Check for updates

Expert Systems

WILEY

Expert Systems

ORIGINAL ARTICLE OPEN ACCESS

The Quality of a Scientific Manuscript Given by a Peer-Reviewed. Report in Three Dimensions: Accessibility, Contribution and Experimentation (AccConExp)

J. J. Montero-Parodi[1] | Rosa Rodriguez-Sánchez[2] | J. A. García[2] | J. Fdez-Valdivia[2]

[1]CITIC-UGR, Universidad de Granada, Granada, Spain | [2]Departamento de Ciencias de la Computación e I. A., CITIC-UGR, Universidad de Granada, Granada, Spain

Correspondence: Rosa Rodriguez-Sánchez (rosa@decsai.ugr.es)

Received: 7 June 2024 | **Revised:** 30 July 2025 | **Accepted:** 29 August 2025

Funding: Funding for open access charge: Universidad de Granada / CBUA.

Keywords: aspect extraction | construct extraction | deep learning | partial least squares-structural equation modelling (PLS-SEM) | peer review | sentiment analysis

ABSTRACT

In this paper, we present a new model (AccConExp) to help the editor evaluate the reviewer's report in the peer-review process. The model provides information about accessibility, contribution and experimentation and analyses the sentiment of these characteristics. Accessibility pertains to the clarity and coherence of the manuscript; Contribution assesses whether the work is original or well justified; and Experimentation reflects the presence of significant comparisons or substance within the paper. For example, with this information, a journal editor can establish whether the paper meets the journal's standards given the polarity of accessibility, contribution or experimentation. The AccConExp model provides a strong and flexible framework for the analysis of reports that emphasise accessibility, contribution and experimentation. Its computational efficiency and scalability with emerging categories render it an essential resource for journal editors and various stakeholders within the academic and research communities. Furthermore, the AccConExp model introduces a novel method for improving the peer-review process by offering a more organised and insightful analysis of reviewers' reports, ultimately resulting in more consistent and high-quality assessments of scientific research. For this, the AccConExp model integrates a theoretical model based on partial least squares-structural equation modelling (PLS-SEM) to acquire new knowledge and a multi-task deep machine learning to explore the knowledge learning with the model PLS-SEM. The PLS-SEM part of the AccConExp model obtains a causal prediction from a set of aspect categories assigned to the reviewer's report to build new knowledge of the report based on accessibility, contribution and experimentation. The causal-exploratory capabilities of the multi-task deep learning model allow the labelling of new report's sentences based on accessibility, contribution, experimentation constructs and sentiment. Once we discover a sentence's construct, a second deep learning machine allows us to obtain its aspect category (clarity, soundness, originality, motivation, substance and meaningful comparison). The AccConExp model has been tested using reviewer reports from ICLR and NeurIPS papers (conferences with high impact in machine learning). The AccConExp model is compared with a multi-task architecture that assigns aspect categories to the report's sentences. The results obtained with the AccConExp model are competitive and allow us to give new information to the reviewer's reports without the effort to generate a new dataset labelled with these new constructs. Also, the AccConExp model's computational efficiency and capacity to adapt

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2025 The Author(s). *Expert Systems* published by John Wiley & Sons Ltd.

to new categories render it an invaluable resource for journal editors and various stakeholders within the academic and research community. The methodology used in this paper can be extended to other research fields to define its constructs, even if the aspects considered in the review's reports area differ from those used in this proposal. We release our codes for more study.

1 | Introduction

Accessibility, new contributions to knowledge and maintaining high-quality experimentation are essential for a scientific paper's recognition of quality.

When a journal's editor receives a scientific manuscript to be evaluated, the editor typically asks for help from experts in the field (reviewers). The reviewer's expertise is of great value in the peer-review process. The reviewers' report usually includes comments on different aspects. Given the journal's profile, the editor can be inclined to high accessibility scores, such as generalist journals, or give more importance to contributions and experimentation, as in technical journals (García et al. 2018).

For example, social and human science readers may prioritise articles with higher accessibility, as the difference in accessibility between less and more complex manuscripts is more salient than the difference in article contribution. In specialised disciplines, such as physics, mathematics and engineering, scholars are more attentive to article contribution and less sensitive to differences in accessibility (García et al. 2018).

The reviewers generate a report with different sentences describing aspects such as clarity, motivation, originality, soundness, meaningful comparison and substance. These sentences can also have a positive or negative polarity (Yuan et al. 2021). For example, the sentence 'This paper incorporates new ideas that have not been analysed before' characterises the research manuscript with an originality-positive. The aspect categories and sentiment analysis of the reviewer's report can help the editor make a final decision (accept or reject).

In this way, integrating accessibility features into scientific manuscripts is pivotal for maximising the dissemination of research findings. For example, some of the key accessibility features to consider are incorporating plain language explanations for complex terms or concepts, enhancing accessibility for readers with diverse backgrounds and levels of expertise (Bralić et al. 2024). Additionally, providing accessible references, such as links to freely available articles or resources, ensures broader accessibility. Cognitive accessibility should also be prioritised by employing concise sentences, avoiding overly complex language and providing clear explanations throughout the manuscript (Alarcón García 2022). Among these prerequisites, clarity and soundness are relevant to manifest accessibility in a manuscript.

A manuscript incorporating new contributions to the knowledge refers to the novelty and uniqueness of the study's findings, methodology, or perspective within the existing body of literature. By prioritising originality in research papers, scholars contribute to advancing knowledge, stimulate intellectual discourse and inspire future investigations in their respective fields. New contributions to knowledge have a primary ingredient, which is originality. A definition of originality is given by Shibayama and Wang (2020) as 'the degree to which a scientific discovery provides subsequent studies with unique knowledge that is not available from previous studies'. Another characteristic associated with originality is motivation. Thus, the relationship between originality and motivation represents how creators engage in their work (Forgeard and Mecklenburg 2013). From this point of view, an author of a scientific manuscript must portray this motivation in the manuscript so that a reader feels the need for the new approach proposed by the author.

High-quality experimentation must be conducted rigorously and well-designed, adhering to established methodologies and best practices in the field to ensure the reliability and validity of the findings (Stach et al. 2021). In this sense, substance and meaningful comparison can be essential to describe good experimentation.

Chakraborty et al. (2020) discovered that sentiments linked to aspects such as clarity, empirical/theoretical validity and the impact of ideas significantly influence the intended recommendation in a review. Therefore, the manuscripts' authors must take care of these aspects.

Helping the peer-review process is a fundamental need for the increased volume of scientific manuscript submissions to journals daily. Each reviewer receives more invitations to review manuscripts and a tight deadline to carry them out. For example, Johnson et al. (2018) observed that more than 3 million articles are published annually. Among them, 33,100 are in English-language peer-reviewed journals, and 9400 are in non-English-language journals. According to Huisman and Smits (2017), the average time to review an article is 17 weeks. In addition to this increase in the volume of submissions, there are other problems in the peer-review process, such as bias (Tomkins et al. 2017; Fernández Pinto 2023), inconsistencies (Cortes and Lawrence 2021) and subjectivity (Brezis and Birukou 2020).

In this paper, we give and propose automatically tagging the sentences' review report using three constructs: accessibility, contribution and experimentation. With this information, a journal's editor can evaluate whether the manuscript matches the accessibility, contribution and experimentation profile required by the journal.

Before building this automatic process, we need a dataset with accessibility, contribution and experimentation annotations. Data annotation is a critical process that requires a great deal of effort and time. Usually, data annotation of a report needs human experts to label the report's sentences.

Figure 1 shows a scheme of the proposed method, called the AccConExp model. In stage one, the sentences of a report are annotated. For this, we will use the aspect categories tagger of reports presented in Yuan et al. (2021). Aspect categories

14680394, 2025, 10, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/exsy.70131 by University Of Granada, Wiley Online Library on [27/10/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

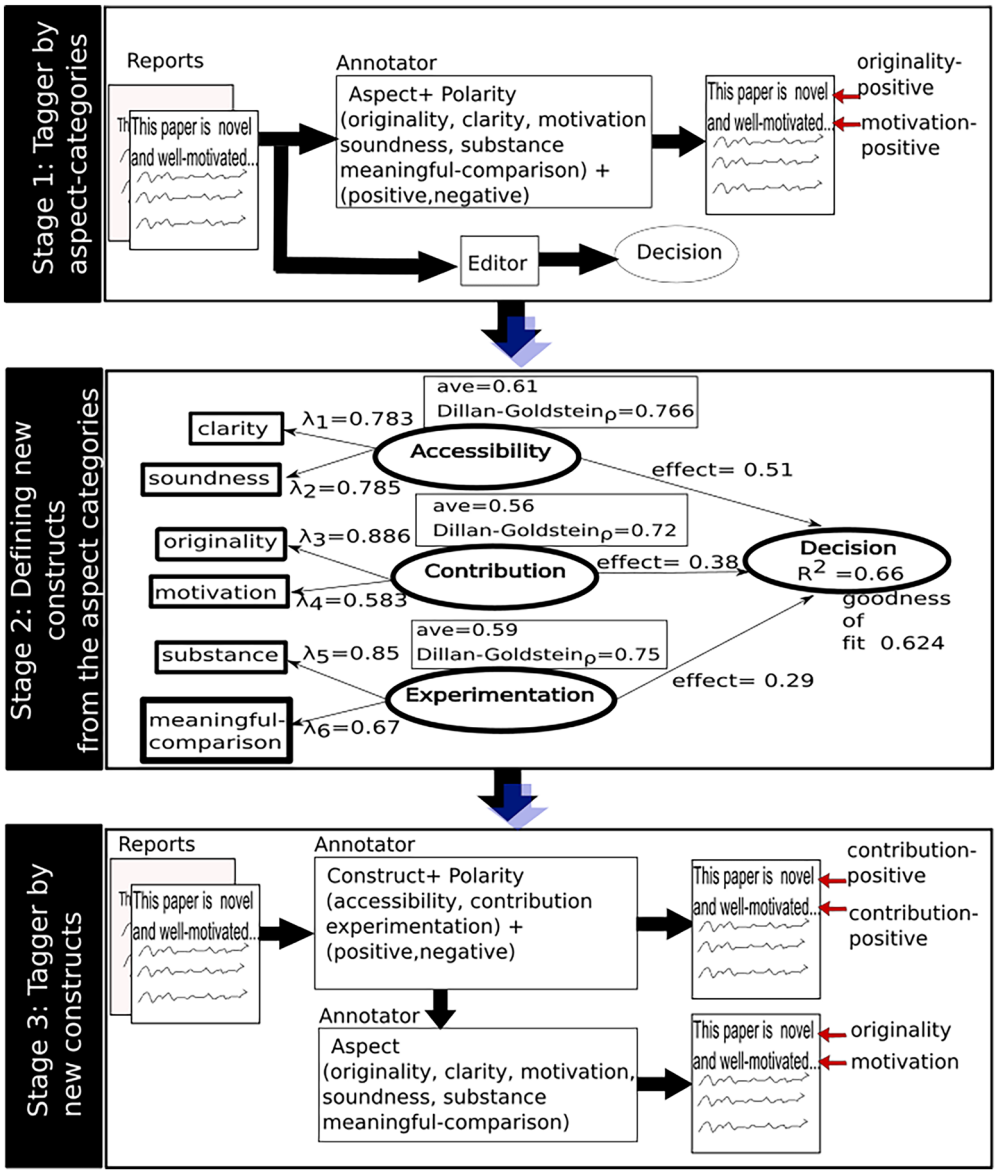

FIGURE 1 | Scheme of the proposed method (AccConExp).

are originality, soundness, clarity, motivation, meaningful comparison and substance. In the next stage, using the statistical technique PLS-SEM (Hair et al. 2019), we will model the relationship between the observed variables (originality, soundness, clarity, motivation, meaningful comparison and substance) and the constructs (accessibility, contribution, experimentation and a score based on the observed variables, we have called decision in Stage 2 of Figure 1). Once the model PLS-SEM has been validated, we relabel the dataset using accessibility, contribution and experimentation constructors. Finally, in the first place, we will train a deep learning machine to tag the review's report by accessibility, contribution, experimentation and the corresponding sentiment. In the second place, we will train three deep learning machines specialised in the aspect-category associated with the construct. Thus, we will build a machine specialised in distinguishing clarity and soundness in sentences labelled with the accessibility construct, a machine to label originality and motivation with sentences labelled with the contribution construct and a machine specialised in labelling substance and meaningful comparison for sentences labelled with the experimentation construct.

This approach saves time and effort in creating a deep learning machine to tag compared to the normal process of building a new dataset with annotations based on accessibility, contribution and experimentation.

In summary, looking at the shortcomings of peer-review processes, one of the big problems is the time spent in the review process. In this line, the AccConExp model could support the journal editor in the following situations:

1. The review processes are temporarily extended due to the number of articles submitted to the journal. Once the reviewers deliver the review reports to the editor, the AccConExp model can categorise the reports into accessibility, contribution and experimentation. The editor can

14680394, 2025, 10, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/exsy.70131 by University Of Granada, Wiley Online Library on [27/10/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

assess the manuscript in these dimensions and, depending on the journal's area and its profile, see the values of accessibility, contribution, or experimentation along with their polarity and check if they match the journal standards.

