HOSTED BY

ELSEVIER

Contents lists available at ScienceDirect

Journal of King Saud University - Computer and Information Sciences

journal homepage: www.sciencedirect.com

Journal of King Saud University – Computer and Information Sciences

Full Length Article

Content-based quality evaluation of scientific papers using coarse feature and knowledge entity network[☆]

Zhongyi Wang[a], Haoxuan Zhang[a], Haihua Chen[b,*], Yunhe Feng[c], Junhua Ding[b]

[a] School of Information Management, Central China Normal University, Wuhan, 430079, China
[b] Department of Information Science, University of North Texas, Denton, 76203, TX, USA
[c] Department of Computer Science and Engineering, University of North Texas, Denton, 76203, TX, USA

ARTICLE INFO

Keywords:
Paper quality evaluation
Knowledge entity
Network analysis
Machine learning
Novelty
Structural entropy

ABSTRACT

Pre-evaluating scientific paper quality aids in alleviating peer review pressure and fostering scientific advancement. Although prior studies have identified numerous quality-related features, their effectiveness and representativeness of paper content remain to be comprehensively investigated. Addressing this issue, we propose a content-based interpretable method for pre-evaluating the quality of scientific papers. Firstly, we define quality attributes of computer science (CS) papers as *integrity*, *clarity*, *novelty*, and *significance*, based on peer review criteria from 11 top-tier CS conferences. We formulate the problem as two classification tasks: *Accepted/Disputed/Rejected* (ADR) and *Accepted/Rejected* (AR). Subsequently, we construct fine-grained features from metadata and knowledge entity networks, including text structure, readability, references, citations, semantic novelty, and network structure. We empirically evaluate our method using the ICLR paper dataset, achieving optimal performance with the Random Forest model, yielding F1 scores of 0.715 and 0.762 for the two tasks, respectively. Through feature analysis and case studies employing SHAP interpretable methods, we demonstrate that the proposed features enhance the performance of machine learning models in scientific paper quality evaluation, offering interpretable evidence for model decisions.

1. Introduction

High-quality research is the engine of scientific and technological progress. Many countries have elevated the identification and management of high-quality research to the national level. In addition to primary funding, many countries conduct large-scale expert assessments of the quality of research and researchers, such as the Research Excellence Framework (REF) in the United Kingdom (Wilsdon, 2016), the Performance-Based Research Fund (PBRF) in New Zealand (Buckle and Creedy, 2019) and Italy (Franceschini and Maisano, 2017), and the Excellence Program in Australia (Hinze et al., 2019). Assessing the quality of papers is often a subjective and time-consuming task (Lin et al., 2023). Peer review is a critical way to evaluate the quality of papers, and it is considered the gatekeeper of publications (Marsh and Bazeley, 1999). However, peer review inevitably has limitations. Differences among reviewing experts in professional knowledge, reviewing environments and emotions, and conflicts of interest can affect the consistency of review results (Lin et al., 2023). Worse still, the proliferation of submissions has become an enormous burden on the effective operation of peer review. The process not only consumes the authors' academic time but also fails to adequately reward reviewers for their efforts (Huisman and Smits, 2017).

Nowadays, many researchers have begun to explore and design various approaches to identify and measure the quality of research. The quality of research is usually reflected in different aspects, such as innovation (Uzzi et al., 2013; Wang et al., 2022), novelty (Wu et al., 2019; Xu et al., 2021; Luo et al., 2022; Hou et al., 2022), impact (Abrishami and Aliakbary, 2019; Hu et al., 2020; Xu et al.,

[☆] This study is supported by National Social Science Foundation of China (Grant Number 22BTQ102).

[*] Corresponding author.

E-mail addresses: wzywzy13579@163.com (Z. Wang), zhanghaox@mails.ccnu.edu.cn (H. Zhang), Haihua.chen@unt.edu (H. Chen), yunhe.feng@unt.edu (Y. Feng), junhua.ding@unt.edu (J. Ding).

Peer review under responsibility of King Saud University.

ELSEVIER Production and hosting by Elsevier

https://doi.org/10.1016/j.jksuci.2024.102119
Received 4 March 2024; Received in revised form 10 June 2024; Accepted 4 July 2024
Available online 9 July 2024

2022), and readability (Vincent-Lamarre and Larivière, 2021; Ante, 2022). The combination of these different attributes constitutes the overall quality of a paper. Some scholars have focused on open review in pursuing a more comprehensive and quantifiable standard for evaluating research quality. Using the corresponding review scores of publicly available papers as a quality criterion achieves a more comprehensive and fairer assessment of the overall quality of a paper (Kang et al., 2018). Researchers are able to analyze the features of papers receiving high review scores to assess the quality of scientific papers. With the advancement of deep learning and natural language processing technologies, researchers have applied them to scientific paper quality assessment. They have used neural networks or pre-trained language models to represent the text of a paper with accurate semantic information (Yang et al., 2018; Wenniger et al., 2020; Xue et al., 2023).

