers’ considerations. We find that a small funding rate in a pool of generally high-quality applications (such as in our SFI:IVP case) signals intensified competition and leads to more accurate predictions of review sentiments than other programs with higher funding rates. Especially, negative reviews can predict rejection decisions in an intensive competing pool with high accuracy. This finding supports the study of Van den Besselaar et al. (2018) on the ERC program (also with a low funding rate) where they found that negative comments had a stronger effect on funding decisions than positive ones.

Our study provides the potential to explore at least three understudied subjects in the peer review literature. The first is the dependency between a reviewer’s grading and commenting behaviors. This can be studied by the correlation between review scores and review sentiments, as we did in RQ1. Strong correlation can be expected if a reviewer gives a score first and then normalizes the comments around the score, or the other way around. In this sense, either of the two proxies (score and sentiment) can be considered as an alternative proxy of the other to represent reviewer opinions. Weak correlation, as we found out, does not necessarily indicate the unreliability of sentiment estimation methods but may indicate heterogeneity between a reviewer’s grading and commenting behaviors.

The second understudied subjects are “grading heterogeneity” (Morgan, 2014) and “commenting heterogeneity” between different peers that have been found in our data sets: The former refers to reviewers of the same proposal with similar opinions but different grades; the latter with different opinions but the same grades. These cases resonate with the literature on interpersonal differences in the understanding of a grading language and of evaluation criteria (Abdoul, Perrey et al., 2012; Lee, Sugimoto et al., 2013). Research suggests that these differences can skew the outcome of the proposals’ evaluations (Feliciani, Moorthy et al., 2020).

Third, such “grading heterogeneity” and “commenting heterogeneity” can also be explored between peers and panelists from different review stages. A generally accurate prediction of postal peer review sentiments in proposals’ funding decisions, across different programs in our data sets, argues that the panelists and decision-makers followed most of the peer reviewer opinions on proposals’ funding worthiness. Only a small proportion of proposals that received “relatively positive” reviews from peer experts were not approved by either panelists who synthesized and discussed the peer reviews or decision-makers who may have other (administrative or policy) considerations. Our study suggests the potential to research (dis)agreement between peers and later stage panelists by measuring not only review scores but also review
sentiments, especially on specific review subjects (Bethencourt, Luo, & Feliciani, 2021). With both sentiment and scores as available research data, we can scrutinize some of the findings of previous research that measure interreviewer reliability by using review scores only (e.g., Pina et al., 2021). Compounding review score data and review sentiment data, especially for individual evaluation subjects, can mitigate the risk of finding false agreement and further support research on interreviewer reliability by considering reviewers’ grading heterogeneity and commenting heterogeneity on different subjects.

Our study also has several limitations. First, many of the differences between the two funding agencies that may have influenced the review data were not addressed in the study, such as the year, country, language of the data resources, and reviewer identities (domestic or international), as well as many other contextual differences between the two agencies. Second, we only applied simple tools for both the SA and statistical analyses. Third, even though the scale of our analyzed data sets (2,456 section-level reviews and 9,532 statement-level reviews) is quite large, especially for manual analysis, the incomplete pools (less than 50%) that this study examined may affect our testing of the reliability of the three methods. Thus, the results from the samples may not generalize to the complete pools. For instance, the difference in funding rates between our sample and the original pool is important to classify the “relative positive” and “relative negative” reviews and to calculate the performance of the prediction model. The behavior of binary classifiers based on SA with different funding rates is untested in this study and deserves further studies with more complete data sets.

7. CONCLUSION

We consider this study as an early exploratory academic work with some potential methodological and practical insights. The grant review process is important and complex but understudied. Data are not widely available, which inhibits research (Lee & Moher, 2017; Shankar, Luo, & Ma, 2020; Squazzoni, Brezis, & Marušić, 2017). Comparative empirical research across systems has been especially difficult. Simulation research on peer review has been a promising approach where the lack of empirical data is hindering progress (Roebber, & Schultz, 2011; Feliciani, Luo et al., 2019). Automated methods for supporting journal peer review have been suggested as far back as the 1980s (Garfield, 1987). For the pre-peer-review screening stage, a trained machine-learning system could quickly check formatting, language, and expression of submissions; the results accurately predicted the “accepted” or “rejected” review outcomes (Checco, Bracciale et al., 2021). The National Natural Science Foundation of China uses automated tools to reduce the load on the selection of review panels (Cyranoski, 2019). But these tools have not been formally adopted (Brezis & Birukou, 2020) in any widespread manner, and we do not intend to suggest that they should be.