2. The editor may have doubts about whether the paper is accessible due to clarity or soundness. Therefore, the editor can now obtain this information using the AccConExp model. The equivalent would be for contribution to distinguish whether originality and motivation appear or for experimentation to distinguish between meaningful comparison and substance.
3. The editor finds it challenging to make a desk decision. The AccConExp model, together with an automatic report generator, could assess whether the manuscript meets the journal's standards measured in accessibility, contribution and experimentation and help the editor perform a desk rejection or pass the manuscript to a peer-review phase. This would allow faster responses to the authors.

This approximation can be applied in other contexts, such as medicine, to diagnose disease from the symptoms described by the patient. For example, in this case, the symptoms of different types of hepatitis (A, B, C and D) are similar, so a new construct can be hepatitis. A bigger analysis of the temporal sequence of symptoms can distinguish the kind of hepatitis. In the botanical world, different species are catalogued with aspects based on their correlation in characteristics, and a deeper classification determines whether they are unique species. Thus, plant species can be classified according to their fruits, leaves, roots and stems. However, other forms of classification could be interesting to discover from the data previously collected, giving rise to new constructs, such as plants that coexist in harmony, plants that require the exact care needs and so on. These concepts can establish new classifications that show intrinsic knowledge existing in the dataset.

This approach can also define new constructs to join data from different datasets. For example, we can use the global economy indicators dataset with a weather dataset to discover new construct relationships between economic indicators and climatological aspects.

In this sense, the approximation given in this paper aims to make a recycling process of the datasets to give new knowledge or concepts to help a user decide.

This paper is structured as follows:

- Section 2 reviews the different works related to assist the peer-review process and how to define new constructs based on the indicators observed to help the peer-review process.
- Section 3 presents the AccConExp model. The first and second stages describe the theoretical model. Once the model is validated, we label the reviewers' reports with the new constructs. In the second stage, we trained a deep-learning machine to label the reports with the construct learned.
- Section 4 shows how to relabel the sentences using the constructs obtained in the theoretical model and also gives the details of implementation.
- Section 5 describes the results obtained by the deep learning machine that labels the report's sentences based on the accessibility, contribution and experimentation constructs, and also assigns aspect categories (clarity, soundness, originality, motivation, substance and meaningful comparison) and sentiment. The results are compared with the multitask deep neural architecture, which labels the sentences with aspect categories and sentiment.
- Finally, we conclude with the main achievements.

2 | Related Works

In the present context, the goal is to explore how theoretical constructs (e.g., accessibility, contribution and experimentation) and their indicators (or their aspects categories) can enhance the peer-review process. According to Weber (2021), a systematic understanding of constructs and indicators can lead to a more robust theoretical model and empirical research. This perspective highlights how a journal's editor can streamline decision-making by focusing on specific values (e.g., accessibility, contribution and experimentation values) associated with reviewer feedback.

Following systematic review guidelines, the related works were evaluated based on their relevance, methodological rigour and contribution to understanding peer-review processes and automated feedback systems. Articles were organised thematically into categories such as:

- Causal-predictive methods: Integration of PLS-SEM and ML to enhance predictive and exploratory modelling.
- Reviewer bias and decision-making: Analytical studies addressing fairness, transparency, behaviour and biases in peer-review processes.
- NLP for peer-review generation: Exploration of automated systems for generating reviews and their implications.
- Deep learning architectures for aspect extraction: Application of advanced neural architectures for detecting review aspects and sentiments, including deep architectures used to generate meta-reviews from the reviewers' reports.
- Structural elements of academic writing: Examination of the influence of manuscript structure on reviewer attention and commentary.

This thematic categorisation provides a structured framework for understanding how diverse methodologies, models and indicators converge to enrich the analysis and automation of peer-review processes. A summary of the related works is shown in Table 1.

The compilation of relevant literature on peer-review enhancements, automated systems and construct-indicator approaches is reflected in Sections 2.1 and 2.2. Finally, we relate the fundamental lines discussed to our approach.

14680394, 2025, 10, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/exsy.70131 by University Of Granada, Wiley Online Library on [27/10/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 1 | Related works and fundamental lines of research.

	Description	Papers
PLS-SEM combined with ML	Causal-predictive methods: Integration of PLS-SEM and ML to enhance predictive and exploratory modelling	Richter and Tudoran (2024), Sanchez-Franco et al. (2019), Leong et al. (2020), Tu et al. (2024)
Automatic Tools assist the peer-review process	Novel metrics for peer-review quality	Ragone et al. (2013), Marcoci et al. (2022)
	Analysis of the behaviour of the peer-review process	Zhou et al. (2024)
	NLP-based solutions to discover potential biases	Gao et al. (2019), Yuan et al. (2021), Verma et al. (2021), Khraisha et al. (2024), Bharti et al. (2024), Ghosal et al. (2022)
	Multi-task deep architectures to facilitate simultaneous detection of aspect categories and sentiment, showing high potential for comprehensive review analysis	Yuan et al. (2021), Kumar et al. (2021), Bharti (2024)
	Structural considerations	Qin and Zhang (2023)
	Analysis of the peer-review process in different rounds	Han et al. (2022)
	Deep architectures to get a meta-review	Kumar et al. (2024)
	NLP for peer-review generation	Khraisha et al. (2024), Chen and Zhang (2025)

2.1 | Partial Least Squares-Structural Equation Modelling (PLS-SEM)

Richter and Tudoran (2024) combined PLS-SEM with selected machine learning (ML) algorithms. The goal of this combination is to leverage the strengths of both approaches: the causal-predictive capabilities of PLS-SEM and the causal-exploratory capabilities of ML algorithms. This integrated approach is intended to enhance the predictive accuracy of research models, deepen the understanding of relationships between variables, assist in discovering new relationships and contribute to theory building. Following this approach, we find different papers that combine PLS-SEM and ML algorithms. Thus, Sanchez-Franco et al. (2019) analysed hotel reviews using ML techniques and PLS-SEM for both exploratory and predictive purposes, revealing that hotels should focus on tangible and intangible features to foster loyalty.

The majority of works that combine PLS-SEM with ML techniques first apply PLS-SEM analysis and then ML algorithms to discover non-linearities or validate the PLS-SEM findings. Thus, Leong et al. (2020) examined determinants of trust in social commerce, focusing on social presence and support. They employed a hybrid SEM-ANN model, which combines the strengths of PLS-SEM and artificial neural networks (ANNs).

Tu et al. (2024) investigated student burnout in blended learning environments, focusing on how social support and self-regulated learning predict its occurrence. Using a combination of PLS-SEM and ML algorithms, the research demonstrates that perceived social support and self-regulated learning significantly reduce burnout.

2.2 | Automatic Tools to Assist the Peer-Review Process

Different models and tools have been developed to assist the peer-review process based on the aspect categories (or indicators) annotated in the reviewer's report.

Analysing aspect categories together with sentiment can reveal different behaviours in the peer-review process. In this line, Zhou et al. (2024) investigated how the sentiment of review aspects correlates with the time taken for peer reviews and examined differences across disciplines and review rounds. Going further into the importance of analysing aspect categories, this process can also help determine whether a report is written in a polite manner. Effective peer-review feedback should not only be objective but also polite and constructive. With this aim, Bharti et al. (2024) presented a multi-task model called 'Multi-Label Critique (MLC)', which utilised ToxicBERT representations and attention mechanisms to evaluate review constructiveness and tone.

In Ragone et al. (2013), the authors developed a theoretical model to assess the quality and efficiency of peer-review processes, introducing new metrics alongside established ones. The model aims to enhance transparency and understanding of peer review by evaluating its reliability, fairness, validity, robustness and the degree of agreement/disagreement among reviewers. Additionally, the model assesses the potential bias in reviewers' decision-making and the ability of peer review to predict the future impact of papers. In this same vein, Marcoci et al. (2022) emphasise the distinct and collective traits that enhance the quality of judgements, as well as the components of elicitation protocols that mitigate bias, foster productive

14680394, 2025, 10, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/exsy.70131 by University Of Granada, Wiley Online Library on [27/10/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

dialogue and facilitate the objective and transparent aggregation of opinions.

Gao et al. (2019) contributed to the understanding of peer review in the NLP[1] community by providing empirical evidence on the effectiveness of the rebuttal phase and highlighting potential biases. It also opened up discussions on how to enhance the peer-review process to ensure a high-quality and fair evaluation of scientific work.

Yuan et al. (2021) indicated that while NLP models can assist in generating initial peer reviews by covering a broad range of paper aspects, there is room for improvement in the constructiveness of the generated text. This research opened up avenues for enhancing the peer-review process through automation, potentially alleviating the burden on subject matter experts and accelerating the dissemination of scientific knowledge. However, it also highlights the need for further refinement of NLP models to ensure that they can provide reviews that are not only comprehensive but also constructive and insightful. The authors in Yuan et al. (2021) built the Review Advisor dataset, which has been used in this work. This dataset is composed of a set of annotated reviews from the papers submitted to ICLR[2] and NeurIPS[3] Proceedings.

Verma et al. (2021) proposed an automated system that uses NLP techniques to extract the implicit aspects of a paper that reviewers comment on in their reviews. The system is designed to identify whether the review addresses important elements such as the paper's motivation, methodological soundness, novelty of its contributions and substance of its content. The authors proposed a deep neural architecture that leverages BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al. 2019) and a neural attention mechanism. BERT is a pre-trained language model that has been shown to be effective in a wide range of NLP tasks. The neural attention mechanism allows the model to focus on different parts of the review text that are most relevant to the aspects being extracted.

Kumar et al. (2021) also experimented with the dataset presented in Yuan et al. (2021), proposing a multi-task deep neural architecture that simultaneously learns the aspect categories and the corresponding sentiment (positive or negative) of the report's sentences. The aspect categories used in these works were Motivation/Impact (MOT), Originality (ORI), Soundness/Correctness (SOU), Substance (SUB), Replicability (REP), Meaningful Comparison (CMP) and Clarity (CLA). Variations of the multi-task deep neural architecture presented in Kumar et al. (2021) are analysed in Bharti (2024). We have used in this work a multi-task deep neural architecture variation that encodes the report's sentence using SciBert (Beltagy et al. 2019), followed by attention layers in order to discover the construct category, aspect category and sentiment. Following Kumar et al. (2021) and Verma et al. (2021), we have used the aspect categories in these papers to define new constructs.

Ghosal et al. (2022) highlighted that peer review is fundamental yet often criticised for its lack of transparency and potential biases, noting that research on this process is limited due to confidentiality constraints. They introduced a unique dataset of 1199 annotated open peer reviews, detailing sections, aspects, functionality, significance and sentiment.

In Kumar et al. (2024), a model was presented to obtain meta-reviews for a manuscript. To achieve this, the model predicted the recommendation and confidence scores for the reviews. This information is used to make a decision, which, in a subsequent stage, generates the meta-reviews using a transformer-based seq2seq architecture.

Structural considerations (Qin and Zhang 2023) and iterative commentary (Han et al. 2022) highlight the importance of manuscript layout and the evolving focus of reviewers, adding another dimension to peer-review research. Thus, Qin and Zhang (2023) investigated the structural composition of academic manuscripts, identifying which sections garner the most attention during the peer-review process. Their findings underscore the importance of manuscript organisation in shaping reviewer scrutiny.

At the same time, identifying which manuscript sections attract the most scrutiny (Qin and Zhang 2023) and mapping how feedback evolves over multiple rounds (Han et al. 2022) are essential to addressing potential biases and improving overall review quality. To this end, Han et al. (2022) investigated the dynamic nature of peer-review commentary across multiple rounds, focusing on how reviewer priorities evolve over time. The study categorises peer-review attributes into clarity, methodological soundness, originality, significance, substance, reproducibility and engagement. These attributes provide a structured optic for understanding the cynosure of reviewer feedback. Key findings highlight that initial review cycles often prioritise methodological soundness and originality, while later cycles focus more on clarity and reproducibility. Furthermore, reviewers may shift their attention to engagement and significance in subsequent rounds as manuscripts are refined. By defining these attributes explicitly, Han et al. (2022) offer a comprehensive framework to understand and enhance the iterative peer-review process, emphasising the importance of adaptability and alignment with field expectations.

Khraisha et al. (2024) and Chen and Zhang (2025) evaluated the performance of large language models (LLMs) in automating systematic reviews, a task traditionally performed by humans. The capabilities and limitations of LLMs in research synthesis are discussed, highlighting their potential to accelerate the review process and the challenges associated with their implementation.

2.3 | Our Approach

By synthesising across these studies, a common thread emerges: the integration of methodological frameworks (e.g., PLS-SEM) with advanced NLP or ML techniques can significantly streamline and enrich the peer-review process.