However, existing research on assessing the quality of scientific papers suffers from two main shortcomings. Firstly, there is a lack of interpretability. Evaluating the quality of scientific papers is a high-stakes task, and the predictions made by models must be accompanied by corresponding evidence. On the other hand, when constructing quality-related features, the focus should be on the content of the paper (Sun et al., 2022), evaluating the value of the knowledge contributed by the paper, rather than merely relying on external metadata. Secondly, the data must be accessible beforehand. Assessing the quality of scientific papers is also a highly time-sensitive task. Data such as citation counts (Thelwall et al., 2023b) and peer review texts (Ghosal et al., 2019) are not available during the peer review stage, and models built using these data are severely limited in value. To achieve this goal, we formulate the following research questions (RQs):

- RQ1: What are the criteria and attributes for the quality evaluation of scientific papers?
- RQ2: What can be used to realistically represent a scientific paper and quantify the predefined criteria and attributes for quality evaluation?
- RQ3: What approaches are effective for evaluating the quality of scientific papers based on their content?

For RQ1, we compiled review guidelines from 11 prominent conferences in computer science, distilling four core attributes – integrity, clarity, novelty, and significance — to gauge paper quality.

For RQ2, scientific papers encapsulate a rich tapestry of valuable knowledge, offering a nuanced reflection of essential content (Zhang et al., 2020). Existing quality metrics, primarily reliant on reference combinations, often neglect this granular knowledge, so lack semantic depth. Methods focusing on topic and term extraction from titles and abstracts overlook the semantic interplay between paper content and referenced knowledge. In contrast, extracting knowledge entities from the full text and constructing co-occurrence networks prove advantageous, revealing characteristics in knowledge generation, utilization, and evolution (Liang et al., 2021). We propose a content representation using a three-level network for each paper – at the paper level, related field level, and paper-related field alignment level – enabling a comprehensive capture of diverse knowledge domains within the article.

For RQ3, assessing the quality of scientific papers based on content requires methods that are comprehensive and interpretable. We fully consider the metadata of scientific papers and the knowledge entity network, constructing features supported by corresponding evidence for integrity, clarity, novelty, and significance of quality. These features encompass various aspects such as text structure, references, citations, readability, network structure characteristics, semantic novelty, and more. In feature analysis, we utilize the SHAP interpretable machine learning method to demonstrate how different features influence the model's decisions.

To validate the effectiveness of the proposed method, we conducted an empirical evaluation based on papers from the International Conference on Learning Representations (ICLR) in 2023. We defined the assessment of scientific paper quality as a binary classification task: *Accepted/Disputed/Rejected* (ADR) and *Accepted/Rejected* (AR). We compared several machine learning models and pre-trained language models. The results indicate that the Random Forest model performed exceptionally well on both classification tasks, with F1 scores of 0.715 and 0.762, respectively. Additionally, we conducted feature analysis and case studies, demonstrating that integrity, clarity, novelty, and significance are all crucial features affecting paper quality. We also summarized the fine-grained feature distributions of papers of different qualities.

The rest of the paper is organized as follows: In Section 2, we review research evaluating the scientific paper's quality. Section 3 details the methodology and research design. In Section 4, we perform an empirical analysis. Section 5 further discusses the results and implications of the study. Section 6 summarizes this paper and outlines future work. Our datasets, code for reproducing the methods, and additional experimental results can be accessed on GitHub with the following link: https://github.com/haihua0913/QualityEval4Papers. The abbreviations and acronyms used in this article are listed in Table A.1.

2. Related work

This section presents an overview of the current literature on the identification of key features and the assessment of quality in high-quality scientific papers.

2.1. The identification of the features of high-quality scientific papers

For high-quality research, numerous scholars have focused on measuring attributes, such as novelty, innovation, impact, and readability, in the quality of papers. These refinements help us understand the connotations of quality in scientific papers and explore more interpretable and pre-accessible features.

In novelty evaluation, the novelty of scientific papers can be seen as an atypical reorganization of different knowledge. Uzzi et al. (2013) analyzed 17.9 million papers by examining the combinations of references in the reference lists and their co-citation frequencies and found that the most influential sciences tended to add novelty to the work of their predecessors. Amplayo et al. (2018) designed both a macro-graph consisting of authors and papers, and a micrograph composed of keywords, topics, and words to detect the novelty of papers, and found that the keyword-topic-word novelty model is more suitable for novelty detection. Luo et al. (2022) regarded the papers as question-method term pairs and used the BERT model to compute semantic similarity between the term pairs to characterize the novelty of the papers. Hou et al. (2022) considered academic papers as different combinations of research questions, research methods, and research results and utilized the Sentence-BERT model to compute the novelty of the knowledge links constituted by the different combinations.

