Peer Reviewing Revisited: Assessing Research with Interlinked Semantic Comments

Cristina-Iulia Bucur
Vrije Universiteit Amsterdam
Amsterdam, The Netherlands
c.i.bucur@vu.nl

Tobias Kuhn
Vrije Universiteit Amsterdam
Amsterdam, The Netherlands
t.kuhn@vu.nl

Davide Ceolin
Centrum Wiskunde & Informatica
Amsterdam, The Netherlands
davide.ceolin@cwi.nl

ABSTRACT

Currently, the field of scientific publishing seems to face a major crisis. On the one hand, the paradigm of publishing has stayed basically the same for 300 years, and now the increasing volume of articles published every day makes it very hard for scientists to stay up to date in their respective fields. On the other hand, many have pointed out serious flaws of current scientific publishing practices, including with respect to the accuracy and efficiency of the reviewing process. To address some of these problems, we apply here the general principles of the Web and the Semantic Web to scientific publishing, focusing on the reviewing process. We want to determine if a fine-grained model of the scientific publishing workflow can help us make the reviewing processes better organized and more accurate, by ensuring that review comments are created with formal links and semantics from the start. Our contributions include a novel model called Linkflows that allows for such detailed and semantically rich representation of reviews and the reviewing processes. We evaluate our approach in the reviewing context on a manually curated dataset from several recent Computer Science journals and conferences that come with open peer reviews. In addition to that, we gathered ground-truth data by contacting the original reviewers and asking them to categorize their own review comments according to our model. Comparing this ground truth to answers provided by model experts, peers, and automated techniques confirms that our approach of formally capturing the reviewers' intentions from the start prevents substantial discrepancies compared to when this information is later extracted from the plain-text comments. In general, our analysis shows that our model is well understood and easy to apply, and it revealed the semantic properties of such review comments.

KEYWORDS

Peer reviewing, scientific publishing, Linked Data, Semantic Web

1 INTRODUCTION

Scientific articles and peer reviews are at the core of communicating and assessing scientific research and discoveries. For the last 300 years, such scientific articles have played a crucial role in facilitating scientific progress, together with the process of peer reviewing for ensuring the quality and integrity of scientific contributions. While communication has changed dramatically in the digital age in almost all fields of life, the paradigm of scientific publishing has remained remarkably stable. Journals, articles, and peer reviews are now mostly produced and consumed in digital form, but their structure and appearance has mostly stayed the same.

Although alternative ways of scientific publishing have been explored, most of the suggested approaches still involve publishing in the form of a large bulk of text in natural language, digitized and maybe semantically enriched, but following the same paradigm of publishing as in the past. This might not be a problem by itself, but looking at the general statistics in the scientific publishing world, we seem to be facing a crisis of information overload [24] with an increasing volume of articles being published every day, as the number of publishing scientists worldwide increases with approximately 4-5% per year [43]. This entails that scientists need more time to stay up to date with the recent developments in their field. For example, in 2004, in the field of primary care, it took over 20 hours per week for epidemiologists to be able to read all the new articles in the field [12].

Moreover, we are also facing challenges with respect to the quality of research. Peer reviewing is a key pillar of the scientific publishing system and the main method of quality assessment. However, despite this established and central role, it is a hotly debated topic in the scientific community due to issues like taking unreasonable amounts of time, lack of transparency, inefficiency, lack of software integration, lack of attribution for reviewers, unsystematic quality assessment, and even poor scientific rigor [4]. Actually, the agreement level between peer reviewers has been found to be only slightly better than what one would expect by chance alone [28].

The ideas and technologies of the Web and the Semantic Web can be used to tackle some of these problems by allowing for a transparent and accessible medium for publication where comments and criticism can be correctly attributed and immediately published in a structured manner. In this research we propose to apply the general principles of the Web and Semantic Web to the reviewing process at a finer-grained scale. These principles entail the use of dereferenceable HTTP URIs to identify, lookup and access resources on the Web, and to interlink them, thereby forming a large network of interconnected resources (or "things"). Or, in Tim Berners-Lee's words [7], forming "a web of data that can be processed directly and indirectly by machines" and humans alike. In such a Web-based system, which does not depend on a central authority, everybody can say anything about anything. To put all these statements into context, we can moreover capture provenance information that will allow us to judge the quality and reliability of the provided information. Therefore, following these principles, we can represent information in the form of immutable nodes in a Web-wide distributed network and accurately track the history and provenance of each snippet of information.

This approach of semantically modeled nodes in a network containing snippets of information can then also use reasoning techniques to aggregate information in dynamic views, instead of the static representations of classical journal articles. In this context, peer reviewing can be performed on a smaller, finer-grained scale, each typically addressing just a small snippet of scientific content. These review comments come with explicit links to the specific part (e.g. a paragraph) of the scientific contribution they are about, thereby making reviews more precise. This structure enables a more detailed and finer-grained quality assessment of both the scientific contributions and of the reviews themselves.

2 RELATED WORK

In the Semantic Web context, considerable research was done in enriching the meaning of a traditional article in the digital publishing environment. This can facilitate the automatic discovery of the article and linking to other related articles or other related parts of articles in a semantic way. Especially notable in this sense are the SPAR ontologies [1], the ontologies central to the task of semantic publishing, PRISM (the Publishing Requirements for Industry Standard Metadata) [2] that uses metadata terms to describe published works, and SKOS (the Simple Knowledge Organization System) [3] that uses a RDFS model for knowledge organization systems.