In the current manuscript, we adopt PLS-SEM to examine the causal-predictive relationships among the constructs of accessibility, contribution and experimentation, while indicators

14680394, 2025, 10, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/exsy.70131 by University Of Granada, Wiley Online Library on [27/10/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

(i.e., aspect categories) serve to label review sentences. A multi-task deep neural architecture is subsequently employed to classify reviewer comments according to construct (accessibility, contribution and experimentation), aspect (e.g., clarity, soundness and originality) and sentiment (positive or negative). By incorporating insights from the literature on bias, automated feedback and structural considerations, our approach aims to bridge the theoretical rigour with advanced computational techniques.

In summary, the reviewed literature supports a multi-faceted framework that combines theoretical modelling (PLS-SEM), ML approaches and structural analysis to enhance peer-review processes. Following established systematic guidelines ensures that the present study's methods are grounded in a rigorously curated and critically assessed body of work, ultimately offering a more transparent, efficient and insightful roadmap for researchers, reviewers and editors alike.

3 | Methodology

The AccConExp model comprises three stages (see Figure 1). In the first stage, we get the reports tagged by aspect categories and sentiment analysis following the method described in Yuan et al. (2021). In the second stage, from the reports annotated with aspect categories, we define a PLS-SEM model with three new constructs: accessibility, contribution and experimentation. In the third stage, we define a new annotator based on deep learning to tag the sentences of the reports using accessibility, contribution and experimentation constructs, aspect categories (clarity, soundness, originality, motivation, substance and meaningful comparison) and the corresponding sentiment.

In the following subsections, we detail every stage.

3.1 | Stage 1: Tagger by Aspect Categories

In this stage, we use the approach of Yuan et al. (2021) to tag the reports' sentences by aspect categories.[4] In Yuan et al. (2021), eight typologies of aspects were presented. The typology of aspects chosen by the authors followed the ACL review guidelines.[5]

The authors proposed Summary, Motivation, Originality, Soundness, Substance, Replicability, Meaningful Comparison and Clarity as aspect categories. They also considered positive and negative polarity for the last seven aspects (not for Summary). The aspects annotation process is composed of three stages. In the first stage, manual annotation is realised. Six experts carry out the review's annotation process. Every review is annotated by two experts. The correlation between the two annotators for every review was obtained to ensure a rational annotation. The authors obtained a subset of reviews annotated by human experts. In the second stage, a deep learning machine was trained to annotate the dataset. In the third stage, the authors added different heuristics to enable problems as interleaving different aspects and inappropriate boundaries.

In our model, we only considered six aspect categories (Motivation, Originality, Soundness, Substance, Meaningful Comparison and Clarity). The *Replicability* aspect category is not considered because the number of times assigned in the dataset is very low. The aspect categories used in this paper are shown in Table 2. Also, every aspect can appear with positive or negative polarity. For each one of the aspects, we show examples of positive and negative polarity in Table 2.

Plus, for every manuscript, we have the final decision (accept or reject).

Finally, every manuscript's report is analysed sentence by sentence. The aim is to obtain a score for every aspect category. Initially, all aspect category scores are put to zero. If a sentence of a report is tagged with a determined aspect category and carries a positive sentiment, a value of one is added to the associated score. Conversely, if the sentence is categorised with a negative sentiment, a value of one is subtracted from the score. Table 3 shows the results of this process for a subset of manuscripts. Table 4 shows the title and URL of these manuscripts. A more significant magnitude of the score indicates a strong sentiment.

3.2 | Stage 2: Defining New Constructors From the Aspect Categories

In Stage 1, we have characterised every manuscript by originality, substance, clarity, motivation, soundness and meaningful comparison. Table 3 shows the score aspect categories for five manuscripts. A value of 0 indicates that this aspect is not considered in the manuscript's reports. By contrast, a high positive value indicates that this aspect, with positive polarity, is frequently observed in the manuscript's reports. Instead, a high-magnitude negative value indicates that this aspect, with negative polarity, is observed a lot of times in the manuscript's reports.

Now, with these features describing every manuscript, we want to translate this description to a new description based on accessibility, contribution and experimentation. For this, we first questioned what aspect categories relate to accessibility, which ones relate to contribution knowledge, and which ones relate to experimentation. To answer these questions, we formulated three hypotheses expressing the relationship between aspect categories and the new constructs or latent variables (accessibility, contribution and experimentation).

The first hypothesis related clarity and soundness to accessibility.

Hypothesis 1. *Elements describing accessibility in a manuscript. If the manuscript receives good (or bad) feedback on clarity or soundness, it will be easy (or hard) to understand for a reader with a minimum level of knowledge in its area.*

The second hypothesis expresses that a manuscript contributing to knowledge reflects originality or motivation.

Hypothesis 2. *Elements describing contribution knowledge in a manuscript. The manuscript is salient in contribution knowledge if it receives good feedback on originality or motivation.*

14680394, 2025, 10, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/exsy.70131 by University Of Granada, Wiley Online Library on [27/10/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 2 | Aspect categories used in our proposal.

Aspect category	Description	Positive example	Negative example
Motivation/impact	Is the issue addressed in the paper of significant importance?/Is it probable that practitioners or researchers will utilise or expand upon these ideas?	The topic investigated in this study holds importance as comprehending the predictive uncertainty of a deep learning model is valuable from both theoretical and practical perspectives.	The approach has constraints regarding practical applicability and providing insight to the reader.
Originality	Is there a new research topic, technique, methodology or insight?	The novel proposed has the potential to enhance the speed of the learning process.	The authors' approach offers no progress compared to similar works.
Soundness/ correctness	Are the assertions in the document adequately substantiated?/Is the proposed approach sound?	The clarity and persuasiveness of illustrations using simulated and actual data are also highly evident.	There is insufficient theoretical backing for the approach.
Substance	Are there sufficient experiments in the paper to showcase the effectiveness of the proposed methods?/ Are there comprehensive result analyses available?/ Does it include significant ablation studies?	This is a comprehensive investigation into a predominantly overlooked issue.	The experimental study appears to lack an adequate number of experiments to showcase the benefits.
Meaningful comparison	Do the comparisons to previous research adequately?	The authors effectively position their study with previous research on double-blind techniques.	The experimental research could involve additional comparisons using challenging datasets that contain a greater number of classes.
Clarity	Can an adequately prepared reader understand the paper's aim and purpose? Is the paper effectively written and organised?	The paper is skilfully crafted and straightforward to comprehend.	The paper lacks proper organisation in its structure.

TABLE 3 | Examples of aspect categories with sentiment analysis scores for each manuscript (identified by ID column).

ID	Originality	Substance	Motivation	Vlarity	Soundness	Meaningful comparison
ICLR_2017_1	2	0	0	2	0	0
ICLR_2017_10	0	−1	2	−1	−5	−2
ICLR_2017_100	2	−2	1	1	−1	−1
ICLR_2017_101	3	1	0	1	1	−1

Note: A score with a negative value represents a negative polarity, and a positive value represents a positive sentiment.

The third hypothesis states that a manuscript's experimentation level reflects substance or meaningful comparison.

Hypothesis 3. *Elements describing a level of experimentation in a manuscript. The manuscript is salient in experimentation if it receives good substance or meaningful comparison feedback.*

Stage 2 in Figure 1 shows a graph representing these hypotheses. A multivariate analytic method assessed the research model hypotheses (see Stage 2 in Figure 1) and estimated the structural or inner model and a measurement or outer model (Fornell and Larcker 1981). The structural or inner model pertains to the connections among constructs or latent variables within the model. In our case, the constructs are accessibility,

14680394, 2025, 10, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/exsy.70131 by University Of Granada, Wiley Online Library on [27/10/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 4 | The titles and URLs of the manuscripts are shown in Table 3.

ID	Title	URL
ICLR_2017_1	Making Neural Programming Architectures Generalize via Recursion	http://openreview.net/pdf/342543971002b3e5f08be11d9a6da60b594a6b47.pdf
ICLR_2017_10	Q-Prop: Sample-Efficient Policy Gradient with An Off-Policy Critic	http://openreview.net/pdf/c210ff1a4868a532ec87ee0da3c6e4254ee567fb.pdf
ICLR_2017_100	Introspection: Accelerating Neural Network Training By Learning Weight Evolution	http://openreview.net/pdf/f5316305b0560db063525a71f36ca95d1932981e.pdf
ICLR_2017_101	Hyperband: Bandit-Based Configuration Evaluation for Hyperparameter Optimization	http://openreview.net/pdf/a9263ffc193997e5041e73ce2230de60001cd855.pdf

contribution, experimentation and decision. The first three are exogenous variables, and decision is an endogenous variable. Therefore, accessibility, contribution and experimentation predict decision. The real values of decision have been taken in the following way (see Equation 1): if the editor decides to accept, and the number of aspects with positive polarity is bigger than the number of aspects with negative polarity, the decision takes the rate between the number of positive aspects and the total aspects. In another case, if the editor decides to reject and the number of aspects with negative polarity is bigger than the number of aspects with positive polarity, the decision takes the negative rate between the number of negative aspects and the number of total aspects. The value in Equation (1) reflects the correlation between the editor's decision and the number of positive and negative aspects considered. If the editor decides to accept (reject) the manuscript, the number of positive (negative) aspects must be more significant than the negative (positive) aspects, and the rate number of positive (negative) aspects over the total is close to 1 (−1).

$$\text{Decision (id)} = \begin{cases} \dfrac{\#\text{aspect}_{\text{pos}}}{\#\text{aspects}} & \text{if decision}_{\text{editor}}(\text{id}) = \text{accept and} \#\text{aspect}_{\text{pos}} > \#\text{aspects}_{\text{neg}} \\ -\dfrac{\#\text{aspect}_{\text{neg}}}{\#\text{aspects}} & \text{if decision}_{\text{editor}}(\text{id}) = \text{reject and} \#\text{aspect}_{\text{pos}} < \#\text{aspects}_{\text{neg}} \\ 0 & \text{in other case} \end{cases} \tag{1}$$

The measurement or outer model comprises the indicators (observed variables) or items and the relationships with their constructs. In the model in Figure 1, clarity, soundness, originality, motivation, substance and meaningful comparison are the indicators, and the associated constructs are accessibility, contribution and experimentation. These indicators are reflective; therefore, they are reflections of the associated construct. Besides, the indicators related to the same construct must have a strong association. Also, every indicator has a strong membership with a constructor and a weak one with the rest.

In Sections 4 and 5, we establish if Hypothesis 1, Hypothesis 2 and Hypothesis 3 are met in our dataset.

3.3 | Stage 3: To Tagger by New Constructs

At the end of Stage 2, we have defined a set of constructs (accessibility, contribution and experimentation) based on the indicators observed in the reviewer's report (clarity, soundness, substance, meaningful comparison, motivation and originality). In this stage, we aim to define the first deep learning machine to label the sentences of the reviewer's report by accessibility, contribution and experimentation construct. Next, a second deep learning machine, based on the accessibility, experimentation and contribution label given to a sentence, will get the aspect category. Thus, if a sentence has been labelled with accessibility, following Hypothesis 1, the second deep learning machine will label the sentence with clarity or soundness. In the same way, for the contribution and experimentation, following Hypothesis 2 and Hypothesis 3, respectively, the second deep learning machine will label the sentence with the related aspects.

Therefore, in this stage, we have two goals:

1. To transform the reports' dataset, labelled by aspect categories and polarity, to reports including labels by accessibility, contribution and experimentation constructs. For this goal, we analyse the dataset, and a sentence labelled by clarity and soundness is also labelled by accessibility, keeping the same polarity. In the same way, sentences labelled with originality or motivation are tagged by contribution, and those labelled with substance or meaningful comparison are tagged as experimentation. This straightforward retagging process is applied, provided Hypothesis 1, Hypothesis 2 and Hypothesis 3 will have been tested by the PLS-SEM model described in Stage 1 in Section 3.
2. To train a new machine based on deep ML. The goal is to label a sentence using the new constructs of accessibility, contribution and experimentation. Once ML is obtained to label the sentences with construct categories and sentiment analysis, the machine learns every sentence's aspect category.

   Thus, we can describe our problem as follows:

   Let $\{\langle i_n, c_n, a_n, s_n\rangle\}^{|I|}_{n=1}$ be a set of quadruples being c_n the construct category (accessibility, contribution and experimentation), a_n the aspect category and s_n sentiment labels for the i_n sentence of the reviewer's report. In this case $s_n \in S$ being $S = \{positive, negative\}$ and a_n take a value between {clarity, soundness, originality, motivation, substance, meaningful comparison}. Given a sentence, the aim is to assign it a proper construct, aspect and sentiment label. For this, we proposed a machine based on deep learning.

14680394, 2025, 10, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/exsy.70131 by University Of Granada, Wiley Online Library on [27/10/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

3.4 | Architecture Description

To train a deep ML model capable of labelling sentences in a report with accessibility, contribution, experiment constructs and polarity, we have followed the architecture presented in Kumar et al. (2021) and variations of this architecture presented in Bharti (2024). The AccConExp model proposed is shown in Figures 2 and 3.