In innovation evaluation, innovations can be classified into two main categories: incremental consolidation and disruptive breakthroughs (Li and Chen, 2022). By counting the citation relationships among focal literature, references to focal literature, and citations to focal literature, Wu et al. (2019) proposed the D index for measuring breakthrough innovation research. Wang et al. (2023) incorporated knowledge entities into calculating the D index, which reveals breakthrough innovations from the perspective of fine-grained knowledge content. In patented technology, Chen et al. (2021) introduced a novel perspective by redefining instability and consolidation as two dimensions of technology. They proposed that highly innovative technology can be characterized as dual technology, emphasizing its ability to consolidate existing technologies and disrupt and destabilize other technologies. From a micro perspective, the degree of research innovation can be reflected by the degree of innovation of core knowledge elements. For instance, Wang et al. (2022) utilized a combination of biterm topic modeling (BTM) and cloud modeling methods to calculate the degree of innovation of knowledge elements in the method category.

Citation count can be used to assess the potential impact of a paper. Abrishami and Aliakbary (2019) used artificial neural networks to predict long-term citations based on the early citation counts. Zhao and Feng (2022) constructed citation networks to predict citation count from the perspective of information cascade. Huang et al. (2022), conversely, combined the structural functions of papers with citation count prediction. On the other hand, the impact is also reflected in changes in the structure of knowledge networks. The emergence of novel topics changes the topology of the network structure (Zhang et al., 2021; Min et al., 2021). Xu et al. (2022) monitored the changes in the structural entropy of the knowledge network of scientific topics to find the tipping point in genetically engineered vaccines.

2.2. *The methods of evaluating quality of scientific papers*

2.2.1. *Evaluation methods based on constructed features*

To construct effective features to quantify and predict decisions in peer review, Kang et al. (2018) proposed two natural language processing (NLP) tasks: paper acceptance classification and review aspect prediction. Paper acceptance classification is a binary classification task used to predict whether a manuscript will be accepted to a conference. Review aspect prediction is a multiclass regression task used to predict the scores of different aspects of a paper to determine the quality of the manuscript. In the first task, Kang et al. (2018) designed two types of features: coarse and lexical. The former includes whether the manuscript contains appendices, title length, number of authors, and references. The latter includes features such as a bag of words extracted from the abstract and TF-IDF weights. In the second task, they performed a regression analysis of eight aspects of the manuscript with the help of expert review texts in OpenReview. Given that some conferences publicly release review comments, which can provide additional features for predictive models, many scholars have conducted sentiment analyses of review comments to explore whether positive reviewer emotions can reflect the quality score of a paper (Wang and Wan, 2018; Ghosal et al., 2019; Ribeiro et al., 2021). Although sentiment analysis of review comments achieved excellent results in the two NLP tasks mentioned above, this post hoc information is ineffective in relieving the pressure on reviewers.

On the other hand, bibliometrics and other indicators can help distinguish the quality of papers, including citation counts, authors' average citation rates, journal impact factors, and so on Thelwall (2022), Thelwall et al. (2023b). High-quality research often receives more citations, and Kousha and Thelwall (2023) systematically reviewed factors that affect research quality or citations, including textual structural properties, journal properties, team size properties, and subject/field properties. Considering the potential bias of relying on citation data alone, Lin and Qilin (2020) argued that sentiment polarity analysis combined with citations can help to differentiate paper quality. However, Xu et al. (2023) found that while high-quality papers received more positive citations, the frequency of negative citations was also significantly higher than general papers. They attributed this phenomenon to the fact that high-quality articles receive more attention, incentivizing more papers to improve their deficiencies. Similarly, the above indicators are still ex-post with a time lag and must be accumulated over time. By considering the citation text in the body of the paper, Basuki and Tsuchiya (2022) argued that the distribution of citation function features reflects the paper's position in the relevant literature and is helpful in evaluating the paper's quality.

2.2.2. *Evaluation methods based on deep learning*

In contrast to the manually constructed features mentioned above, deep learning-based models automatically extract semantic features from papers. Yang et al. (2018) proposed a modular hierarchical convolutional neural network to achieve paper rating by the deep representation of the full-text content. Since training recurrent neural networks on long texts such as papers can lose important information about text structure, Wenniger et al. (2020) combined hierarchical attention networks with structural labels to improve paper quality prediction tasks. However, many papers do not support open access and the full paper can contain more redundant and noisy information. Xue et al. (2023) proposed the DGC-BERT model, which uses a dual-view convolutional network to enhance the in-depth representation of the paper's title and abstract.

In the field of deep learning for text generation, generating review comments for papers is a more direct way to evaluate their quality. Wang et al. (2020) generated review comments for target papers by constructing knowledge graphs for the target papers and their related fields. However, their experimental results have not been widely validated, and only 50 papers were manually evaluated for the constructive and effective review comments generated. Yuan et al. (2022) constructed the ASAP-Review dataset for generating review comments on different aspects of papers and trained the BART model to generate review comments. Extensive experimental results showed that compared with human-written reviews, the reviews generated by their model could summarize the core ideas of the paper more comprehensively. However, there still needs to be improvement in the constructive and accurate nature of the feedback.