However, all these ontologies mainly consider the scientific article still as a big bulk of text, so problems like addressing the "information overload" issue or "Knowledge Burying" [31] in the article text itself still remain. To solve this, certain solutions were proposed based on data mining techniques like extracting patterns from text, the most common in the field of publishing being association rule mining, while for large knowledge bases that use RDF, one of the most frequently used methods is Inductive Logic Programming [21]. SWARM allows the automated mining of Semantic Association Rules from RDF data [2]. There is also a research that uses two Semantic Web models, the Micropublications ontology and the Open Annotations Data Model, on knowledge bases containing information about the drug-drug interactions and links the assertions from these knowledge bases to evidence (statements in documents and scientific articles, data, materials and methods) [34]. Nanopublications were created for the explicit representations of statements made in scientific literature to provide "fine-grained" information in RDF form [20] and extensions to these simple statements, nanopublications, were done to enlarge their application domain by using AIDA English sentences [14]. When more detailed and nuanced assertions are needed, the micropublications model can be complementary used together as it is compatible with the statement-based model of nanopublications [42]. All these support the idea of "genuine semantic publishing" [23] and publishing at a more finer-grained level.

In terms of assessing the quality of scientific publications, one of the most widely used indicator in the last 40 years is the Journal Impact Factor (JIF) [18, 26], but this metric has been the subject of extensive debates, as it was shown that it can be manipulated [22], and has problems like skewness of citations, false precision, absence of confidence intervals, and the asymmetry in the calculation [8]. Also, the JIF can be biased towards journals that publish high number of non-research items (e.g. research notes, comments) and have higher publishing numbers [1]. So, new methods for rating the quality of scientific publications are needed. Semantic Web technologies with ontologies like the Dataset Quality Information (daQ) [9] can support better and unbiased measures of quality, while new dimensions of quality that consider these Linked Data and Semantic Web technologies need to be taken into account [11].

A deciding factor for the publication of a scientific article is the evaluation by the peers in the field, the peer reviews. While this system for evaluating the quality of research has been used for a long period of time, researchers have also outlined the flaws of this process [36], advocating for a change. Some point out that there is no conclusive evidence in favor of peer reviews [27], that the existing evidence actually corroborates flaws in this system [37], and that biases and preconceptions in judgments [3] affect the overall evaluation of quality that the peer-reviewers make. While proposals like structuring more the peer reviews [13, 40] or even the suggestion to pay the peer reviewers [10] as to ensure a higher quality of peer-reviews exist, there is still increasing debate about this and no clear solution. One approach hints at making reviews more fine-grained [33], while the FAIR Reviews ontology [4] can be used for more structured reviews.

As datasets, documents and knowledge in general are spread on the Web, where everything can be shared and reused and linked, decentralization is a key concept. Decentralization implies that no central authority — like a publishing house — has global control over the content or participants of the system. A large amount of research has investigated the prerequisites and consequences of decentralization in general [35], and to a lesser extent specifically in the context of scientific publishing. The former includes recent initiatives like Solid [5], a completely decentralized linked-data framework over the Web. The latter include approaches to ensure the functioning of a secure and decentralized global file system [30], the application of the BitTorrent peer-to-peer file sharing protocol to distribute and access scientific data [29], and decentralized data publishing approaches specifically designed for RDF data [16], nanopublications [15], Dokieli [6], a decentralized client editor that allows for the decentralized article editing, and blockchain-based initiatives.

3 APPROACH

In this work, we aim to answer the following research question: *Can an approach for scientific publishing based on a fine-grained semantic model help to make reviewing better structured and more accurate?*

3.1 General Approach

In our approach, we apply the general principles of the Web and the Semantic Web to the field of scientific publishing. While the Web consists of a distributed network of documents that link to each other, the Semantic Web adds to that a more fine-grained network at the level of data and knowledge, where the nodes are concepts, domain entities, not documents. Following the principles of the Web and the Semantic Web, we argue for fine-grained publication of scientific knowledge in the form of small distributed knowledge snippets that form the nodes in a network, to replace the long

[1] http://www.sparontologies.net/
[2] https://www.idealliance.org/prism-metadata
[3] https://www.w3.org/2004/02/skos/
[4] http://fairreviews.linkeddata.es/def/core/index.html
[5] https://solid.inrupt.com/

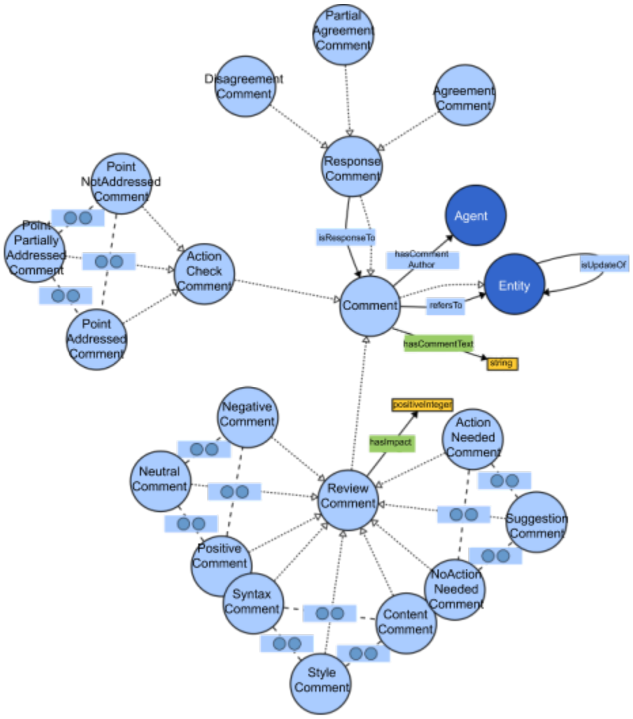


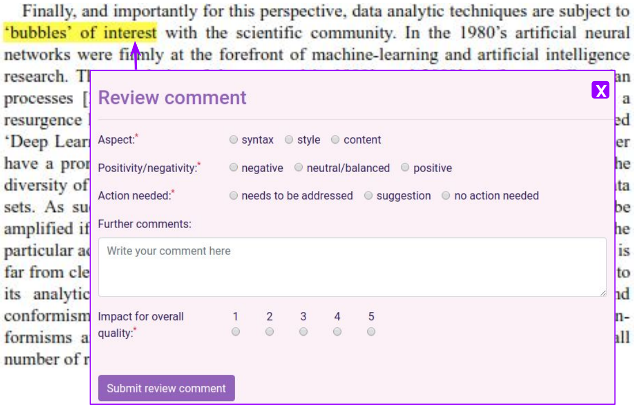
Figure 1: Linkflows model for reviewing. Visualization with the online WebVOWL tool [7].