The architecture presented in Kumar et al. (2021) and Bharti (2024) is a multi-task learning model (Caruana 1998) for detecting aspect categories and sentiments. In our case, we have taken this architecture to tagger construct categories and sentiments. Thus, the AccConExp model is composed of two stages: in the first stage, we label the sentences of the report by accessibility, contribution, experimentation and sentiment (see Figure 2). In the second stage (see Figure 3), we annotate the sentences by an aspect category conditioned to the construct category obtained in the first stage.

In the first stage, the AccConExp model splits the input report into different sentences $\{s_1, s_2, \cdots, s_n\}$, and every sentence is composed of different words $s_i = \{w_1^i, w_2^i, \cdots, w_t^i\}$. Besides, the [CLS] and [SEP][6] tokens are used to establish the bounds of the sentences (see Figure 2). These sentences are encoded by using SciBert (Beltagy et al. 2019). The output for every sentence is a d-dimensional embedding vector $e_i \in R^d$. Until this stage, the set of report's sentences is represented by a matrix $Rep \in R^{N \times d}$, being N the number of sentences, and $Rep = e_0 \otimes e_1 \otimes \cdots \otimes e_N$.

The BiLSTM layers (Bidirectional Long Short-term Memory) establish the relationships between the subsequences in every sentence, providing a semantic context (Hochreiter and Schmidhuber 1997; Li et al. 2016). BiLSTM layers have as an underlying idea to know when to forget or remember the context of a word. Being bidirectional, the analysis is done from left to right and vice versa.

In the BiLSTM stage (see Figure 2), the output hidden representation $(h_1, h_2, \cdots, h_n)$ feeds the following attention layers. In this case, n is the number of BiLSTM units.

In the first stage, the attention layer is associated with a construct category, and in the second stage (see Figure 3), it is associated with an aspect category. Attention layers create a weight to weigh the outputs obtained from the BiLSTM layer. As we have said, the attention layer is associated with a constructor, so words related to this constructor and its contexts (semantic and syntactic relationships with other words) should have a higher weight. Every attention layer is characterised by a code $c^i = (c_1^i, c_2^i, \cdots, c_n^i)$. Therefore, the set of codes $C_a = \{c^1, c^2, \cdots, c^K\}$ has as many elements as the number of aspect categories or construct categories. C_a is learned in the training.

In the *i-th* attention layer, the correlation between the layer's code c^i and the output hidden representation h_k is first calculated, followed by a softmax operation (see Equation 2).

$$\alpha_i^k = \frac{\exp(c^i \cdot h_k)}{\sum_{j=1}^{n} \exp(c^i \cdot h_j)} \tag{2}$$

The α_i^k value expresses the relevance of the construct category (or aspect category) in the output hidden representation. Finally,

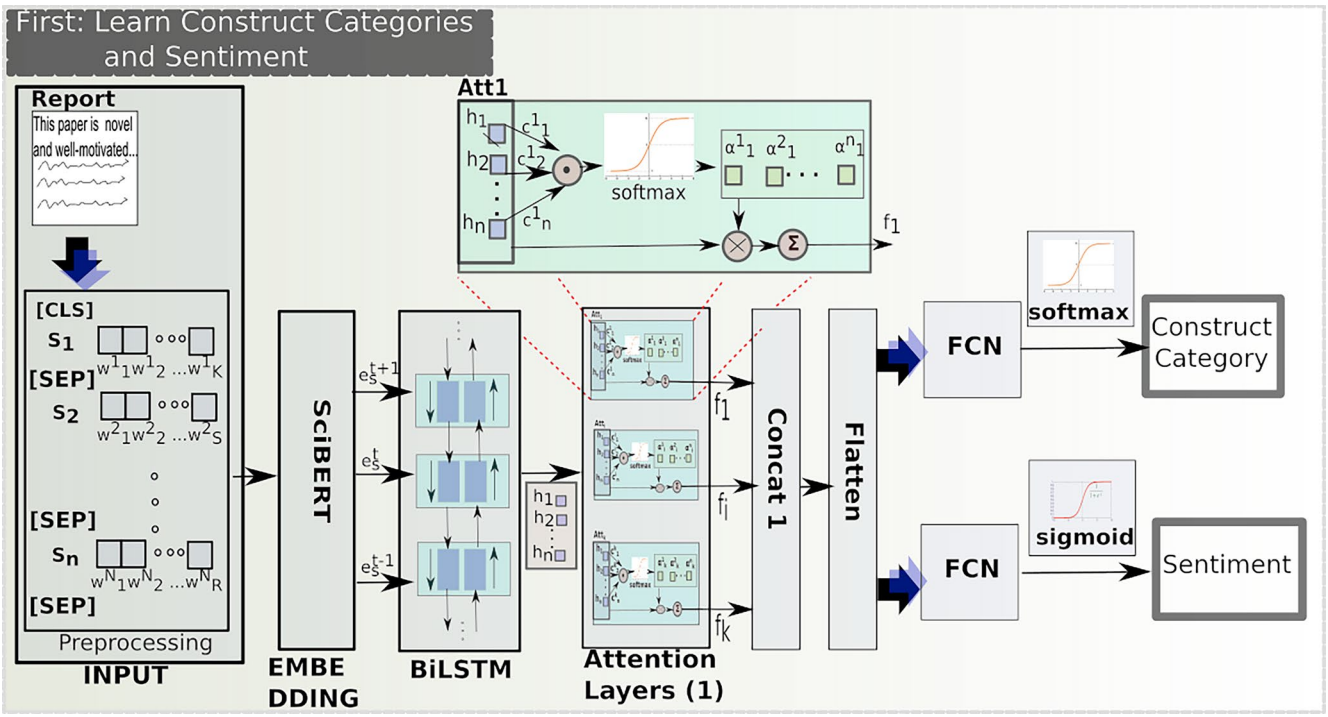

FIGURE 2 | AccConExp deep learning machine scheme to learn to tag reviewer report sentences using construct categories and sentiment analysis.

14680394, 2025, 10, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/exsy.70131 by University Of Granada, Wiley Online Library on [27/10/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

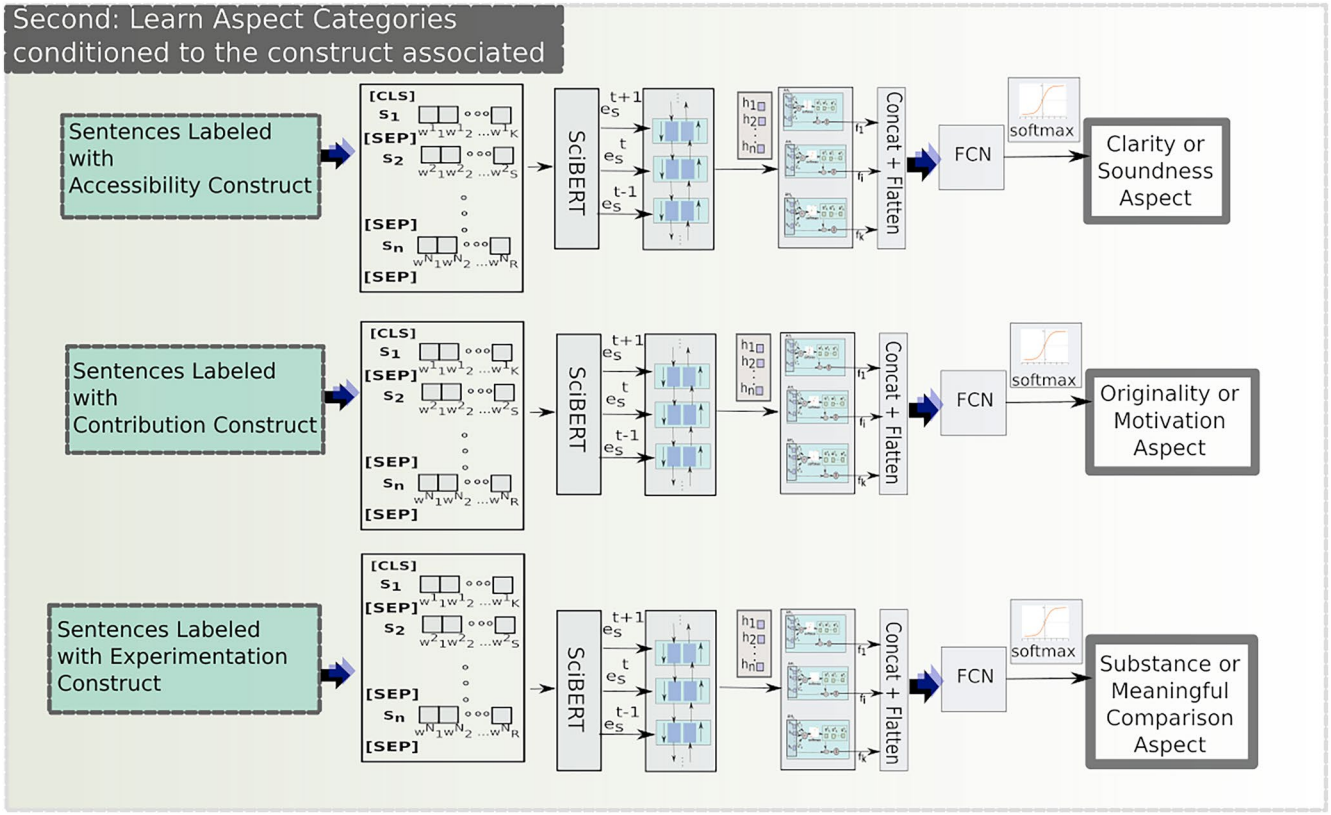

FIGURE 3 | AccConExp deep learning machine's scheme to assign an aspect category conditioned to the sentence's construct category.

the output value f_i (see Equation 3) can be interpreted as the amount of the i-th construct category (or aspect category) in the sentence.

$$f_i = \sum_{k=1}^{n} \alpha_i^k h_k \tag{3}$$

In the next block, the outputs f_* are concatenated and flattened. In Figure 2, two branches carry out two of the goals of the model. The first branch (Branch 1) pretends to assign a construct category to the report, while the second branch (Branch 2) tries to give a sentiment to the construct category of the report. Both branches begin with a dense layer. Next, Branch 1 applies a nonlinear function by a softmax operation, while Branch 2 uses a sigmoid operation. As the loss function, Branch 1 uses a categorical cross-entropy and Branch 2 a binary cross-entropy function. Both branches train simultaneously, using a loss function total defined as the sum of the weighted loss functions of every branch.

In our case, the output's Branch 1 will contain three values associated with the probability of accessibility, contribution and experimentation. The output's Branch 2 will be 0 (negative) or 1 (positive).

Figure 3 describes the three architectures to label the aspect category of the sentence. The three architectures are identical and similar to the architecture of Branch 1 in Figure 2. Depending on the construct category given to the sentence, one of the three architectures is chosen. For example, if the sentence $\boldsymbol{i}$ was labelled with the accessibility category and positive sentiment, the machine in the second stage obtains the label between clarity and soundness. Thus, two quadruples are possible for the sentence $\boldsymbol{i}$: $< i, accessibility, clarity, positive >$, and $< i, accessibility, soundness, positive >$.

4 | Experimental Setup and Results

In this section, we explained the dataset used and the details of the implementation. The dataset used and code can be downloaded at https://github.com/rosadecsai/AccConExp.

Next, we will obtain the results of the PLS-SEM model to test if the hypothesis formulated in Section 3 is met.

Once we have tested that the hypothesis is met, we relabel the sentences of our dataset with the new constructs: accessibility, contribution and experimentation. With this modified dataset, we will train the deep ML described in the previous section. Finally, the results of this training process are shown.

4.1 | Dataset

We have used the ASAP-Review dataset proposed by Yuan et al. (2021) to test the hypothesis. The papers in this dataset are articles submitted to the ICLR (International Conference on Learning Representations) from 2017 to 2020 and the NIPS (Neural Information Processing Systems) from 2016 to 2019. The total number of papers considered is 8742. The dataset keeps the reviewers' reports for every article with the tagger, with aspect categories and polarity.

14680394, 2025, 10, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/exsy.70131 by University Of Granada, Wiley Online Library on [27/10/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

To train and test the proposed model, we also used the ASAP-Review dataset[7] (Yuan et al. 2021). Proceeding has annotations about aspect category and polarity.

This dataset was updated in the following way:

- If a sentence is labelled with clarity or soundness, it is added a label of accessibility. The sentiment will remain the same.
- If a sentence is labelled with originality or motivation, the sentence is also labelled with contribution. The sentiment will be the same.
- If a sentence is labelled with substance or meaningful comparison, the sentence is also labelled with experimentation. The sentiment will be the same.

The total number of sentences in the dataset is 172,752. Every aspect category has a frequency of 28,792. The positive sentiment is counted 76,847 times, and the negative sentiment is 95,905.