In summary, we highlight the following gaps: (1) Clarity in quality standards. There is a need for clearly defined and subdivided quality standards for evaluating scientific papers. While peer review decisions or scores offer comprehensive assessments, specific quality properties, such as innovation, novelty, impact, and readability, need explicit definition and quantification. This finer granularity can enhance the precision of scientific paper quality assessment based on overall peer review results. (2) Content-focused predictive features. Features for predicting scientific paper quality should center on the content dimension. Relying solely on metadata or bibliometric information neglects the substance of scientific content, potentially introducing bias or randomness and leading to unfair evaluations. (3) Interpretability and pre-accessibility of features. Features used for predicting paper quality should be interpretable and accessible beforehand. Peer review is a high-stakes scenario; only interpretable features can guide review experts. Since peer review is a pre-evaluation process, post-information like citation counts, journal impact factors, and sentiment in review comments cannot be utilized.

3. Methodology

This section first gives a formal definition of two different tasks for the quality evaluation of scientific papers. We then introduce the proposed framework, which is shown in Fig. 1. The framework consists of three main steps: (1) establishing criteria for scientific paper quality assessment, (2) constructing feature representations for machine learning models, and (3) machine learning model development and evaluation, which will be described in Sections 3.2, 3.3, and 3.4, respectively.

3.1. *Problem definition*

Evaluating the quality of scientific papers can be defined as a classification task. For conference papers, the determination made by the area chair serves as the benchmark for classifying paper quality. Area chairs, who are recognized experts in their fields, utilize anonymous review comments and their extensive expertise to categorize a paper as accepted or rejected. Acceptance signifies a high level of

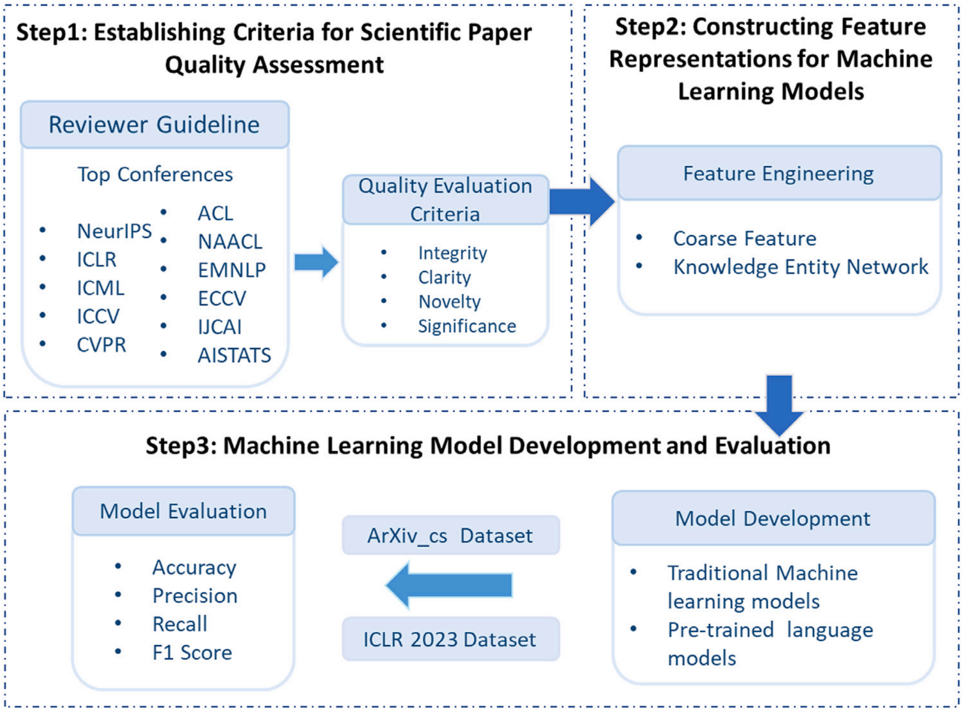

Step1: Establishing Criteria for Scientific Paper Quality Assessment
Reviewer Guideline
Top Conferences
NeurIPS
ICLR
ICML
ICCV
CVPR
ACL
NAACL
EMNLP
ECCV
IJCAI
AISTATS
Quality Evaluation Criteria
Integrity
Clarity
Novelty
Significance
Step2: Constructing Feature Representations for Machine Learning Models
Feature Engineering
Coarse Feature
Knowledge Entity Network
Step3: Machine Learning Model Development and Evaluation
Model Evaluation
Accuracy
Precision
Recall
F1 Score
ArXiv_cs Dataset
ICLR 2023 Dataset
Model Development
Traditional Machine learning models
Pre-trained language models

Fig. 1. The overall architecture for evaluating the quality of scientific papers.

quality, while rejection implies a deficiency in this aspect. However, we have observed instances of misalignment between review scores and decision categories. Some papers with high scores face rejection, while others with low scores get accepted, indicating a divergence in quality assessment between anonymous reviewers and area chairs, as shown in Fig. A.1.