prose texts in formats such as PDF. These knowledge snippets are represented with formal semantics, and identified and located with the help of dereferenceable URIs. In this way, scientific knowledge can be shared and accessed as a continuously growing decentralized network of machine-readable snippets of scientific contributions, instead of a static and machine-unfriendly collection of long texts.

3.2 Linkflows Model of Reviewing

At the core of our approach, we propose an ontology for granular and semantic reviewing. Figure 1 shows the main classes and properties of this ontology that we call the Linkflows model of reviewing. The formal ontology specification can be found online.[6]

The main class of the ontology is the *Comment* class, which includes review comments (subclass *ReviewComment*), on which we focus here, but the class also includes general text annotations or any kind of comment about a text snippet that comes with a dereferenceable URI. The general properties of the model are *refersTo*, which connects a comment to the entity the comment is about, *isResponseTo*, which declares that a comment is a response to another comment, *isUpdateOf*, which connects an entity such as a comment to its previous version, *hasCommentText*, which links to the text content of a comment, and *hasCommentAuthor*, which declares the person who wrote a comment.


Finally, and importantly for this perspective, data analytic techniques are subject to ‘bubbles’ of interest with the scientific community. In the 1980’s artificial neural networks were firmly at the forefront of machine-learning and artificial intelligence

Review comment

Aspect:* ○ syntax ○ style ○ content

Positivity/negativity:* ○ negative ○ neutral/balanced ○ positive

Action needed:* ○ needs to be addressed ○ suggestion ○ no action needed

Further comments:

Write your comment here

Impact for overall quality:* 1 ○ 2 ○ 3 ○ 4 ○ 5 ○

Figure 2: Mock interface implementing the Linkflows model for reviewing.

Our model defines three dimensions for review comments with different categories, each defined as subclasses of the class *ReviewComment*, which form the core of our semantic representation of reviews. The first dimension is about whether the point raised in the review comment is about individual spelling or grammar issues (*SyntaxComment*), the general style of the text including text structure, text flow, text ordering, and consistent wording (*StyleComment*), or the content of the text, e.g. about missing literature, the presented arguments, or the validity of the findings (*ContentComment*). The second aspect is the positivity/negativity of the review comment: *PositiveComment* for review comments that mainly raise positive points, *NeutralComment* for neutral or balanced points raised, and *NegativeComment* for the cases with mainly negative points. The third dimension captures whether an action is needed in response to the review comment (according to the reviewer): *ActionNeededComment* means that the reviewer thinks his or her comment necessarily needs to be addressed by the author; *SuggestionComment* stands for comments that may or may not be addressed; and *NoActionNeededComment* represents the comments that do not need to be addressed, such as plain observations. On top of that, we define a datatype property *hasImpact* that takes an integer from 1 to 5 to represent the extent of the impact of the point raised in the review comment on the overall quality of the article according to the reviewer. This can be on a scale from 1 to 5. For negative comments this score indicates what would be the positive impact if the point is fully addressed, while for positive points it indicates what would be the negative impact if this point were not true. To further clarify these dimensions and how they could be implemented in an actual reviewing system, we show a mockup of such an interface in Figure 2.

For representing the interaction between reviewers and authors and follow-up actions that are necessary or requested by the reviewer, the *ResponseComment* and *ActionCheckComment* were created as subclasses of *Comment*. The author of the text snippet that was commented upon can agree, disagree or partially agree with the review comment of the reviewer and they can indicate this classifying their response comment accordingly as *AgreementComment*, *PartialAgreementComment*, or *DisagreementComment*.

[6] https://github.com/LaraHack/linkflows_model/blob/master/Linkflows.ttl
[7] http://www.visualdataweb.de/webvowl/#iri=https://raw.githubusercontent.com/LaraHack/linkflows_model/master/Linkflows.ttl

Finally, reviewers or editors can indicate in another follow-up comment whether they think the the author indeed addressed the point raised by the reviewer (to which they might have agreed, disagreed or partially agreed). This can be expressed by the subclasses *PointAddressedComment*, *PointPartiallyAddressedComment*, and *PointNotAddressedComment*.

4 EVALUATION DESIGN

We evaluate our approach based on recently published articles that come with open peer reviews. We use these data to simulate how the Linkflows model for reviewing would have worked, had this fine-grained semantic model been applied to these manuscripts from the start. Specifically, we performed the following steps: First, we created a dataset consisting of 35 articles with open reviews and rebuttal letters (where available). Second, the review snippets were manually annotated by model experts by applying our model. Third, we asked reviewers of the selected articles to rate one of their own review comment with regard to our model, thereby giving us ground-truth data. Fourth, we asked experts via a questionnaire to apply our model to the review snippets for which we have ground-truth answers. Fifth and lastly, we applied different automated sentiment analysis methods on the review comments to see how well these automated methods would have been able to determine their positivity or negativity expressed by the review comments for comparison. These five steps are explained in more detail below.