4.2 | Implementation Details

To obtain the PLS-SEM model, we have used the library plspm[8] for Python.

The architectures have been coded in Python, adapting the code in Bharti (2024).

The libraries needed are Keras with TensorFlow-GPU. The executions were realised under Tesla V100 GPUs with 16GB VRAM.

For the architectures presented in Section 3.4, we used an epoch number of 150 and a batch size of 64. For the architecture in Figure 2, the number of attention layers was 3, since each layer of attention is associated with a construct. For every machine in Figure 3, the number of attention layers was 2, since every construct is associated with two aspects.

To train the model, the dataset was split 70% for training and 30% for validation.

4.3 | Results

In this section, we will describe the results of the two stages of the AccConExp model. First, we will describe the PLS-SEM results, which will test whether the hypothesis described in Section 3 is met. Second, we will train deep ML to label sentences based on accessibility, contribution and experimentation. Next, we will also train the machine associated with the construct to distinguish the aspect.

4.3.1 | PLS-SEM Model

To evaluate the measurement or outer model, in Table 5, we show the different constructs and the indicators. The value of loading represents the correlation between a construct and its indicators. Indicators with a loading value of 0.7 or higher are considered highly satisfactory (Henseler et al. 2009; Götz et al. 2010). A loading value of 0.5 is regarded as acceptable; the indicators with a loading value of less than 0.5 should be erased (Chin 1998; Hair et al. 2013).

As it is shown in Table 5, there is a strong correlation between clarity and soundness with the Accessibility construct. Concerning the Contribution construct, Table 5 shows that the motivation loading value of 0.583 is acceptable, and the originality indicator has a loading value of 0.886. The substance and meaningful comparison indicators, associated with the Experimentation construct, have more than an acceptable loading value. We have obtained the cross-loadings to assess that every indicator has a high-level membership with its construct and a weak membership with the rest (see Table 6).

We also examined each construct's composite reliability to determine the internal consistency in scale items. Composite reliability is measured by taking into account the shared variance among items and the errors in measurement. Higher composite reliability suggests that the construct's components are closely related and measure the same fundamental construct. The acceptable composite reliability is usually 0.7 or higher (Hair et al. 2020). Table 5 shows that different constructs have acceptable composite reliability. Also, in Table 5, the average variance extracted (AVE) is shown. AVE measures the average variance extracted between the construct and its individual indicators. Following Hair et al. (2020), an acceptable value must be 0.5 or higher. In Table 5, the AVE values are 0.563, 0.590 and 0.610.

Once we have evaluated the measurement model, the next step is to evaluate the structural or inner model. To evaluate the structural model, a criterion is to observe the value, which is a

TABLE 5 | Constructs, indicators, loadings, composite reliability and average variance extracted.

Constructs	Indicators	Loading	Composite reliability	AVE
Accessibility	Clarity	0.783	0.766	0.610
	Soundness	0.785		
Contribution	Originality	0.886	0.720	0.563
	Motivation	0.583		
Experimentation	Substance	0.850	0.750	0.590
	Meaningful comparison	0.670		

14680394, 2025, 10, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/exsy.70131 by University Of Granada, Wiley Online Library on [27/10/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 6 | Cross-loadings.

	Accessibility	Contribution	Experimentation
Clarity	**0.783**	0.101	0.134
Soundness	**0.785**	0.157	0.263
Originality	0.096	**0.887**	0.060
Motivation	0.183	**0.584**	0.059
Substance	0.215	0.031	**0.855**
Meaningful comparison	0.175	0.101	**0.679**

Note: In blue, the highest value of the indicators.

TABLE 7 | Bootstrap results for R^2 and effect size.

	Original	Mean bootstrap	Std. error	perc.0.25	perc.975
R^2					
Decision	0.6594	0.6589	0.0051	0.6501	0.6656
Effects					
Accessibility	0.5139	0.5144	0.0055	0.5047	0.5226
Contribution	0.3889	0.3884	0.0055	0.3784	0.3976
Experimentation	0.2914	0.2896	0.0064	0.2722	0.2993

TABLE 8 | Bootstrap results for loadings.

	Original	Mean bootstrap	Std. error	perc.0.25	perc.975
Clarity	0.7831	0.7798	0.0093	0.7584	0.7954
Soundness	0.7851	0.7869	0.0102	0.7672	0.8073
Originality	0.8868	0.8862	0.0099	0.8663	0.9005
Motivation	0.5835	0.5835	0.0177	0.5572	0.6201
Substance	0.8545	0.8543	0.0096	0.8367	0.8733
Meaningful comparison	0.6794	0.6799	0.0161	0.6514	0.7079

measure that indicates the endogenous constructs' predictive ability (only for the sample of data). In our model, the endogenous construct is decision, and the value is 0.6, which is acceptable. The Goodness of Fit (GoF) index is a simulated measure of GoF that considers the quality of the model in terms of both the measurement and structural aspects. In our case, GoF is 0.624.

The effect size is another interesting measure of the structural model's predictive ability. This measure estimates the predictive ability of each independent construct in the model. Thus, in the model in Table 7, the effect values are 0.51 for accessibility, 0.38 for contribution and 0.29 for experimentation. A value of 0.35 and above is labelled as a large effect, and values of 0.15 and up to 0.35 are medium (Cohen 1988).

We have verified that the results obtained from both the inner and outer models are consistent. Next, we proceed to validate the model using the Bootstrap method, applying 100 bootstrap samples. The Bootstrap process involved applying the PLS-SEM model to each of the 100 samples.

TABLE 9 | Mean, standard deviation and percentiles of the values predicted by the model for the accessibility, contribution and experimentation constructs for the manuscripts accepted.

	Mean	Std	perc.0.25	perc.75
Accessibility	0.609	0.097	0.561	0.659
Contribution	0.495	0.105	0.422	0.555
Experimentation	0.579	0.100	0.533	0.635

Tables 7 and 8 show the bootstrap results obtained for R^2, effect size and loadings. To test the results' robustness, we have calculated the mean, standard deviation, 25th percentile, and 97.5th percentile across the bootstrap samples.

14680394, 2025, 10, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/exsy.70131 by University Of Granada, Wiley Online Library on [27/10/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 10 | The number of training parameters and accuracy for the training validation set for the AccConExp model (see Figure 2).

Model	#Par.	Construct-category accuracy	Construct-category validation accuracy	Sentiment accuracy	Sentiment validation accuracy
AccConExp	3,733,023	0.99	0.93	0.99	0.98

TABLE 11 | Accurate and F1 metrics for every construct and sentiment for the AccConExp model (see Figure 2).

		AccConExp model	
		ACC	F1
Construct category	Accessibility	0.90	0.91
	Contribution	0.95	0.94
	Experimentation	0.95	0.94
Sentiment	Positive	0.98	0.98
	Negative	0.98	0.97

The original column in Tables 7 and 8 shows the results obtained with the datasets. The mean Bootstrap column is the mean value obtained across the 100 bootstrap samples. To prove that this mean Bootstrap value represents the samples, we have also shown the standard error, 25th and 97.5th percentiles. The mean Bootstrap values are similar to the original values.

Next, we show in Table 9 the mean, standard deviation, 25th percentile and 75th percentile for the accessibility, contribution and experimentation of the predicted data when the final decision was to accept the manuscript.

The total number of manuscripts accepted is 5309. Of these, 3783 have a value of accessibility around the accessibility mean value (mean ± std), 3831 around the contribution mean value and 3862 around the experimentation mean value. The number of manuscripts with values in accessibility, contribution and experimentation around the mean value simultaneously is 2098.

4.3.2 | Construct and Aspect Tagger

In this section, we will analyse whether the AccConExp model can annotate the dataset based on the new constructs defined (accessibility, contribution and experimentation) while maintaining the accuracy of the annotation based on aspect categories and sentiment analysis.

For this, the results obtained for the AccConExp model labelling accessibility, contribution, experimentation and polarity (see Figure 2) for every sentence of the dataset's report are shown in Tables 10 and 11. The AccConExp model got a construct category validation accuracy of 0.93 and a sentiment validation accuracy of 0.98. Table 11 shows the accuracy and F1 metrics for every construct category and sentiment.

Once the validity of the AccConExp model to label the sentences through accessibility, contribution and experimentation has been verified, we will test whether, with this information, we can obtain the aspect category conditioned to the corresponding construct category (see Figure 3). Table 12 shows the results of tagging every report's sentences with aspect categories and sentiment. The AccConExp model was compared with the model in Bharti (2024). The architecture proposed in Bharti (2024) is the same architecture shown in Figure 2, and the only difference is that the construct-category box is replaced by the aspect-category box; in this case, the number of attention layers is 3 instead of 6 in Bharti (2024). The results in Table 12 show that the AccConExp model gets the precision, recall and F1 values with few variations from the results obtained by the model in Bharti (2024). We have considered variations of the model in Bharti (2024) to compare with the proposed model as follows:

- Variation 1: We have eliminated the BiLSTM layers, keeping the six layers of attention.
- Variation 2: We have eliminated the attention layers, keeping the BiLSTM layers
- Variation 3: We have eliminated the attention layers and the BiLSTM layers

Tables 13–15 compare the AccConExp model with the benchmark Variations 1–3. After analysing all the comparisons, we see that our proposal is very competitive.

Also, Table 16 shows some correct predictions obtained for the AccConExp model.

5 | Discussion

The results obtained by the AccConExp model manifest a new methodology for discovering new constructs in a dataset, thereby saving the effort required by a human expert to relabel the dataset with these new constructs. At the same time, the model proposes a scalable mechanism for tagging the dataset with the new constructs as well as the aspect categories.

The two parts of the model are:

- PLS-SEM model that allows the proof of the existence of new constructs in a dataset based on the observed information (aspect categories).
- An ML tagger to tag the sentences of a reviewer's report by using the new constructs and sentiment analysis.

14680394, 2025, 10, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/exsy.70131 by University Of Granada, Wiley Online Library on [27/10/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 12 | Comparison between the AccConExp model and the model in Bharti (2024).

		AccConExp			Bharti et al.		
		Precision	**Recall**	**F1**	**Precision**	**Recall**	**F1**
Aspect categories	Clarity	0.93	0.92	0.92	0.94	0.92	0.93
	Soundness	0.84	0.82	0.83	0.85	0.82	0.83
	Motivation	0.90	0.92	0.91	0.91	0.92	0.91
	Originality	0.92	0.91	0.92	0.92	0.91	0.92
	Meaningful comparison	0.93	0.96	0.94	0.93	0.95	0.94
	Substance	0.88	0.90	0.89	0.88	0.91	0.89
Sentiment	Positive	0.98	0.98	0.98	0.98	0.98	0.98
	Negative	0.97	0.98	0.97	0.97	0.97	0.97

TABLE 13 | Comparison between the AccConExp model and benchmark Variation 1.

		AccConExp			Variation 1		
		Precision	**Recall**	**F1**	**Precision**	**Recall**	**F1**
Aspect categories	Clarity	0.93	0.92	0.92	0.91	0.88	0.90
	Soundness	0.84	0.82	0.83	0.71	0.73	0.72
	Motivation	0.90	0.92	0.91	0.84	0.82	0.83
	Originality	0.92	0.91	0.92	0.85	0.86	0.85
	Meaningful comparison	0.93	0.96	0.94	0.87	0.89	0.88
	Substance	0.88	0.90	0.89	0.79	0.79	0.79
Sentiment	Positive	0.98	0.98	0.98	0.97	0.95	0.96
	Negative	0.97	0.98	0.97	0.94	0.97	0.95

TABLE 14 | Comparison between the AccConExp model and benchmark Variation 2.

		AccConExp			Variation 2		
		Precision	**Recall**	**F1**	**Precision**	**Recall**	**F1**
Aspect categories	Clarity	0.93	0.92	0.92	0.94	0.92	0.93
	Soundness	0.84	0.82	0.83	0.85	0.83	0.84
	Motivation	0.90	0.92	0.91	0.91	0.92	0.92
	Originality	0.92	0.91	0.92	0.92	0.92	0.92
	Meaningful comparison	0.93	0.96	0.94	0.93	0.96	0.95
	Substance	0.88	0.90	0.89	0.88	0.91	0.90
Sentiment	Positive	0.98	0.98	0.98	0.98	0.98	0.98
	Negative	0.97	0.98	0.97	0.97	0.98	0.97

14680394, 2025, 10, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/exsy.70131 by University Of Granada, Wiley Online Library on [27/10/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 15 | Comparison between the AccConExp model and benchmark Variation 3.

		AccConExp			Variation 3		
		Precision	**Recall**	**F1**	**Precision**	**Recall**	**F1**
Aspect categories	Clarity	0.93	0.92	0.92	0.93	0.91	0.92
	Soundness	0.84	0.82	0.83	0.79	0.78	0.79
	Motivation	0.90	0.92	0.91	0.88	0.88	0.88
	Originality	0.92	0.91	0.92	0.90	0.89	0.89
	Meaningful comparison	0.93	0.96	0.94	0.92	0.94	0.93
	Substance	0.88	0.90	0.89	0.84	0.86	0.85
Sentiment	Positive	0.98	0.98	0.98	0.97	0.96	0.97
	Negative	0.97	0.98	0.97	0.96	0.96	0.96

TABLE 16 | Some sentence examples classified by the AccConExp model.