Therefore, we bifurcate the task of evaluating paper quality into two classification tasks: the *Accepted/Disputed/Rejected (ADR) task* and the *Accepted/Rejected (AR) task*. We will assess paper quality for both classification tasks concurrently in subsequent experiments. The formal mathematical formulation of the two classification tasks is as follows:

For the ADR task:

$$y = f(x_1, x_2, \ldots, x_n) \tag{1}$$

where y represents the paper category, which can be Accepted, Disputed or Rejected. $x_1, x_2, \ldots, x_n$ represent the input crude features and knowledge entity network features. $f(\cdot)$ represents the classification model.

For the AR task:

$$y' = g(x_1, x_2, \ldots, x_n) \tag{2}$$

where y' represents the paper category, which can be Accepted or Rejected. $g(\cdot)$ represents another classification model, which may be different from the one used in the ADR task.

3.2. Establishing criteria for scientific paper quality assessment

To evaluate the quality of scientific papers, it is essential to clarify the definition of quality and its attributes. The peer review process, divided into pre-publication and post-publication stages (Spezi et al., 2018; Checco et al., 2021), plays a pivotal role in this evaluation. In the pre-publication stage, papers undergo initial screening for plagiarism, formatting, scope, and presentation. Reviewers then assess the papers based on criteria, such as Novelty/Originality, Importance/Significance, Scope/Relevance, and Soundness/Rigor. Automated review systems also follow a two-stage process (Lin et al., 2023).

Table 1
The statistics on the different criteria mentioned in the review guidelines.

Top conferences	Criteria mentioned in the review guidelines						
	Area1	Area2	Area3	Area4	Area5	Area6	Area7
NeurIPS	✓	✓	✓	✓	✓	✓	
ICLR	✓	✓	✓	✓	✓	✓	✓
ICML	✓	✓	✓	✓		✓	
ICCV		✓	✓	✓		✓	✓
CVPR		✓	✓	✓		✓	
ACL/NAACL/EMNLP	✓	✓	✓	✓	✓	✓	✓
ECCV		✓	✓	✓		✓	
IJCAI	✓	✓	✓	✓	✓	✓	
AISTATS	✓	✓	✓	✓	✓	✓	✓

Note: Area1 refers to “Related Work/Relationship to Previous Work”, Area2 refers to “Strengths and Weaknesses/Contribution”, Area3 refers to “Innovation/Novelty/Originality”, Area4 refers to “Rigor/Rationality/Repeatability/Interpretability/Scalability”, Area5 refers to “Readability/Clarity/Organization of Writing”, Area6 refers to “Importance/Application Prospects/Impact/Significance”, and Area7 refers to “Ethical Concerns/Ethical Review”. The statistical date is as of July 10, 2023.

As the field of computer science is experiencing rapid expansion and a substantial influx of submissions, there is an increasing demand for a more efficient peer review process. Consequently, we have placed our emphasis on articles within the realm of computer science. In this particular field, conference papers serve as the predominant format for scholarly works, thus prompting us to examine the quality criteria delineated in the latest review guidelines of esteemed conferences. Through manual analysis of criteria and literature review (Yuan et al., 2022; Kousha and Thelwall, 2023; Shi et al., 2024), we identified four main criteria for evaluating computer science papers: integrity, clarity, novelty, and significance (see Tables 1 and 2). In the subsequent sections, these criteria will be the foundation for our paper quality evaluation.

Table 2
The criteria for evaluating the quality of scientific papers.

Evaluation criteria	Description
Integrity	The proposed methodology, technique, or justification should be comprehensive, detailed, precise, reproducible, and take full account of existing research.
Clarity	The quality of language (word spelling, grammar) are up to standard, and the writing style and organization of paper's sections are in line with academic norms.
Novelty	The scientific paper is an innovative way to advance science by recombining existing knowledge or proposing new knowledge (including methods, theories, discoveries, results, etc.) in an unprecedented way.
Significance	The scientific paper changes or complements an existing body of knowledge and has an impact on the field of study or social practice.

3.3. Constructing feature representations for machine learning models

3.3.1. Coarse feature and knowledge entity network construction

(1) **Coarse feature**

When assessing the quality of scientific papers, we can extract useful metadata from the paper itself to construct coarse features. These features typically include the textual structure, such as charts, formulas, and word count; the readability of the abstract text; the citation intent of the cited content; and the quantity, recency, and citation frequency of the references in the bibliography. These features can not only be obtained in advance, but also play a crucial role in identifying the quality of the paper (Kang et al., 2018; Checco et al., 2021; Kousha and Thelwall, 2023).

(2) **Knowledge entity network construction**

Knowledge entities are fundamental elements of scientific papers (Zhang et al., 2020). These entities are categorized into macro entities (e.g., authors, journals, papers), meso entities (e.g., keywords), and micro entities (e.g., datasets, methods, domain entities) (Ding et al., 2013). We focus on micro-knowledge entities as they are essential components of scientific papers.