4.1 Dataset

In order to study how well our model can capture the reviewing process, we needed a dataset of manuscripts and their reviews. For that, we selected journals and conferences in Computer Science that make their reviews openly available in a non-anonymous way. Because this data preparation involves a lot of manual work, we had to restrict ourselves to a subset of all available articles, and a subset of the reviewers they had.

Specifically, we started by considering all the 38 articles published in 2018 in the Semantic Web Journal (SemWeb)[8], the 13 articles published in the first edition of the journal Data Science (DS) [9], and the 25 articles published in 2018 in the PeerJ in Computer Science Journal (PJCS) [10]. Additionally, we collected data from two conferences where article versions and reviews are openly accessible via the openreview.net platform [39]. The only conferences in openreview.net that had a complete number of submissions (all articles and reviews) uploaded on the platform at the time we created our dataset were two workshops in 2018: Decentralizing the Semantic Web from the International Semantic Web Conference (ISWC-DeSemWeb) [11] and the International Workshop on Reading Music Systems from the International Society for Music Retrieval Conference (ISMIR-WoRMS) [12]. ISWC-DeSemWeb had 10 submissions and ISMIR-WoRMS had a total of 12 submissions.

From this set of articles, we first filtered out the ones that were eventually rejected (we had to focus on accepted papers, because open-review data for rejected submissions is even much more scarce), had only anonymous reviewers (because of the ground truth step below), or where one of the authors of this paper was editor or reviewer (for objectivity reasons). After this first filtering step, we selected seven articles from each of the five data sources (the three journals and the two workshops), resulting in a dataset of 35 articles. This selection was done randomly by applying the SHA-256 hash function on their title strings and picking the seven articles with the smallest hash values. We then applied the same procedure on the reviewer names to select a random non-anonymous peer review for each of the selected papers. We always picked the first-round reviews for the papers that went over multiple rounds of reviews.

Finally, we took the resulting reviews and split them into smaller, finer-grained snippets of review comments according to our model. Each of these snippets covers a single point raised by the reviewer, and they often correspond to an individual paragraph or an individual list item of the original review. We did an initial annotation of the review comments corresponding to the 35 selected articles and reviews with regard to the target of the review comments. This target of a review comment can be (a) the entire article, (b) a section or (c) a paragraph (or a similar type of structure in terms of the granularity of the ideas expressed like figure, table, diagram, sentence, footnote, listing, reference, etc.). The level of granularity that the review comments address in the text of the article can give indications, for example, of the pieces of text (whether paragraph-level ones or sections) that were modified the most or that were required to be modified the most, the details added or deleted, and the possibility to track if review comments were addressed or not as intended by the reviewer. This also aligns with the representation of the Linkflows model as inter-connected distributed nodes in a network. Basically, a paragraph-like structure and a review comment that addresses it are nodes connected together in a network of scientific contributions, while an article section and the entire article is composed of multiple inter-connected nodes.

In summary, therefore, our dataset covers 35 pairs of articles and reviews, where the reviews are split into smaller snippets that allow us now to further simulate the application of our model by manual annotation.

4.2 Ground-Truth and Manual Annotation

In order to find out how well the Linkflows model would have worked had it been applied to the papers of our dataset in the first place, we approached the specific peer reviewers who wrote the reviews in our dataset (which is why it was important to exclude anonymous reviewers) and asked them in a questionnaire to categorize one of their existing review comments according to our new model. We can then use their responses as ground-truth data. Specifically, they were asked whether a review comment was about syntax, style or content, whether they raised a positive or negative point, whether an action was mandatory or suggested, what the impact level was for the overall quality of the article, and whether the author eventually addressed the point.

With our envisaged approach, the various actors (most importantly the reviewers) would in the future directly contribute to the network of semantically represented snippets. This kind of

[8] http://www.semantic-web-journal.net/issues#2018

[9] https://content-iospress-com.vu-nl.idm.oclc.org/journals/data-science/1/1-2

[10] https://peerj.com/computer-science/

[11] https://openreview.net/group?id=swsa.semanticweb.org/ISWC/2018/DeSemWeb&mode=edit

[12] https://openreview.net/group?id=ISMIR.net

annotation activity is therefore only needed for our evaluation methodology but would not be needed in the future when our approach is applied.

Considering the limited time people have for filling in questionnaires, we selected just one review comment for each peer reviewer. For that, we again applied SHA-256 hashing on the review text and chose the one with the smallest hash value. For simplicity, we only considered paragraph-level comments here. To optimize the questionnaire that asks reviewers to provide the classifications according to our model, we conducted a small pilot study in the Semantic Web groups at the Vrije Universiteit Amsterdam. We received 12 answers and incorporated the feedback in the further design of the questionnaire before we sent it to the peer reviewers. Out of 35 contacted peer reviewers (corresponding to the 35 selected articles in our dataset), eleven replied. This therefore resulted in a ground-truth dataset of eleven review comments.

In order to further assess the applicability of the model, the three authors of this paper independently annotated the review comments of the ground-truth datasets as well. This will allow us to compare our answers to the ones of the ground truth, but also to quantify the agreement among us model experts by using Fleiss' Kappa, which can be seen as an indication of the clarity and quality of the model.

For comparison, we also calculate a random baseline of equally distributed and fully uncorrelated responses that correspond to the outcome of random classification. This will serve as an upper baseline for the disagreement with the ground truth.