Sentence examples	**Construct category**	**Aspect category**	**Sentiment**
The proofs are generally written very clearly, and even when they are not in the main body, enough explanations are given to make the results pretty intuitive	Accessibility	Clarity	Positive
The mathematical notations are overly verbose, which makes the paper harder to understand	Accessibility	Clarity	Negative
The presented method achieves excellent results compared with other state-of-the-art algorithms	Accessibility	Soundness	Positive
The method appears a bit ad hoc in that it has many components, but the effect of each component is not evaluated individually.	Accessibility	Soundness	Negative
This paper presents a nice way of generating saliency maps from activations inside a network	Contribution	Originality	Positive
It is not clear to me that the work contributes any new ideas beyond those already introduced in the following paper	Contribution	Originality	Negative
The paper is, in general, very interesting because it shows that a unit trace-constrained PSD cone may be easier to solve than the PSD cone	Contribution	Motivation	Positive
I believe that if the paper could provide more evidence about its potential influence in real applications	Contribution	Motivation	Negative
The authors also provide experiments that confirm the theory and also provide examples highlighting the gap	Experimentation	Substance	Positive
The experiment results could be discussed more	Experimentation	Substance	Negative
The relation to previous works is detailed and clear	Experimentation	Meaningful comparison	Positive
The paper currently does not contain some very relevant baselines	Experimentation	Meaningful comparison	Negative

14680394, 2025, 10, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/exsy.70131 by University Of Granada, Wiley Online Library on [27/10/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

5.1 | PLS-SEM Model

Our first aim has been to test the three hypotheses defined in Section 3. These hypotheses establish the association between the new constructs (accessibility, contribution and experimentation) and the aspect categories.

To test the hypotheses, we have obtained Table 5. This table shows that the PLS-SEM model is coherent (see values of AVE, composite reliability in Table 5). The hypotheses are also proven by analysing Table 6:

- Hypothesis 1 is validated as the maximum loadings associated with accessibility are clarity and soundness (see Table 6). These two aspects, clarity and soundness, exhibit low loading values when they are associated with contribution and experimentation. Table 8 shows the result of applying bootstrapping to 100 samples to test the robustness of the association between the new constructs and aspects. The average values are similar to the original values obtained for the entire dataset with minimal standard deviations.
- Hypothesis 2 associates contribution with originality and motivation. Analysing Table 6 reveals that this hypothesis is fulfilled since the aspects with the highest loading values are originality and motivation. Table 8 also validates this result by using 100 samples and applying bootstrapping.
- Finally, Hypothesis 3 associates experimentation with meaningful comparison and substance. Again, analysing Table 6, we see that the higher loading values with the experimentation construct are given by meaningful comparison and substance. The robustness of these results is validated by bootstrapping (see Table 8).

After assessing the measurement model, the subsequent phase involves evaluating the structural or inner model. A criterion is employed to examine the R^2 value for this purpose. The R^2 value serves as an indicator of the predictive capability of the endogenous constructs. Therefore, we have studied whether the new construct accessibility, contribution and experimentation, associated with their respective aspects, can predict decision (defined as in Equation 1, in order to accept or reject a paper). In our model, the endogenous construct is decision, and the R^2 value is 0.6 (see Table 7), which is deemed acceptable. Also, to validate these results, we have applied bootstrap (see Table 7).

Another analysis that deserves attention is the average values that the constructs adopt when the editor accepts the article. These data are shown in Table 9. For our database, the average value of accessibility is 0.61, that of contribution is 0.5 and that of experimentation is 0.6. These data are coherent since the articles are engineering papers (ICLR and NeurIPS papers).

To extend this study, we have created a new dataset that contains the ICLR papers published from 2018 to 2023. We have replicated this study for this new dataset (see Appendix A).

5.2 | Tagger

Once we had proven the three hypotheses, we developed a system based on deep ML to tag reports' sentences (seen or unseen before) by accessibility, contribution and experimentation constructs, positive and negative polarity and aspects categories (clarity, soundness, originality, motivation, substance and meaningful-comparison). The system proposed in the first stage tagged the sentences in accessibility, contribution, experimentation and polarity (see Figure 2). In the second stage, conditioned to the construct assigned to a sentence, this is labelled by the associated aspects of the construct (see Figure 3).

The results of tagging the dataset by using the construct and polarity are presented in Tables 10 and 11. The AccConExp model achieved a validation accuracy of 0.93 for construct categories and 0.98 for sentiment analysis. Table 11 presents the accuracy and F1 metrics for each construct category and sentiment. To test the goodness of the AccConExp model, we obtained the aspect categories using the scheme shown in Figure 3 and the results were compared (see Table 12) with the results obtained by the model in Bharti (2024). Also, we have compared the AccConExp model with other benchmarks (see Tables 13–15).

The results demonstrate that the proposed model is compelling.

Another problem that deep learning architectures face is their behaviour when used with data with unseen labels. We have tested the AccConExp model with sentences[9] from the original dataset labelled with the replicability aspect category. This category was not used by the training because the number of sentences labelled with this category was low, and there is an unbalanced relationship between positive and negative replicability. A subset of sentences labelled with replicability is shown in Table 17. Also, we can observe the constructs, polarities and aspects given by the AccConExp model.

Figure 4 shows the distribution of the sentences labelled with replicability. This aspect has not been seen in the training of the AccConExp model. The top graph in Figure 4 shows the distribution of the sentences in construct and polarity. As can be observed in Table 18, the most frequent construct and polarity is accessibility with negative polarity (56.7%), followed by experimentation with negative polarity (31%). The less frequent construct and polarity is contribution with positive polarity (1.9%) followed by experimentation and positive polarity (2%). The rest of the graphs in Figure 4 show the distribution of the sentences in construct, polarity and their aspect categories. Table 19 shows the distribution of the sentences by construct, polarity and aspect category.

5.3 | Contexts Where It Could Be Helpful to the AccConExp Model

With the results obtained, some contexts in which it will be adequate to use the AccConExp model are the following:

14680394, 2025, 10, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/exsy.70131 by University Of Granada, Wiley Online Library on [27/10/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

- If a journal editor needs to analyse reports with a granularity of information based on accessibility, contribution, experimentation and the corresponding sentiment, they can obtain this information from the reports' sentences tagged and make a decision. In this situation, at the computational level, the number of training parameters is less than that used (Bharti 2024). In Bharti (2024), the number of trainable parameters is 5,365,725, and the AccConExp model to tagger in accessibility, contribution and experimentation is 3,733,023.
- For certain reports, once the editor obtains information on accessibility, contribution and experimentation information, greater granularity may be required based on aspect categories such as clarity, motivation, originality, substance, meaningful comparison, or soundness, to support the decision-making process. At a computational level, this second architecture of the AccConExp model requires fewer examples to learn due to the previous classification into construct categories.
- If the dataset is increased with new sentences classified into additional aspect categories, and these new categories result in the creation of novel constructs, the model would only need to retrain the architecture shown in Figure 2 and add a mini architecture for each new construct that determines the aspect categories associated with it.

5.4 | Limitations

The AccConExp model has been tested over a dataset belonging to engineering (ICLR and NeurIPS are ML conferences). In this research field, the contribution, experimentation and clear exposition (accessibility) can determine the quality of the paper. Depending on the research field, the journal's editor can give more weight to accessibility and less to contribution and experimentation, as it can in journals in the Humanities area.

In other research areas, such as medicine, the aspect categories can differ from those treated in this model. For example, in medicine, the aspect categories must consider other aspects, in addition to those taken into account in this paper, such as those related to the patient, problem or population, intervention, appropriate comparisons and outcome measures (Gülpınar and Güçlü 2013). These aspect categories could give others definitions of the construct. In the future, a possible line of research could involve exploring different research areas where the indicators obtained from reviewers' reports differ, and studying the corresponding constructs to develop a machine capable of labelling review reports in that area. Alternatively, building on the approach in Han et al. (2022) for future lines of work, we should investigate the application of an algorithm to identify

TABLE 17 | Some results of the AccConExp model for sentences with the replicability aspect. The replicability aspect was unseen in the training process of the AccConExp model.

Sentence	Construct	Polarity	Aspect category
The experiments do not provide enough details on the implementation to judge their significance; for example, 2× gain in speed could be achieved by better software implementation of the same algorithm.	Accessibility	Negative	Soundness
The paper is not self-contained, important methodological aspects of the method are insufficiently described.	Accessibility	Negative	Clarity
It is great that the authors provided the hyperparameter search details for PTN, CNN2 and Capsule nets.	Accessibility	Positive	Soundness
Experimental details are given in Appendix A, facilitating reproducibility of the results.	Accessibility	Positive	Clarity
Several figure captions should be updated to clarify which model and dataset are studied.	Experimentation	Negative	Substance
Details for the stacking process are provided.	Experimentation	Positive	Substance
The model has been constructed, and details of the model have been provided as the novelty of the proposed model.	Experimentation	Positive	Meaningful comparison
The details of experimental design, architecture and hyperparameter choice are not provided in the earthquake application.	Experimentation	Negative	Meaningful comparison
Overall, there are interesting new ideas, a new model, insufficient model description and experimental details.	Contribution	Positive	Originality
I am not sure how this is achieved in this work.	Contribution	Negative	Originality
The intuition and feasibility of identifying 'good' matrices (Defs 0.1 and 2) should be detailed.	Contribution	Positive	Motivation

14680394, 2025, 10, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/exsy.70131 by University Of Granada, Wiley Online Library on [27/10/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

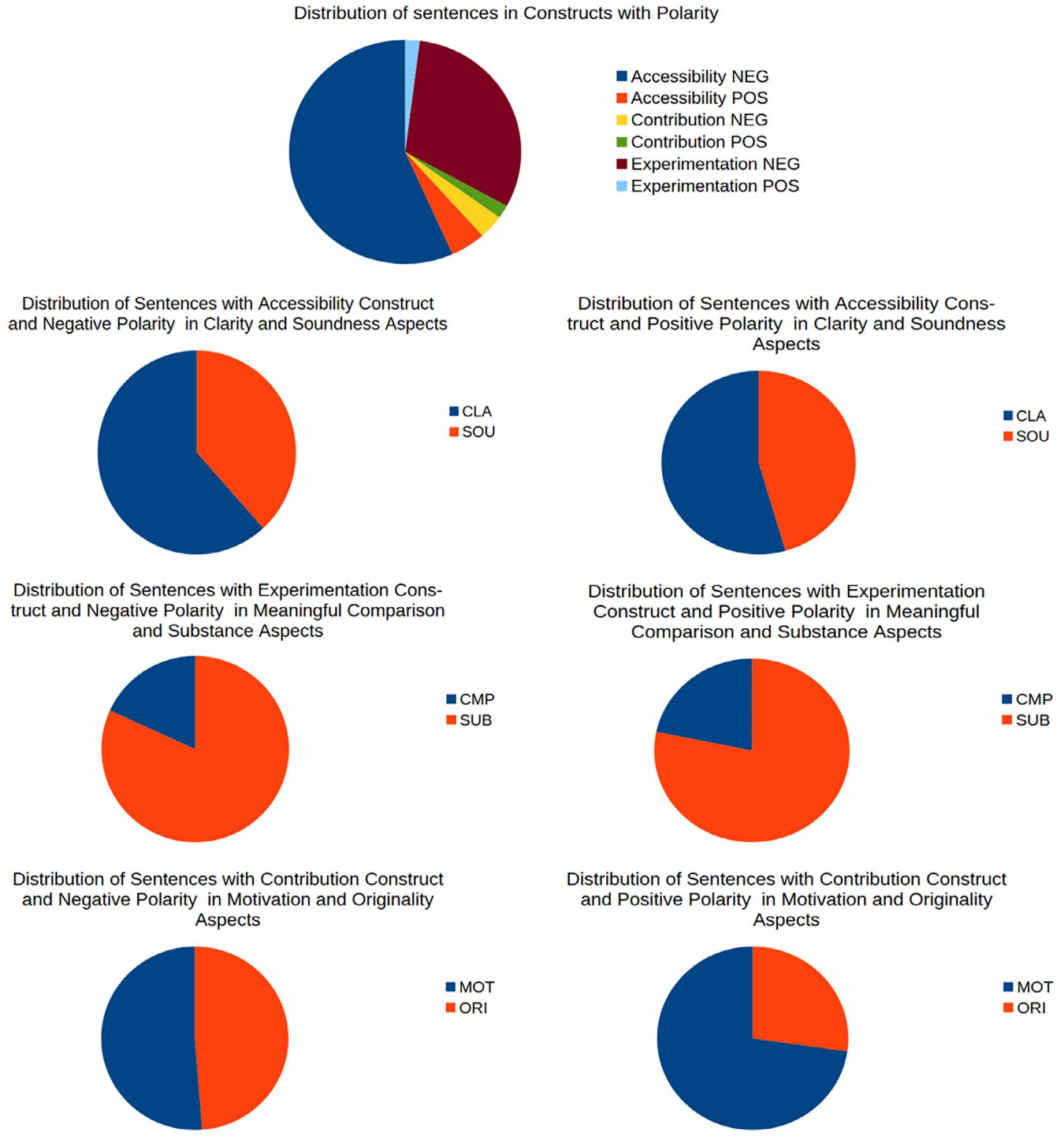

FIGURE 4 | Distribution of the results of the AccConExp model for sentences with the replicability aspect. The replicability aspect was unseen in the training process of the AccConExp model. The top graph shows the distribution of the sentences in construct and polarity. The other graphs show the distribution of the sentences in construct, polarity and their associated aspect categories.

aspect categories and subsequently conduct a PLS-SEM analysis to derive the different constructs. In this way, the aspect categories will be dependent on the corpus considered.