In the field of artificial intelligence, we adopt the entity labeling scheme introduced by the SciERC dataset (Luan et al., 2018), encompassing entity types like Task, Method, Metric, Material, Other-Scientific Term, and Generic. We employ the SciNERTopic model (Roman Jurowetzki, 2022) to extract knowledge entities. This model combined Sentence Transformers and BERTopic and was fine-tuned on the SciERC dataset, enabling efficient extraction of knowledge entities. The extracted entities form the basis for constructing knowledge entity co-occurrence networks, providing a more detailed representation of content than previous approaches relying on references, keywords, or topics (Liang et al., 2021; Xu et al., 2022). These networks reveal specific reference knowledge and association strengths through co-occurrence relationships.

We design a three-level network of knowledge entity co-occurrence, comprising the paper, related field, and paper-related field alignment levels. Precisely, the paper-level network ($Network_p = (K_p, E_p)$) consists of knowledge entities at the paper level (K_p) and their co-occurrence relationships (E_p), with edge weights indicating co-occurrence frequency. The related field level network ($Network_r = (K_r, E_r)$) encompasses literature in the related field containing entities from the paper (K_r) and their co-occurrence relationships (E_r), with edge weights representing co-occurrence frequency. The paper-related field alignment level network is defined as $Network_a = (K_a, E_a)$, where K_a is obtained by aligning K_p and K_r, and E_a is the union of E_p and E_r. Utilizing these three network levels enables a nuanced assessment of novelty and significance within knowledge entity networks.

3.3.2. Integrity features construction

We first design integrity features based on the citation intent of the cited content. By analyzing the distribution of citation intents, we can gain insights into how scientific papers build upon and improve upon existing research (Lin and Sui, 2020; Basuki and Tsuchiya, 2022). Specifically, following the classification in Lauscher et al. (2021), we categorize the citation intents in the paper into seven types: background, difference, extends, future work, motivation, similarity, and uses. Background citations can explain the research problem's context and related work. Difference citations can highlight the study's innovations and contrast it with prior work. Extends citations can demonstrate the research's continuity. Future work citations can indicate the study's limitations and improvement directions. Motivation citations can elucidate the rationale and justification for the current study. Similarity citations can supplement the argumentation by referencing analogous studies, reflecting the study's coherence. Uses citations introduce the methods, theories, or models adopted in the research.

In addition, the references are also a reflection of the paper's integrity, including the quantity and recency of the references. These features indicate whether the scientific paper is built upon a comprehensive foundation of existing research, whether it keeps up with the latest research progress, and whether it is at the forefront of the field. Furthermore, for conference papers, the presence of appendices is often an important feature in assessing the paper's quality (Kousha and Thelwall, 2023).

3.3.3. Clarity features construction

One way to construct clarity features is to measure the readability of the scientific paper. Various readability indices, such as the Flesch–Kincaid Grade Level, SMOG Index, Coleman-Liau Index, Automated Readability Index (ARI), Linsear Write Formula, and Gunning Fog Index (Kousha and Thelwall, 2023), are commonly used to calculate the readability of the text. These indices consider factors like sentence length, word length, complex words, and syllable counts. There has been extensive empirical research conducted by many scholars exploring the relationship between readability and the quality of scientific papers (Ante, 2022; Vincent-Lamarre and Larivière, 2021). While the findings may vary across different domains, these readability measures can provide valuable insights into the clarity of the paper's quality (Lu et al., 2019). By incorporating these readability features, we can assess the extent to which the scientific paper is presented in a clear, concise, and easy-to-understand manner. This can be an important indicator of the paper's overall quality and its potential impact on the target audience.

Another aspect of constructing clarity features is to consider the textual structure of the paper, including the presence of charts, formulas, and word count. To effectively convey the research process, present the findings, and provide thorough argumentation, authors need to include appropriate visual aids, mathematical expressions, and textual explanations within the paper (Kousha and Thelwall, 2023). The inclusion of well-designed charts, informative formulas, and a suitable word count can reflect the author's efforts to enhance the clarity and comprehensibility of the scientific paper. These structural elements can help readers better understand the concepts, methods, and conclusions presented in the work.

3.3.4. Novelty features construction

We leverage knowledge entity networks to measure the novelty of scientific papers. Specifically, we first assign weights to the entity pairs based on their co-occurrence frequency in the paper, giving higher weights to core knowledge entities and reducing the influence of redundant information. Next, we calculate the semantic similarity between the knowledge entity pairs present in the paper and those representing the domain-level knowledge. This allows us to ultimately derive a novelty score for the paper. Beyond the numerical novelty score, the number of new nodes and edges in the paper's knowledge network can also reflect its novelty. The presence of new knowledge entities and novel connections between them suggests that the paper introduces new concepts or reorganizes existing knowledge in innovative ways. By analyzing the characteristics of the knowledge entity network constructed from the paper, we can gain insights into the degree of novelty and originality of the research work (Hou et al., 2022; Luo et al., 2022). This provides a quantifiable way to assess the paper's contribution to advancing the state-of-the-art in the field.