4.3 Peer Questionnaire

Next, we wanted to find out how important it is to let reviewers themselves semantically express their reviews (with proper tool support), instead of extracting this representation afterwards from classical plain-text based reviews. Specifically, we wanted to find out to what extent the interpretation of a review comment by a peer researcher would differ from the meaning that the reviewer had in mind. We hypothesize that there is a substantial level of disagreement, which would demonstrate that we cannot reliably reconstruct the reviewer's intention in detail if we don't let him or her express these explicitly when authoring the review. This in turn, can then be seen an indication of the benefit and value of our approach.

For this part, we used the same kind of questionnaire that we also used for the ground-truth collection, now covering all eleven review snippets that are covered by the ground-truth data. We first conducted again a small pilot study, to see how long it will take to fill in such a questionnaire. Due to the assumed time constraints of potential respondents, we decided to split the questionnaire randomly in two parts, such that filling in one part of the questionnaire would take somewhere between 10–15 minutes. We then sent the two parts of the questionnaire to various mailing lists in the field of Computer Science.

4.4 Automated Sentiment Analysis

Finally, we wanted to compare the results we get from human experts with what we can achieve with fully automated methods. Specifically, we applied off-the-shelf sentiment analysis tools to extract the positivity/negativity from the text of the review comments. We expect that these automated sentiment analysis tools will perform worse than the ground truth (i.e. the classifications by the original reviewers) because such comments can have various targets of positivity and negativity (e.g. negativity towards mentioned related work vs. negativity towards the article being reviewed) and because even very critical points are sometimes phrased very politely, thus expressing a negative point with positive language. We hypothesize that these subtle distinctions are not correctly identified by automated sentiment analysis tools.

We used a benchmark sentiment analysis platform called iFeel [13] [32] to automatically detect whether the eleven review comments in our ground-truth dataset express negative, neutral/balanced, or positive points. We ran all the 18 lexicon-based sentiment analysis methods available of the benchmark platform and collected the results. Finally, we calculated the accuracy of these automated sentiment analysis methods versus the ground-truth data.

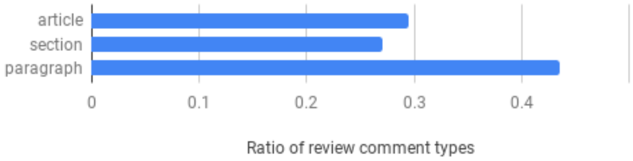


Figure 3: The part of the article that a review comment targets.

5 RESULTS

We can now have a look at the results we obtained by conducting the studies outlined above.

5.1 Descriptive Analysis

Figure 3 shows the distribution of the total of 421 review comments in our dataset with respect to the type of article structure they target. 29% of them are about the article as a whole, 27% are at the section level, and the remaining 44% target a paragraph or an even smaller part of the article. This indicates that the review comments indeed often work at a granular level and that the application of a finer-grained model, such as the Linkflows model, therefore indeed seems appropriate and valuable.

Figure 4 shows the results of the manual annotation by us model experts on the 450 review comments that correspond to the articles included in the ground-truth subset (articles for which we have one review snippet annotated by one of its peer reviewers). We notice that most of the review comments were about the content of the text, while comments about style and syntax were less common. In terms of the positivity/negativity, we see that — unsurprisingly — most of the review comments were rated as negative and that there is a relatively balanced distribution of the neutral/balanced and positive. The impact of the review comment for the overall quality of the article is rarely assigned the extreme values of 1 or 5, but the remaining three values of 2–4 are all common (2 and 3 more so than 4). The action that needs to be taken according to

[13] http://blackbird.dcc.ufmg.br:1210/

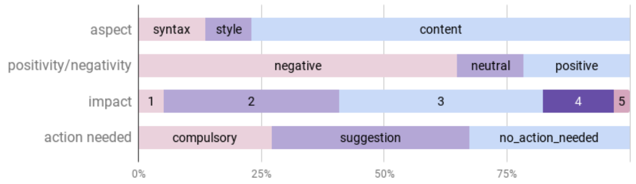


Figure 4: The results of the model expert annotations.

the reviewer, finally, shows a balanced distribution of the three categories (compulsory, suggestion, and no-action-needed).

For all of the 450 review snippets that we rated, the average degree of agreement based on Fleiss' kappa [17] had a value of 0.42 which indicates moderate agreement between raters for all dimensions of the Linkflows model [25]. We notice, however, a considerable variation across dimensions: the aspect and positivity/negativity dimensions had the highest agreement with values of 0.68 and 0.62 respectively (substantial agreement), the action needed had a value of 0.37 (fair agreement), while the impact dimension gave a value of just 0.03 (slight agreement). This low value for the impact dimension is not so surprising given that it consists of just a numerical scale without precise definitions of the individual categories. Moreover, it has a larger number of categories which is known to lead to worse kappa values. But, altogether, the substantial inter-annotator agreement shows that our model is well applicable and sufficiently precise.

For the peer questionnaire study, we gathered 79 responses in total: 43 responses for the first part of the questionnaire containing 5 review comments (215 answers in total) and 36 answers for the second part with 6 review comments (216 answers in total). The participants had to answer a few questions about their background, which revealed that 58.2% of the respondents have a university degree and that most of them are working in academia (89.9%). The majority (75.9%) has advanced knowledge of Computer Science, while only 5% consider themselves beginners in this field. This confirms that they can indeed be seen as peers of the reviewers and authors of our dataset of Computer Science papers.

When answering the questions, these peers could also choose the options "More context would be needed; it is not possible to answer" or "The review comment is confusing; it is not possible to answer", when they didn't feel confident to give an answer. We can now take this as an indication of how well the model worked out from their point of view. Table 1 shows these results. We see some variation across the different dimensions, but at a very low level. The two options combined were chosen in less than 7% of the cases for any of the dimension, and overall in just a bit over 4% of the cases was one of these two options chosen. This indicates that the dimensions of the Linkflows model are overall very well understood and easy to apply.