Finally, to make the model more flexible and to deal with review reports in a language other than English, it would be feasible first to use a machine translation system to convert the reports into English. After this translation, the analysis could be performed in terms of accessibility, contribution and experimentation along with the polarity. Alternatively, a more resource-intensive option would be to develop a separate AccConExp model for each language. However, this approach would require vastly more significant resources to establish the corresponding constructs and build database annotations in the desired language.

14680394, 2025, 10, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/exsy.70131 by University Of Granada, Wiley Online Library on [27/10/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 18 | Distribution of sentences by the AccConExp model with unseen replicability aspect.

Construct	Polarity	Number of sentences	Percentage
Accessibility	NEG	1297	56.69%
	POS	112	4.90%
Contribution	NEG	80	3.50%
	POS	44	1.92%
Experimentation	NEG	709	30.99%
	POS	46	2.01%

TABLE 19 | Distribution of sentences by the AccConExp model with unseen replicability aspect. The distribution of constructs and aspect categories is shown.

Construct	Polarity	Aspect category	Number of sentences
Accessibility	NEG	Clarity	800
		Soundness	497
	POS	Clarity	61
		Soundness	51
Contribution	NEG	Motivation	41
		Originality	39
	POS	Motivation	32
		Originality	12
Experimentation	NEG	Meaningful comparison	129
		Substance	580
	POS	Meaningful comparison	10
		Substance	36

Future work lines could also deal with non-textual information (graphs or tables) by analysing the sentences or captions that describe or comment on that non-textual information. In such cases, a search would need to be conducted for sentences related to the non-textual content, allowing them to be evaluated in terms of accessibility, contribution and experimentation. If such sentences do not exist, in the case of images, automatic image description systems could be used to facilitate the model's later use.

6 | Conclusions

The paper proposes a new model called AccConExp, which aims to assist editors in evaluating reviewers' reports during the peer-review process. The model focuses on three key constructs: accessibility, contribution and experimentation, and it also analyses the sentiment associated with these constructs.

The AccConExp model integrates two main components:

- A theoretical model based on PLS-SEM: This component is used to acquire new knowledge and to build a causal prediction of a set of construct categories assigned to the reviewer's report. It helps in understanding the review based on the three main constructs mentioned above.
- Deep ML: This component is employed to learn and explore the knowledge acquired by the PLS-SEM model. It uses multi-task deep learning to label the sentences in the reviewer's report according to accessibility, contribution, experimentation and sentiment.

The process involves the following steps:

- The PLS-SEM part of the AccConExp model identifies and predicts the constructs of accessibility, contribution and experimentation in the reviewer's report, previously labelled with aspects-categories (clarity, soundness, originality, motivation, substance and meaningful comparison).
- The multi-task deep learning model then labels the sentences based on these constructs and their sentiment.
- A second deep ML model is used to discover and assign aspect categories such as originality, soundness, motivation, substance, and meaningful comparison to the sentences.

The AccConExp model is compared with a multi-task architecture that directly assigns aspect categories to the sentences. The results show that AccConExp provides competitive performance and offers new insights to the reviewer's reports, which can be valuable for editors in making informed decisions during the peer-review process.

The AccConExp model, which is designed to analyse reports based on accessibility, contribution and experimentation, can be particularly interesting in the following contexts:

- Journal editing and review process: Journal editors need to analyse submitted reports with a high level of detail, focusing on key constructs like accessibility, contribution and experimentation. The AccConExp model facilitates this by extracting relevant information and sentiment from the reports, enabling editors to make informed decisions. The model is computationally efficient, requiring fewer training parameters compared to previous models.
- Enhanced granularity for specific reports: For certain reports, editors might need a more granular breakdown of information based on specific aspects such as clarity, motivation, originality, substance, meaningful comparison or soundness. The AccConExp model's architecture supports this by allowing the extraction of detailed aspect-based information from the reports. This refined granularity helps editors evaluate the reports with fine detail.
- Adaptability to New Categories (*Scalability and Adaptation*): When the database is updated with new sentences categorised under new aspects, the model can adapt by retraining

14680394, 2025, 10, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/exsy.70131 by University Of Granada, Wiley Online Library on [27/10/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

only the upper architecture and adding a mini-architecture for each new construct. This modular approach ensures that the model remains flexible and scalable, efficiently incorporating new categories without the need for extensive retraining.

The AccConExp model's design emphasises computational efficiency and effective learning:

- Fewer training parameters: Compared to other models, the AccConExp model uses fewer training parameters to obtain a characterisation based on accessibility, contribution, experimentation and the corresponding sentiment, which means it can be trained faster and with less computational resource consumption.
- Focused learning on categorised examples: By classifying reports into constructs and further into aspect categories, the model can learn more effectively with fewer examples. This targeted learning approach enhances the model's ability to analyse and categorise new reports accurately.

The AccConExp model can be applied in various editorial and review processes:

- Academic journals: Editors can use the model to streamline the review process and ensure that reports meet high standards of accessibility, contribution and experimentation.
- Peer-review systems: The model can assist peer reviewers by providing detailed analyses of the reports, highlighting key aspects that need attention.
- Research evaluation: Institutions and organisations can use the model to evaluate research outputs, ensuring they align with desired quality metrics and standards.

In summary, the AccConExp model offers a robust and adaptable framework for analysing reports with a focus on accessibility, contribution and experimentation. Its computational efficiency and ability to scale with new categories make it a valuable tool for journal editors and other stakeholders in the academic and research community. Besides, the AccConExp model represents an innovative approach to enhancing the peer-review process by providing a more structured and informed analysis of reviewers' reports, which can lead to more consistent and high-quality evaluations of scientific work.

Data Availability Statement

The data that support the findings of this study are available in PLS-SEM data: https://drive.google.com/file/d/1pZNhjlFV7QmXlf7tTtajzJBusm9ZlII0/view?usp=sharing, Training Machine Learning data: https://drive.google.com/file/d/1ZKAAqKyaHHA_7qk6_XZ9FGLVJZDOxSob/view?usp=sharing, ICLR (2018–2023): https://www.kaggle.com/datasets/juanjomontero/iclr-papers-and-reviews-data-2018-2023, Replicability dataset: https://drive.google.com/file/d/1qCkBlTAqljkG5CeFLxLPzIhbpGNs8FHb/view?usp=sharing. These data were derived from the following resources available in the public domain: (1) https://github.com/neulab/ReviewAdvisor and (2) https://openreview.net/.

Endnotes

[1] Natural language processing.

[2] International Conference on Learning Representations.

[3] Neural Information Processing Systems.

[4] The annotator proposed by Yuan et al. (2021) can be downloaded at https://github.com/neulab/ReviewAdvisor/.

[5] https://acl2018.org/downloads/acl_2018_review_form.html.

[6] Scibert was pretrained with the format [CLS] sen 1 [SEP] sen 2 [SEP]. [SEP] is a separator token and [CLS] is a token that will be used to predict whether or not sen 2 is a sentence that directly follows sen 1.

[7] Download from https://github.com/PrabhatkrBharti/Aspect-category-and-sentiment-extraction.

[8] https://github.com/GoogleCloudPlatform/plspm-python.

[9] The sentences with replicability aspect can be downloaded at https://drive.google.com/file/d/1qCkBlTAqljkG5CeFLxLPzIhbpGNs8FHb/view?usp=sharing.

[10] The dataset can be downloaded at https://www.kaggle.com/datasets/juanjomontero/iclr-papers-and-reviews-data-2018-2023.

References

Alarcón García, R. 2022. "Lexical Simplification for the Systematic Support of Cognitive Accessibility Guidelines." http://hdl.handle.net/10016/35140.

Beltagy, I., K. Lo, and A. Cohan. 2019. "SciBERT: A Pretrained Language Model for Scientific Text." ArXiv. https://doi.org/10.48550/arXiv.1903.10676.

Bharti, P. K., M. Agarwal, and A. Ekbal. 2024. "Please Be Polite to Your Peers: A Multi-Task Model for Assessing the Tone and Objectivity of Critiques of Peer Review Comments." *Scientometrics* 129: 1377–1413. https://doi.org/10.1007/s11192-024-04938-z.

Bharti, P. K. 2024. "Identifying Aspect Categories and Their Sentiments in Scientific Peer Reviews." https://github.com/PrabhatkrBharti/Aspect-category-and-sentiment-extraction.

Bralić, N., A. Mijatović, A. Marušić, and I. Buljan. 2024. "Conclusiveness, Readability and Textual Characteristics of Plain Language Summaries From Medical and Non-Medical Organizations: A Cross-Sectional Study." *Scientific Reports* 14, no. 1: 6016. https://doi.org/10.1038/s41598-024-56727-6.

Brezis, E. S., and A. Birukou. 2020. "Arbitrariness in the Peer Review Process." *Scientometrics* 123, no. 1: 393–411. https://doi.org/10.1007/s11192-020-03348-1.

Caruana, R. 1998. "Multitask Learning." In *Learning to Learn*, edited by S. Thrun and L. Pratt, 95–133. Springer US. https://doi.org/10.1007/978-1-4615-5529-2_5.

Chakraborty, S., P. Goyal, and A. Mukherjee. 2020. "Aspect-Based Sentiment Analysis of Scientific Reviews." In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*, 207–216. IEEE. https://doi.org/10.1145/3383583.3398541.

Chen, X., and X. Zhang. 2025. "Large Language Models Streamline Automated Systematic Review: A Preliminary Study." ArXiv. https://arxiv.org/abs/2502.15702.

Chin, W. W. 1998. "The Partial Least Squares Approach for Structural Equation Modeling." In *Modern Methods for Business Research*, Methodology for Business and Management, 295–336. Lawrence Erlbaum Associates.

Cohen, J. 1988. *Statistical Power Analysis for the Behavioral Sciences*. 2nd ed. Routledge. https://doi.org/10.4324/9780203771587.

14680394, 2025, 10, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/exsy.70131 by University Of Granada, Wiley Online Library on [27/10/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Cortes, C., and N. D. Lawrence. 2021. "Inconsistency in Conference Peer Review: Revisiting the 2014 NeurIPS Experiment." ArXiv. https://arxiv.org/pdf/2109.09774.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. 2019. "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding." ArXiv. https://doi.org/10.48550/arXiv.1810.04805.

Fernández Pinto, M. 2023. "Methodological and Cognitive Biases in Science: Issues for Current Research and Ways to Counteract Them." *Perspectives on Science* 31, no. 5: 535–554. https://doi.org/10.1162/posc_a_00589.

Forgeard, M. J. C., and A. C. Mecklenburg. 2013. "The Two Dimensions of Motivation and a Reciprocal Model of the Creative Process." *Review of General Psychology* 17, no. 3: 255–266. https://doi.org/10.1037/a0032104.

Fornell, C., and D. F. Larcker. 1981. "Evaluating Structural Equation Models With Unobservable Variables and Measurement Error." *Journal of Marketing Research* 18, no. 1: 39–50. https://doi.org/10.2307/3151312.

Gao, Y., S. Eger, I. Kuznetsov, I. Gurevych, and Y. Miyao. 2019. "Does My Rebuttal Matter? Insights From a Major NLP Conference." In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, edited by J. Burstein, C. Doran, and T. Solorio, 1274–1290. Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1129.

García, J. A., R. Rodriguez-Sánchez, and J. Fdez-Valdivia. 2018. "Competition Between Academic Journals for Scholars' Attention: The 'Nature Effect' in Scholarly Communication." *Scientometrics* 115, no. 3: 1413–1432. https://doi.org/10.1007/s11192-018-2723-9.

Ghosal, T., S. Kumar, P. K. Bharti, and A. Ekbal. 2022. "Peer Review Analyze: A Novel Benchmark Resource for Computational Analysis of Peer Reviews." *PLoS One* 17, no. 1: e0259238. https://doi.org/10.1371/journal.pone.0259238.