However, we believe in the potential benefits of combining qualitative research and algorithmic tools (such as computational simulations) for peer review research. As our results demonstrated, even simple algorithmic SA tools have the potential to provide a quick addition to the existing toolkit of quantitative, qualitative, and modeling approaches to study peer review further. Far more research is needed before making any concrete policy recommendations for the use of automated tools for practice. In peer review practice, human analyzers and decision-makers can never be replaced but can potentially be informed and supported by automated tools that have been trained, tested, and calibrated to work reliably under specific conditions and for specific situations.

Future research can further explore ways to test and calibrate the reliability and generalizability of our empirical findings with more complete and broader data from agencies with
different funding strategies and review processes. With sentiments as another supplementary proxy of reviewer opinions beyond numeric scores, interreviewer reliability between different subjects, between independent peer reviewers of the same proposals, and between postal and panel reviewers, can be quickly reported to funding agencies. Consequently, sentiment as supplementary data has the potential to help improve the validity and transparency of grant peer review systems. Of course, the costs (to conduct textual analysis), risks (e.g., algorithms’ lower accuracy than humans), and challenges (e.g., applicants may not trust sentiment analysis at all) would have to be addressed in future studies and practical pilots if possible.

ACKNOWLEDGMENTS

We thank Pablo Lucas and Lai Ma for their comments on earlier versions of the paper; Niamh Russell for her methodological advice; Nikita Sorgatz for his support in data curation; and the Science Foundation Ireland (SFI) and the Swiss National Science Foundation (SNSF) for access to archival data. We also thank the two anonymous reviewers who made constructive comments and suggestions.

AUTHOR CONTRIBUTIONS

Junwen Luo: Conceptualization, Data curation, Formal analysis, Investigation, Writing—original draft, Writing—review & editing. Thomas Feliciani: Conceptualization, Formal analysis, Investigation, Visualization, Writing—review & editing. Martin Reinhardt: Conceptualization, Data curation (SNSF data), Investigation, Writing—review & editing. Judith Hartstein: Conceptualization, Data curation (SNSF data), Investigation, Writing—review & editing. Vineeth Das: Data curation (SFI and SNSF data), Investigation. Olalere Alabi: Data curation (SFI data), Investigation. Kalpana Shankar: Conceptualization, Investigation, Project administration, Writing—review & editing.

COMPETING INTERESTS

The authors have no competing interests. The two funding agencies, SFI and SNSF, did not influence the research design, analyses, or writing of this article.

FUNDING INFORMATION

This study is funded by Science Foundation Ireland (SFI) under Grant No.17/SPR/5319.

DATA AVAILABILITY

We are not allowed to share the review data due to our research agreement with the two funding agencies, but our scripts for data analysis are available on GitHub at https://github.com/thomasfelici/SFI-and-SNSF-review-sentiment.

REFERENCES

Abdoul, H., Perrey, C., Amiel, P., Tubach, F., Gottot, S., & Durand-Zaleski, I. (2012). Peer review of grant applications: criteria used and qualitative study of reviewer practices. *PLOS ONE*, 7(9), 1–15. https://doi.org/10.1371/journal.pone.0046054, PubMed: 23029386

Bethencourt, A. M., Luo, J., & Feliciani, T. (2021). Bias and truth in science evaluation: A simulation model of grant review panel discussions. *Proceedings of the Workshop Reducing Online Misinformation through Credible Information Retrieval (ROMCIR 2021) Co-Located with the 43rd European Conference on Information Retrieval (ECIR 2021)* (pp. 16–24). Retrieved from https://ceur-ws.org/Vol-2838/paper2.pdf

Brezis, E. S., & Birukou, A. (2020). Arbitrariness in the peer review process. *Scientometrics*, 123, 393–411. https://doi.org/10.1007/s11192-020-03348-1

Buljan, I., Garcia-Costa, D., Grimaldo, F., Squazzoni, F., & Marušič, A. (2020). Meta-research: Large-scale language analysis of peer review reports. *eLife*, 9, e53249. https://doi.org/10.7554/eLife.53249, PubMed: 32678065
Analyzing sentiments in peer review reports