5.2 Accuracy of Sentiment Analysis Methods

Before we move on to compare the results from the different sources, we first analyze the performance of the different sentiment analysis tools here, in order to be able to chose the best ones for inclusion in the comparative analysis described below.

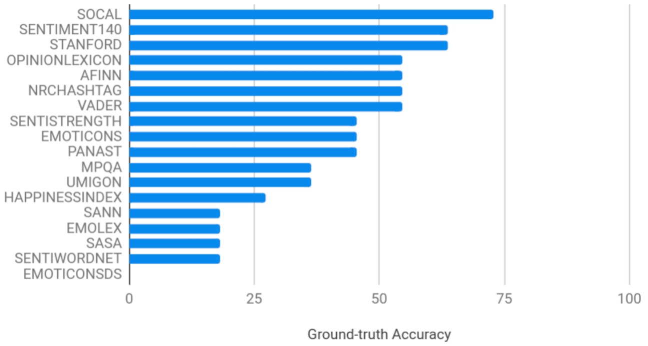


Figure 5: Accuracy of 18 Sentiment Analysis methods vs. ground-truth data.

The results we get from the iFeel platform are already normalized to the values of negative, neutral, and positive, which directly map to the positivity/negativity dimension of our model. We then define accuracy in this context as the ratio of correctly classified instances when compared to the ground truth we collected from the reviewers.

In Figure 5 we summarized the results of the 18 sentiment analysis methods used to rate the eleven review comments for which we have ground-truth data. We see a wide variation of performance with a maximum accuracy of 72.8% for the SOCAL method [41]. As expected, most of these methods perform quite poorly. In general, we can observe that the sentiment analysis methods that have more complex rules perform best, even if their lexicon size is not large. Apart from SOCAL, we also select the tools ranked second and third for our comparative analysis below: Sentiment140 [19] and Stanford Recursive Deep Model [38] (both 63.6% accuracy).

5.3 Comparative Analysis

We can now have a closer look into the differences between the answers we got from the different sources, to better understand the nature and extent of the observed disagreement. We quantify this disagreement by applying a variation of the Mean Squared Error metric. To compare the responses of two groups, we calculate normalized (0–1) within-group averages of their responses and then take the square root of the mean squared differences between the groups. For the nominal dimensions, we calculate the squared differences from the ratio for each category separately.

These disagreement scores between reviewers, model experts, and peers for all dimensions of the model are shown in Table 3 (for the positivity/negativity dimension, which includes the sentiment analysis tools) and Table 2 (for all other dimensions).

Looking at Tables 2 and 3 we see that the agreements between the reviewers and the groups of model experts and peers range from 0.12 to 0.38, well above perfect agreement (0) but also well below the random baseline (0.5 for the ordinal dimension *impact*, 0.58 for the ordinal dimension *positivity/negativity*, and 0.67 for the nominal dimensions). It is notable that the model experts and the peers always agree with each other more than they agree with the ground truth in the form of the original reviewer. This seems to indicate that they misinterpret the review comments in a relatively

Table 1: Percentages of questions where it was not possible to answer (Peers experiment).

	aspect	positivity/negativity	action needed	impact	action taken	Overall
“more context needed”	0.75%	1.50%	2.64%	4.90%	2.64%	2.49%
“confusing”	1.89%	1.70%	1.32%	2.07%	0.94%	1.58%
Total	2.64%	3.20%	3.96%	6.97%	2.58%	4.07%

Table 2: Disagreement scores for four of the Linkflows model dimensions.

	aspect			action needed			impact			action taken		
	R	M	P	R	M	P	R	M	P	R	M	P
Reviewer (R)	0			0			0			0		
Model experts (M)	0.32	0		0.26	0		0.21	0		0.38	0	
Peers (P)	0.29	0.12	0	0.29	0.17	0	0.31	0.14	0	0.34	0.21	0
Random baseline	0.67			0.67			0.50			0.67		

Table 3: Disagreement scores for the positivity/negativity dimension.

	R	M	P	SA1	SA2	SA3
Reviewer (R)	0					
Model experts (M)	0.32	0				
Peers (P)	0.16	0.24	0			
SOCAL (SA1)	0.26	0.41	0.24	0		
SENTIMENT140 (SA2)	0.30	0.16	0.22	0.34	0	
STANFORD (SA3)	0.30	0.16	0.22	0.34	0	0
Random baseline	0.58					

small but consistent manner. The highest disagreement scores for all pairs of groups come from the *action taken* dimension, which seems in general slightly more difficult than the others.

[modify text to reflect the results in Table 3] With respect to the positivity/negativity dimension (Table 3), we see that the best automated sentiment analysis tool surprisingly performed a bit better than the model experts (0.26 versus 0.28 disagreement score, compared to the ground truth). A further surprising outcome is that the peers performed much better than the model experts and the sentiment analysis tools. We will further investigate this effect below. Moreover we see that the second and third best sentiment analysis tools provide identical results (and therefore zero disagreement), but that they have considerable disagreement with the best tool (SOCAL), which might hint at complementary information given by these tools and therefore the potential for an ensemble method.

[modify text to reflect the results in Table 4] Table 4 focuses on the comparison of the two main groups of model experts and peers. We see that for three of the five dimensions, the peers had a higher agreement (i.e. lower disagreement) with the ground truth than the model experts. Testing these observed differences for statistical significance with a two-tailed Wilcoxon signed-rank test test, only the difference in favor of the peers for the positivity/negativity dimension turned out to be significant (p-value of 0.02).