Götz, O., K. Liehr-Gobbers, and M. Krafft. 2010. "Evaluation of Structural Equation Models Using the Partial Least Squares (PLS) Approach." In *Handbook of Partial Least Squares: Concepts, Methods and Applications*, edited by V. E. Vinzi, W. W. Chin, J. Henseler, and H. Wang, 691–711. Springer. https://doi.org/10.1007/978-3-540-32827-8_30.

Gülpınar, Ö., and A. G. Güçlü. 2013. "How to Write a Review Article?" *Turkish Journal of Urology* 39, no. Suppl 1: S44–S48. https://doi.org/10.5152/tud.2013.054.

Hair, J. F., W. C. Black, B. J. Babin, and R. E. Anderson. 2013. *Multivariate Data Analysis*. Pearson Education Limited.

Hair, J. F., M. C. Howard, and C. Nitzl. 2020. "Assessing Measurement Model Quality in PLS-SEM Using Confirmatory Composite Analysis." *Journal of Business Research* 109: 101–110. https://doi.org/10.1016/j.jbusres.2019.11.069.

Hair, J. F., J. J. Risher, M. Sarstedt, and C. M. Ringle. 2019. "When to Use and How to Report the Results of PLS-SEM." *European Business Review* 31, no. 1: 2–24. https://doi.org/10.1108/EBR-11-2018-0203.

Han, R., H. Zhou, J. Zhong, and C. Zhang. 2022. "Characterizing Peer Review Comments of Academic Articles in Multiple Rounds." *Proceedings of the Association for Information Science and Technology* 59, no. 1: 89–99.

Henseler, J., C. M. Ringle, and R. R. Sinkovics. 2009. "The Use of Partial Least Squares Path Modeling in International Marketing." In *New Challenges to International Marketing*, Advances in International Marketing, edited by R. R. Sinkovics and P. N. Ghauri, vol. 20, 277–319. Emerald Group Publishing Limited. https://doi.org/10.1108/S1474-7979(2009)0000020014.

Hochreiter, S., and J. Schmidhuber. 1997. "Long Short-Term Memory." *Neural Computation* 9, no. 8: 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735.

Huisman, J., and J. Smits. 2017. "Duration and Quality of the Peer Review Process: The Author's Perspective." *Scientometrics* 113, no. 1: 633–650. https://doi.org/10.1007/s11192-017-2310-5.

Johnson, R., A. Watkinson, and M. Mabe. 2018. "The STM Report an Overview of Scientific and Scholarly Publishing." https://www.stm-assoc.org/2018_10_04_STM_Report_2018.pdf.

Khraisha, Q., S. Put, J. Kappenberg, A. Warraitch, and K. Hadfield. 2024. "Can Large Language Models Replace Humans in Systematic Reviews? Evaluating GPT-4's Efficacy in Screening and Extracting Data From Peer-Reviewed and Grey Literature in Multiple Languages." *Research Synthesis Methods* 15: 616–626. https://doi.org/10.1002/jrsm.1715.

Kumar, A., T. Ghosal, S. Bhattacharjee, and A. Ekbal. 2024. "Towards Automated Meta-Review Generation via an NLP/ML Pipeline in Different Stages of the Scholarly Peer Review Process." *International Journal on Digital Libraries* 25: 493–504. https://doi.org/10.1007/s00799-023-00359-0.

Kumar, S., T. Ghosal, P. K. Bharti, and A. Ekbal. 2021. "Sharing Is Caring! Joint Multitask Learning Helps Aspect-Category Extraction and Sentiment Detection in Scientific Peer Reviews." In *2021 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, 270–273. IEEE. https://doi.org/10.1109/JCDL52503.2021.00081.

Leong, L. Y., T.-S. Hew, K.-B. Ooi, and A. Y. L. Chong. 2020. "Predicting the Antecedents of Trust in Social Commerce: A Hybrid Structural Equation Modeling With Neural Network Approach." *Journal of Business Research* 110: 24–40.

Li, Z., J. Huang, Z. Zhou, H. Zhang, S. Chang, and Z. Huang. 2016. "LSTM-Based Deep Learning Models for Answer Ranking." In *2016 IEEE First International Conference on Data Science in Cyberspace (DSC 2016)*, 90–97. IEEE. https://doi.org/10.1109/DSC.2016.37.

Marcoci, A., A. Vercammen, M. Bush, et al. 2022. "Reimagining Peer Review as an Expert Elicitation Process." *BMC Research Notes* 15, no. 1: 127. https://doi.org/10.1186/s13104-022-06016-0.

Qin, C., and C. Zhang. 2023. "Which Structure of Academic Articles Do Referees Pay More Attention to?: Perspective of Peer Review and Full-Text of Academic Articles." *Aslib Journal of Information Management* 75, no. 5: 884–916.

Ragone, A., K. Mirylenka, F. Casati, and M. Marchese. 2013. "On Peer Review in Computer Science: Analysis of Its Effectiveness and Suggestions for Improvement." *Scientometrics* 97, no. 2: 317–356. https://doi.org/10.1007/s11192-013-1002-z.

Richter, N. F., and A. A. Tudoran. 2024. "Elevating Theoretical Insight and Predictive Accuracy in Business Research: Combining PLS-SEM and Selected Machine Learning Algorithms." *Journal of Business Research* 173: 114453. https://doi.org/10.1016/j.jbusres.2023.114453.

Sanchez-Franco, M. J., G. Cepeda-Carrion, and G. Roldan. 2019. "Understanding Relationship Quality in Hospitality Services. A Study Based on Text Analytics and Partial Least Squares." *Internet Research* 29, no. 3: 478–503.

Shibayama, S., and J. Wang. 2020. "Measuring Originality in Science." *Scientometrics* 122, no. 1: 409–427. https://doi.org/10.1007/s11192-019-03263-0.

Stach, E., B. DeCost, A. G. Kusne, et al. 2021. "Autonomous Experimentation Systems for Materials Development: A Community Perspective." *Matter* 4, no. 9: 2702–2726. https://doi.org/10.1016/j.matt.2021.06.036.

Tomkins, A., M. Zhang, and W. D. Heavlin. 2017. "Reviewer Bias in Single- Versus Double-Blind Peer Review." *Proceedings of the National Academy of Sciences of the United States of America* 114, no. 48: 12708–12713. https://doi.org/10.1073/pnas.1707323114.

Tu, Y., C. Huang, Q. Wang, Y. Zhou, Z. Han, and Q. Huang. 2024. "Predicting Student Burnout in Blended Environments: A Complementary PLS-SEM and Machine Learning Approach."

14680394, 2025, 10, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/exsy.70131 by University Of Granada, Wiley Online Library on [27/10/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Interactive Learning Environments 33, no. 3: 2703–2717. https://doi.org/10.1080/10494820.2024.2415446.

Verma, R., K. Shinde, H. Arora, and T. Ghosal. 2021. "Attend to Your Review: A Deep Neural Network to Extract Aspects From Peer Reviews." In *Neural Information Processing*, edited by T. Mantoro, M. Lee, M. A. Ayu, K. W. Wong, and A. N. Hidayanto, 761–768. Springer International Publishing. https://doi.org/10.1007/978-3-030-92310-5_88.

Weber, R. 2021. "Constructs and Indicators: An Ontological Analysis." *MIS Quarterly* 45, no. 4: 1644–1678. https://doi.org/10.25300/MISQ/2021/15999.

Yuan, W., P. Liu, and G. Neubig. 2021. "Can We Automate Scientific Reviewing?" ArXiv. https://doi.org/10.48550/arXiv.2102.00176.

Zhou, H., R. Han, J. Zhong, and C. Zhang. 2024. "Which Review Aspect Has a Greater Impact on the Duration of Open Peer Review in Multiple Rounds?—Evidence From Nature Communications." *Aslib Journal of Information Management* 2024. https://doi.org/10.1108/AJIM-02-2024-0158.

Appendix A

In this appendix, we describe the PLS-SEM model's results for the reviewer reports of ICLR articles from 2018 to 2023[10] and review the different concepts used in it.

The PLS-SEM model is employed to analyse complex relationships between latent variables (constructs) and their observable indicators. In this context, the constructs and their respective indicators are defined as follows:

- Accessibility: Measured by clarity and soundness.
- Contribution: Measured by originality and motivation.
- Experimentation: Measured by substance and meaningful comparison.
- Decision: Value defined as in Equation (1).

Summary of the Structural Model

The structural model provides insights into the quality of the measurement and structural models. Key components include:

- Type: Indicates whether a construct is exogenous (not explained by other constructs in the model) or endogenous (explained by other constructs in the model).
- R^2 (coefficient of determination): Represents the proportion of variance explained for the endogenous construct.
- Block communality: Reflects the shared variance between the indicators and their construct.
- Average redundancy: Indicates the amount of variance in the endogenous construct explained by the exogenous constructs through their indicators.
- AVE (average variance extracted): Represents the average percentage of variance explained by the construct's indicators.

Acceptable values:

- R^2: Values above 0.67 are considered substantial; between 0.33 and 0.67, moderate; between 0.19 and 0.33, weak.
- Communality and AVE: Values above 0.50 are deemed acceptable, signifying that the construct explains more than half of its indicators' variance.

As shown in Table A1, the R^2 for decision is 0.6956, indicating that 69.56% of the variance in decision is explained by the exogenous constructs, which is considered moderate to high and acceptable.

Path Coefficients

The structural model reveals the direct relationships between constructs, including:

- Estimate: The path coefficient indicates the strength and direction of the relationship.
- Standard error: The standard deviation of the estimate.
- *t* value: A statistic for testing the significance of the coefficient.
- Significance: Indicates whether the coefficient is significantly different from 0.

Acceptable values:

- Path coefficients: Values between −1 and 1; coefficients closer to −1 or 1 indicate stronger relationships.
- *t* values: Absolute values greater than 1.96 (95% confidence level) indicate statistical significance.
- Significance: A *p* value less than 0.05 is commonly considered significant.

Table A2 shows that all relationships are statistically significant ($p < 0.001$) and positive, indicating that increases in the exogenous constructs are associated with increases in decision.

Cross-Loadings

Cross-loadings evaluate the discriminant validity of the indicators by measuring their association with their own construct versus others. Indicators should load more strongly on their construct than on others, with the difference being significant.

TABLE A2 | Path coefficients.

Relationship	Estimate	Standard error	*t*	Significance
EXP->Decision	0.2993	0.0075	40.0850	$p < 0.001$
CON->Decision	0.4384	0.0072	60.8976	$p < 0.001$
ACC->Decision	0.5264	0.0075	70.3547	$p < 0.001$

TABLE A1 | Structural model.

Construct	Type	R^2	Communality	AVE
Accessibility	Exogenous	0.0000	0.6112	0.6112
Contribution	Exogenous	0.0000	0.5380	0.5380
Experimentation	Exogenous	0.0000	0.6015	0.6015
Decision	Endogenous	0.6956	1.0000	N/A

14680394, 2025, 10, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/exsy.70131 by University Of Granada, Wiley Online Library on [27/10/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE A3 | Cross-loadings.

Indicator	Experimentation	Contribution	Accessibility
Clarity	0.1600	0.162	**0.7900**
Soundness	0.2744	0.1098	**0.7736**
Originality	0.0596	**0.8796**	0.0391
Motivation	−0.0020	**0.5498**	0.0980
Substance	**0.8988**	0.0345	0.2489
Meaningful comparison	**0.6285**	0.0479	0.1732

TABLE A4 | Unidimensional.

Construct	Composite reliability	AVE
Experimentation	0.7593	0.6015
Contribution	0.7039	0.5380
Accessibility	0.7588	0.6112
Decision	N/A	N/A

From Table A3, the highest loadings (in bold) confirm that each indicator is more strongly associated with its respective construct, meeting the criteria for discriminant validity.

Unidimensionality (Reliability and Validity)

Unidimensionality is assessed using Dillon–Goldstein's rho (composite reliability), which is more appropriate than Cronbach's alpha in PLS-SEM and average variance extracted (AVE) to evaluate convergent validity.

Conclusion

The general model, based on data from reviews of ICLR articles from 2018 to 2023, shows that all constructs are significantly related to Decision, with an R^2 of 0.6956. This demonstrates the robustness of the model in explaining the relationships between accessibility, contribution, experimentation and decision.

- Accessibility reflects clarity or soundness, as indicated by high factor loadings (>0.75).
- Contribution reflects originality or motivation, supported by strong path coefficients between contribution and decision (>0.45).
- Substance and meaningful comparisons describe experimentation, validated by high loadings (>0.7).

Table A4 indicates that composite reliability values exceed 0.7 for all constructs, confirming reliability and AVE values above 0.5 validate convergent validity.

GoF

The GoF index is a global measure of model fit, calculated as the square root of the product of the average communality and the average R^2.

Acceptable values:

- Small: 0.1
- Medium: 0.25
- Large: 0.36 or higher

For the general model, the GoF value is 0.6371, indicating a large and satisfactory model fit.

14680394, 2025, 10, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/exsy.70131 by University Of Granada, Wiley Online Library on [27/10/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License