Feliciani, T., Moorthy, R., Lucas, P., & Shankar, K. (2020). Grade Hutto, C., & Gilbert, E. (2014). VADER: a parsimonious rule-based
H2020 Evaluation Guidelines
Feliciani, T., Luo, J., Ma, L., Lucas, P., Squazzoni, F., Maru
Cicchetti, D. (1991). The reliability of peer review for manuscript
and grant submissions: A cross-disciplinary investigation. Behavioral and Brain Sciences, 14(1), 119–135. https://doi.org/10.1017/S0140525X00065675
Cole, S., Cole, J. R., & Simon, G. A. (1981). Chance and consensus in peer review. Science, 214(4523), 881–886. https://doi.org/10.1126/science.7302566, PubMed: 7302566
Cyranoski, D. (2019). Artificial intelligence is selecting grant reviewers in China. Nature, 569, 316–317. https://doi.org/10.1038/d41586-019-01517-8, PubMed: 31089232
Devlin, J., Chang, M. W., Lee, K., Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 Conference of the North America Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 1 (pp. 4171–4186. https://doi.org/10.18653/v1/N19-1423
European Commission. (2015). H2020 Evaluation Guidelines. https://ec.europa.eu/research/participants/data/ref/h2020/wp/2014_2015/annexes/h2020-wp1415-annex-h-esacrit_en.pdf. Accessed September 2021.
Feliciani, T., Luo, J., Ma, L., Lucas, P., Squazzoni, F., Marušić, A., & Shankar, K. (2019). A scoring review of simulation models of peer review. Scientometrics, 121(1), 555–594. https://doi.org/10.1007/s11192-019-03167-z, PubMed: 31564758
Feliciani, T., Moomty, R., Lucas, P., & Shankar, K. (2020). Grade language heterogeneity in simulation models of peer review. Journal of Artificial Societies and Social Simulation, 23(3), 8. https://doi.org/10.18564/jasss.4284
Fogelholm, M., Leppinen, S., Auvinen, A., Raitanen, J., Nuutinen, A., & Väänänen, K. (2012). Panel discussion does not improve reliability of peer review for medical research grant proposals. Journal of Clinical Epidemiology, 65(1), 47–52. https://doi.org/10.1016/j.jclinepi.2011.05.001, PubMed: 21831594
Garfield, E. (1987). Refereeing and peer review: How the peer review of research grant proposals works and what scientists say about it. Essays of an Information Scientist, 10, 21–26. Retrieved from https://www.garfield.library.upenn.edu/essays/v10p027y1987.pdf
Hassan, S., Aljohani, N., Idrees, N., Sarwar, R., Nawaz, R., ... Herrera, F. (2020). Predicting literature’s early impact with sentiment analysis in Twitter. Knowledge-Based Systems, 192, 105383. https://doi.org/10.1016/j.knosys.2019.105383
Hartmann, I., & Neidhardt F. (1990). Peer review at the Deutsche Forschungsgemeinschaft. Scientometrics, 19, 419–425. https://doi.org/10.1007/BF02020704
Hirschauer, S. (2010). Editorial judgments: A praxeology of “Voting” in peer review. Social Studies of Science, 40(1), 71–103. https://doi.org/10.1177/0306312709335405
Hutto, C., & Gilbert, E. (2014). VADER: a parsimonious rule-based model for sentiment analysis of social media text. Proceedings of the International AAAI Conference on Web and Social Media, 8(1), 216–225. Retrieved from https://ojs.aaai.org/index.php/ICWSM/article/view/14550
Kohavi, R., & Provost, F. (1998) Glossary of terms. Machine learning—Special issue on applications of machine learning and the knowledge discovery process. Machine Learning, 30, 271–274. https://doi.org/10.1023/A:1017181826699
Kretzenbacher, H. L., & Thurmai, M. (1992). Textvergleich als Grundlage zur Beschreibung einer wissenschaftlichen Textsorte: Das Peer Review. In Klaus D. Baumann & Hartwig Kalverkämper (Eds.), Kontrastive Fachsprachenforschung, pp. 135–146. Forum für Fachsprachenforschung.
Kretzenbacher, H. L., & Thurmai, M. (1995). ‘... sicherlich von Interesse, wenngleich...’ Das Peer Review als bewertende Textsorte der Wissenschaftssprache. In Heinz L. Kretzenbacher & Harald Weinrich (Eds.), Linguistik der Wissenschaftssprache, pp. 175–216. Berlin: de Gruyter.
Kretzenbacher, H. L. (2017). “The wording is on occasion somewhat emotional”: A qualitative study of English and German peer reviews for a chemical journal. Fachsprache, 39(1–2), 59–73. https://doi.org/10.24989/fs.v34i1-2.1261
Lamont, M. (2010). How professors think: Inside the curious world of academic judgment. Cambridge, MA: Harvard University Press. https://doi.org/10.1177/0162243908329381