This somewhat counter-intuitive result that peers perform better than model experts when compared to the ground truth could be explained by the fact that the model experts shared more detailed information and acquired background knowledge through in-depth discussions with each other, and therefore they had information neither the reviewer nor the peers had. This might have made them settle on similar choices that are different from the model experts’. Because the group of peers consists of more than ten times as many individuals as compared to the model expert group, an alternative explanation is that this effect is due to a wisdom of the crowd effect, where averaging over a larger number of individual responses gets us closer to the truth.

In order to test this wisdom of the crowd hypothesis, we divided the answers from the peers into twelve groups of three peers each (thereby matching the size of the model expert group) and calculated the disagreement scores for each of these smaller groups. The bottom part of Table 4 shows the average results. We see that the average disagreement of the smaller peer groups is always larger than the disagreement of the large peer group, confirming that there seems to have been a wisdom of the crowd effect at play. Except for positivity/negativity, where the difference was largest to start with, the average disagreement of the smaller peer groups is larger than for the model experts. Therefore, peers seem to perform slightly worse than model experts, and a substantial wisdom of the crowd effect makes larger groups of individuals to perform better.

6 DISCUSSION AND CONCLUSION

The results of this study suggest that the Linkflows model is able to express at a finer-grained level the properties of review comments and their relation to the article text. We propose to let reviewers create semantically represented review comments from the start. Our evaluation showed that a substantial level of disagreement arises if other actors like peer researchers or model experts try to reconstruct the intended nature of these review comments afterwards, thereby underlying the importance of capturing this information right at the source. Through a wisdom of the crowd effect, larger groups of peers can achieve lower disagreement with the ground-truth provided by the original reviewers, but the disagreement is still substantial and larger groups of annotators require more collective effort. Existing automated methods like sentiment analysis tools can reach the same level of agreement like model experts, but a substantial level of disagreement remains for both of them. In summary, this confirms the value of our approach to precisely capture the network and type of review comments directly from the reviewers.

Table 4: Model experts versus peers.

	aspect	positivity/negativity	action needed	impact	action taken
Model experts	0.32	0.32	0.26	0.21	0.38
Peers	0.29	0.16	0.29	0.31	0.34
Difference	0.03	0.16	−0.03	−0.10	0.04
p-value (Wilcoxon signed-rank test)	0.21	0.28	0.42	0.09	0.66
Average of groups of 3 peers	0.34	0.22	0.31	0.34	0.44
Standard deviation of groups of 3 peers	0.026	0.010	0.119	0.055	0.016

While we have provided a simple mock-up of an interface for reviewers, an actual prototype still has to be implemented and evaluated as future work. Specifically, we plan to experiment with applying the Dokieli environment [6] and Linked Data Notifications [5]. To publish reviews (and also the articles themselves) as a network of semantically annotated snippets in a decentralized way, we will furthermore investigate the application of the nanopublication technology and infrastructure [15] for reliable semantic publishing.

REFERENCES

[1] 2011. Dissecting our impact factor. *Nature Materials* (2011). https://www.nature.com/articles/nmat3114

[2] Molood Barati, Quan Bai, and Qing Liu. 2016. SWARM: An Approach for Mining Semantic Association Rules from Semantic Web Data. In *Trends in Artificial Intelligence, Lecture Notes in Computer Science (PRICAI'16)*, Vol. 9810. 30–43. https://doi.org/10.1007/978-3-319-42911-3_3

[3] Wim G.G. Benda and Tim C. E. Engels. 2011. The predictive validity of peer review: A selective review of the judgmental forecasting qualities of peers, and implications for innovation in science.

[4] John Bohannon. 2013. Who's Afraid of Peer Review? *Science* 342, 6154 (2013), 60–65. https://doi.org/10.1126/science.342.6154.60

[5] Sarven Capadisli, Amy Guy, Christoph Lange, Sören Auer, Andrei Vlad Sambra, and Tim Berners-Lee. 2017. Linked Data Notifications: A Resource-Centric Communication Protocol. In *ESWC*.

[6] Sarven Capadisli, Amy Guy, Ruben Verborgh, Christoph Lange, Sören Auer, and Tim Berners-Lee. 2017. Decentralised Authoring, Annotations and Notifications for a Read-Write Web with dokieli. In *ICWE*.

[7] Tom Heath Christian Bizer and Tim Berners-Lee. 2009. Linked Data - the story so far. *International Journal on Semantic Web and Information* 5, 3 (2009), 1–22. https://doi.org/10.4018/jswis.2009081901

[8] Sergio Copiello. 2019. On the skewness of journal selfâĂŘcitations and publisher selfâĂŘcitations: Cues for discussion from a case study. *Learned Publishing* 32, 3 (2019), 249–258.

[9] Jeremy Debattista, Christoph Lange, and Sören Auer. 2014. daQ, an Ontology for Dataset Quality Information. In *Workshop on Linked Data on the Web*.

[10] Eleftherios P Diamandis. 2017. The Current Peer Review System is Unsustainable-Awaken the Paid Reviewer Force! *Clinical biochemistry* 50, 9 (2017). https://doi.org/10.1016/j.clinbiochem.2017.02.019

[11] Amrapali Zaveri et al. 2015. Quality assessment for Linked Data: A Survey. *Semantic Web* 7, 1 (2015). https://doi.org/10.3233/SW-150175

[12] Brian S. Alper et al. 2004. How much effort is needed to keep up with the literature relevant for primary care? *Journal of the Medical Library Association* 92, 2 (2004), 429–437.