When calculating the semantic similarity, we employ the SPECTER pre-trained language model, which is suitable for computing semantic similarity in the context of scientific papers. SPECTER (Cohan et al., 2020) is based on the SciBERT model and further trained on a corpus that includes citation relationships. Importantly, the word embeddings generated by SPECTER can be used directly for semantic similarity calculations without the need for additional fine-tuning.

The formula for calculating the novelty of knowledge entity pairs is expressed as follows:

$$Novelty(Entity\_pairs_i) = 1 - \frac{\sum_{j=1}^{n} Sim(Entity\_pairs_i, Entity\_pairs_j)}{n} \tag{3}$$

In Eq. (3), $Novelty(Entity\_pairs_i)$ represents the novelty of the ith entity pair, $Sim()$ represents the cosine similarity between the word embeddings of the computed entity pairs, and n represents the number of knowledge entity pairs in the related field level network.

Subsequently, the novelty of the paper is calculated in two steps. Firstly, we compute the weights of each entity pair in the paper. Then, we determine the novelty of all entity pairs in the paper based on their respective weights, summing them to obtain the overall novelty of the paper. The calculation process is depicted in the following equations:

$$Weight(Entity\_pairs_i) = \frac{Co(Entity\_pairs_i)}{\sum_{j=1}^{m} Co(Entity\_pairs_j)} \tag{4}$$

$$Novelty(Paper_i) = \sum_{j=1}^{m} Weight(Entity\_pairs_j) \times Novelty(Entity\_pairs_j) \tag{5}$$

In the above equation, $Weight(Entity\_pairs_i)$ denotes the weight of the ith entity pair, $Co(Entity\_pairs_i)$ denotes the frequency of co-occurrence of entity pairs, m denotes the number of entity pairs in the paper, and $Novelty(Paper_i)$ denotes the novelty of the ith paper.

3.3.5. Significance features construction

The significance of a scientific paper can be measured by the changes in its knowledge entity network. In the dynamic landscape of scientific development, different knowledge domains interweave and integrate, forming a complex network. To quantify the dynamic changes in the structure of this complex network, researchers have proposed metrics such as degree structure entropy, betweenness structure entropy, and structure entropy ratio (Xu et al., 2022; Zhang and Li, 2022).

Degree structure entropy calculates the distribution of node degrees within the network. The more uniform the degree distribution, the higher the degree structure entropy, and the more complex the network structure. Betweenness structure entropy measures the centrality of the nodes, and the more uniform the betweenness distribution, the more complex the network structure. The structure entropy ratio is the ratio of the two, which reflects the relative complexity of the degree and betweenness distributions in the network. It incorporates both local and global information, and can identify cases of similar distributions, providing a more comprehensive measure of network complexity. By comparing the changes in structure entropy between the paper-level network and the paper-related field alignment level network, we can quantify the influence of the new knowledge introduced in the scientific paper. This allows us to assess the significance of the paper within the broader context of the research field. The schematic diagram of this process is illustrated in Fig. 2:

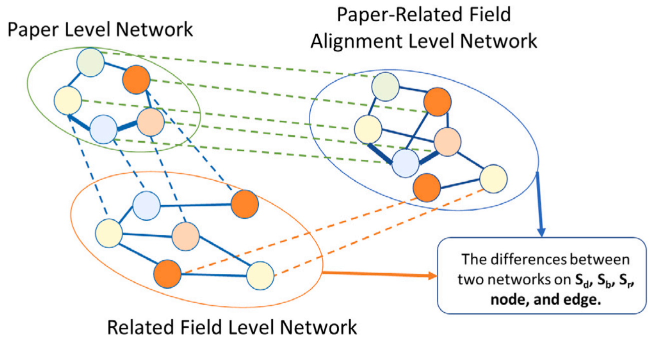

Paper Level Network
Paper-Related Field Alignment Level Network
The differences between two networks on S_d, S_b, S_r, node, and edge.
Related Field Level Network

Fig. 2. Quantifying the significance of the paper based on knowledge entity networks. Different colored circles represent different knowledge entities, dotted lines represent the same knowledge entities, solid lines represent the existence of co-occurrence between two knowledge entities and the thickness of the solid line reflects the degree of co-occurrence.

The formulas for degree structural entropy (S_d), betweenness structural entropy (S_b), and structural entropy ratio (S_r) are shown below:

$$S_d = -\sum_{i=1}^{n} \frac{d_i}{n} \log_2 \frac{d_i}{n} \tag{6}$$

$$S_b = -\sum_{i=1}^{n} \frac{b_i}{n} \log_2 \frac{b_i}{n} \tag{7}$$

$$S_r = \frac{S_d - S_b}{S_d} \tag{8}$$

The above formulas use S_d, S_b, and S_r to represent degree structural entropy, betweenness structural entropy, and structural entropy ratio, respectively. Here, n represents the number of knowledge entities, and d_i and b_i represent the degree and betweenness of an entity, respectively. Here are the formulas for degree (d_i) and betweenness (b_i) of a knowledge entity:

$$d_i = \sum_{j=1}^{n} A_{ij} \tag{9}$$

where A_{ij} represents the adjacency matrix element between entity i and entity j.

$$b_i = \sum_{j \neq i \neq k} \frac{\sigma_{jk}(i)}{\sigma_{jk}} \tag{10}$$

where σ_{jk} is the total number of shortest paths from entity j to entity k, and $\sigma_{jk}(i)$ is the number of those paths that pass through entity i.