Langefeld, L., & Scordato, L. (2016). Efficiency and flexibility in research funding. A comparative study of funding instruments and review criteria. Report published by Nordic Institute for Studies in Innovation, Research and Education (NIFU), ISBN 978-82-327-0184-4
Lee, C. J., Sugimoto, C. R., Zhang, G., & Cronin, B. (2013). Bias in peer review. Journal of the American Society for Information Science and Technology, 64(1), 2–17. https://doi.org/10.1002/asi.22784
Lee, C. J., & Moher, D. (2017). Promote scientific integrity via journal peer review data. Science, 357(6348), 256–257. https://doi.org/10.1126/science.aan4414, PubMed: 28729501
Liu, B. (2010). Sentiment analysis and subjectivity. In N. Indurkhya & F. J. Damara (Eds.), Handbook of Natural Language Processing (2nd ed.). Chapman and Hall.
Liu, X. Z., & Fang, H. (2017). What we can learn from tweets linking to research papers. Scientometrics, 111, 349–369. https://doi.org/10.1007/s11192-017-2279-0
Loria, S., Keen, P., Honnibal, M., Yankovsky, R., Karesh, D., & Dempsey, E. (2014). TextBlob: Simplified text processing. Secondary TextBlob: Simplified Text Processing. https://textblob.readthedocs.io/en/dev/
Loper, E., & Bird, S. (2002). NLTK: The natural language toolkit. arXiv preprint cs/0205028.
Luo, J., Alabi, O., Feliciani, T., Lucas, P., & Shankar, K. (2020). Peer reviews’ prediction in proposals’ funding success: A sentiment analysis of grant reviews at Science Foundation Ireland. PEER International (Virtual) Conference on Peer Review 2020, 30 September–1 October.
Ma, L., Luo, J., Feliciani, T., & Shankar, K., (2020) How to evaluate ex ante impact of funding proposals? An analysis of reviewers’ comments on impact statements. Research Evaluation, rva022, 1–10. https://doi.org/10.1093/reseval/rva022
Mallard, G., Lamont, M., & Guetzkow, J. (2009). Fairness as appropriateness: Negotiating epistemological differences in peer review. Science Technology & Human Values, 34, 573–606. https://doi.org/10.1177/0162243908329381
Morgan, M. G. (2014). Use (and abuse) of expert elicitation in support of decision making for public policy. Proceedings of the National Academy of Sciences, 111(20), 7176–7184. https://doi.org/10.1073/pnas.1319946111, PubMed: 24821779
Obrecht, M., Tibelius, K., & D’Aloisio, G. (2007). Examining the value added by committee discussion in the review of applications for research awards. Research Evaluation, 16(2), 79–91. https://doi.org/10.1177/09582502070223785
Pier, E. L., Brauer, M., Filut, A., Kaatz, A., Raclaw, J., ... Carnes, M. (2018). Low agreement among reviewers evaluating the same NIH grant applications. Proceedings of the National Academy of Sciences, 115(25), 6408–6411. https://doi.org/10.1073/pnas.1716490115
...
Figures A1 and A2 show the overall distribution (in violin plots) and the quartiles (in boxplots) of the review scores and sentiment scores of SNSF reviews under two disciplinary panels (A1) and three review languages (A2).

Figure A1 involved 60 reviews from Humanities and Social Sciences panels and 65 reviews from Mathematics, Natural and Engineering Sciences panels. Note that for the Biology and Medicine panel none of the reviews are section-separated, so we only include this panel for statement-level analysis.

SNSF reviews by panel

Figure A1. Violin plots and boxplots showing the distribution and quartiles of the SNSF review scores and sentiments as measured by three methods under the two disciplinary panels.
Figure A2 involved 36 English reviews, 47 German reviews, and 42 French reviews. Note that review scores and VADER scores were only available for SNSF reviews written in English.

Figure A3 shows the linear regression between manually coded sentiment and the two SA algorithms for all review sections for each of the three programs (A: SNSF, B: SFI:IVP, C: SFI:IF). Together with Figure 3, Figure A3 shows that VADER correlated more strongly than TextBlob across all the programs and almost all the review sections.
Analyzing sentiments in peer review reports

Figure A3. Linear regression between manually coded sentiment and algorithmic SA for section-level reviews for three programs. Confidence interval: 95%.