[13] Daniel Leung et al. 2014. How to review journal manuscripts: A lesson learnt from the world's excellent reviewers. *Tourism Management Perspectives* 10 (2014), 46–56. https://doi.org/10.1016/j.tmp.2014.01.003

[14] Tobias Kuhn et al. 2013. Broadening the Scope of Nanopublications *(ESWC'13)*, Vol. 7882. ESWC: European Semantic Web Conference, Springer, 487–501. https://doi.org/10.1007/978-3-642-38288-8_33

[15] Tobias Kuhn et al. 2016. Decentralized provenance-aware publishing with nanopublications. *PeerJ Computer Science* 2 (2016), e78. https://doi.org/10.7717/peerj-cs.78

[16] Imen Filali, Francesco Bongiovanni, Fabrice Huet, and Francoise Baude. 2011. A Survey of Structured P2P Systems for RDF Data Storage and Retrieval. *LNCS* 6790 (2011). https://doi.org/10.1007/978-3-642-23074-5_2

[17] Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin* 76, 5 (1971), 378–382. https://doi.org/10.1037/h0031619

[18] Eugene Garfield. 2006. The History and Meaning of the Journal Impact Factor. *JAMA* 295, 1 (2006), 90–93. https://doi.org/10.1001/jama.295.1.90

[19] Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. (2009). Technical Report.

[20] Paul Groth, Andrew Gibson, and Jan Velterop. 2010. The anatomy of a nanopublication. *Information Services and Use* 30, 1-2 (2010). https://doi.org/10.3233/ISU-2010-0613

[21] Sumaiya Kabir, Shamim Ripon, Mamunur Rahman, and Tanjim Rahman. 2013. Knowledge-based Data Mining Using Semantic Web. In *ICACC*.

[22] Tobias Kiesslich, Silke B. Weineck, and Dorothea Koelblinger. 2016. Reasons for Journal Impact Factor Changes: Influence of Changing Source Items. *PLoS ONE* 11, 4 (April 2016). https://doi.org/doi:10.1371/journal.pone.0154199

[23] Tobias Kuhn and Michel Dumontier. 2017. Genuine semantic publishing. *Data Science* 1-2 (2017), 139–154. https://doi.org/10.3233/DS-170010

[24] Esther Landhuis. 2016. Scientific literature: Information overload. *Nature* 535 (2016), 457–458. https://doi.org/10.1038/nj7612-457a

[25] Richard J. Landis and Gary G. Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics* 33, 1 (1977), 159–174. https://doi.org/10.2307/2529310

[26] Vincent Lariviere and Cassidy Sugimoto. 2018. The Journal Impact Factor: A brief history, critique, and discussion of adverse effects. (2018).

[27] Faina Linkov, Mita Lovalekar, and Ronald LaPorte. 2006. Scientific Journals are ''faith based": is there science behind peer review? *Journal of the Royal Society of Medicine* 99 (2006), 596–598. Issue 12. https://doi.org/10.1258/jrsm.99.12.596

[28] Stephen Lock. 1985. *A Difficult Balance: Editorial Peer Review in Medicine.* The Nuffield Provincial Hospitals Trust, London.

[29] Chris Markman and Constantine Zavras. 2014. BitTorrent and Libraries: Cooperative Data Publishing, Management and Discovery. *D-Lib Magazine* 20, 3-4 (2014). https://doi.org/10.1045/march2014-markman

[30] David Mazières and M.Frans Kaashoek. 1998. Escaping the evils of centralized control with self-certifying pathnames *(ACM SIGOPS'98)*. https://doi.org/10.1145/319195.319213

[31] Barend Mons. 2005. Which gene did you mean? *BMC Bioinformatics* 6, 1 (2005). https://doi.org/10.1186/1471-2105-6-142

[32] Filipe N. et al. Ribeiro. 2016. SentiBench - a benchmark comparison of state-of-the-practice sentiment analysis methods. *EPJ Data Science* 5, 1 (07 Jul 2016), 23. https://doi.org/10.1140/epjds/s13688-016-0085-1

[33] Afshin Sadeghi, Sarven Capadisli, Johannes Wilm, Christoph Lange, and Philipp Mayr. 2019. Opening and Reusing Transparent Peer Reviews with Automatic Article Annotation. *Publications* (2019). https://doi.org/10.3390/publications7010013

[34] Jodi Schneider, Paolo Ciccarese, Tim Clark, and Richard David Boyce. 2014. Using the Micropublications Ontology and the Open Annotation Data Model to Represent Evidence within a Drug-Drug Interaction Knowledge Base. In *LISC@ISWC*.

[35] Nigel Shadbolt, Tim Berners-Lee, and Wendy Hall. 2006. The Semantic Web Revisited. *IEEE Intelligent Systems* (2006). https://doi.org/10.1109/MIS.2006.62

[36] Richard Smith. 1988. Problems With Peer Review And Alternatives. *British Medical Journal (Clinical Research Edition)* 296, 6624 (1988), 774–777.

[37] Richard Smith. 2010. Classical peer review: an empty gun. *Breast cancer research* 12 (2010). https://doi.org/10.1186/bcr2742

[38] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Y. Ng, and C. Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. (2013).

[39] David Soergel, Adam Saunders, and Andrew McCallum. 2013. Open Scholarship and Peer Review: a Time for Experimentation.

[40] Gina S. Sucato and Cynthia Holland-Hall. 2018. Reviewing Manuscripts: A Systematic Approach. *Journal of Pediatric and Adolescent Gynecology* 31, 5 (2018).

[41] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational Linguistics* 37 (2011). https://doi.org/10.1162/COLI_a_00049

[42] Paolo N Ciccarese Tim Clark and Carole A Goble. 2014. Micropublications: a semantic model for claims, evidence, arguments and annotations in biomedical communications. *Journal of Biomedical Semantics* 5, 28 (2014). https://doi.org/10.1186/2041-1480-5-28

[43] Mark Ware and Michael Mabe. 2015. The STM Report: An overview of scientific and scholarly journal publishing.