The intrinsic network characteristics of the scientific paper itself can also reflect its quality. Papers with groundbreaking innovations tend to exhibit unique network topological structures (Min et al., 2021). Therefore, we can employ network analysis metrics to evaluate the structural features of the paper's knowledge entity network, including: number of nodes, number of edges, average degree, network density, average clustering coefficient, maximum betweenness centrality, maximum closeness centrality, maximum eigenvector centrality and number of connected components. These metrics can provide insights into the complexity, connectivity, and centrality of the knowledge entities and their relationships within the paper's network. The detailed calculation formulas for these network measures can be found in the work by Min et al. (2021).

The references cited in the paper can also help determine its significance. The higher the average citation count of the references, the more significant the research topic addressed in the paper, and the potentially higher the quality of the paper (Kousha and Thelwall, 2023). Therefore, we also analyze the citation counts of the references included in the paper. This metric provides further insights into the prominence and impact of the prior work that the paper builds upon.

In summary, we have constructed a set of features to assess the quality of scientific papers, including integrity, clarity, novelty, and importance. These features are derived from the paper's metadata and knowledge entity network, and are supported by empirical evidence from related studies, while being readily available a priori. For detailed explanations of each feature, please refer to Appendix A.3.

3.4. Machine learning model development and evaluation

In scientific paper quality evaluation, conventional machine learning models have traditionally been utilized to predict paper quality based on manually crafted features, often offering a certain level of interpretability (Kang et al., 2018; Thelwall, 2022). Noteworthy machine learning models include support vector machines (SVM), decision trees (DT), random forests (RF), k-nearest neighbors (KNN), extreme gradient boosting (XGBoost), gradient boosting for classifiers (GBC), and adaptive boosting for classifiers (ABC), among others.

The advent of deep learning algorithms has witnessed remarkable performance in text representation, with pre-trained language models gaining significant traction in quality assessment studies. A prominent pre-trained language model, BERT (Devlin et al., 2018), undergoes extensive data pre-training, capturing global word and sentence dependencies through a bidirectional self-attention mechanism. It can be fine-tuned for specialized tasks with limited data. Building on BERT's foundation, subsequent models trained on scientific literature, such as SciBERT (Beltagy et al., 2019) and SPECTER (Cohan et al., 2020), as well as models with expanded parameter sizes and training data, like RoBERTa (Liu et al., 2019), have emerged. However, models like BERT have token limitations, requiring chunking paper content (van Dongen et al., 2020) or inputting only the title and abstract data (Xue et al., 2023). Diverging from traditional machine learning models, deep learning models, exemplified by pre-trained language models, can automatically extract features from scientific papers without manual feature engineering. Nonetheless, this characteristic also poses challenges to model interpretability. This study will compare traditional machine learning models with pre-trained language models in the classification task of assessing the quality of scientific papers.

4. Empirical evaluation

In this section, we describe the datasets, statistical analysis of the networks, experimental setting, evaluation metrics, and results of the experiments.

4.1. Datasets and gold standard

We collected 4922 papers submitted to ICLR 2023. Following the screening, 1098 articles were excluded due to formatting issues, scope mismatches, or preprints. The remaining 3824 papers were parsed into XML format using the GROBID[1] tool. Out of these, 14 papers encountered parsing problems and were subsequently excluded. Ultimately, we obtained a dataset of 3810 papers, comprising 1567 accepted and 2243 rejected papers.

To develop the gold standard for evaluating paper quality, we considered the necessity of establishing consensus among reviewers and area chairs. To achieve this, we utilized the confidence scores provided by anonymous reviewers in their reviews. These confidence scores were used to adjust the impact weighting of the review results, assigning less weight to reviews with lower confidence levels. The goal was to minimize the influence of unconfident reviews on the final assessment of paper quality. For more details on combining paper quality scores, self-confidence levels, and their division, please refer to Appendix A.2. Descriptive information for the experimental data is provided in Table 3.

Table 3
The descriptive information for the ICLR 2023 dataset.

	Number	Mean	Std	Min	Max
Accepted paper	1409	6.37	0.74	5.22	9.18
Disputed paper	737	5.44	0.47	3.43	7.07
Rejected paper	1664	4.14	0.77	1	5.21

Table 4
The top 10 most frequent pairs of entities and their entity types in the arXiv_cs network.

Entity1 (Type)	Entity2 (Type)	Number of co-occurrence
Deep neural network (Method)	Neural network (Method)	606
Artificial intelligence (Task)	Machine learning (Method)	593
Deep learning (Method)	Deep neural network (Method)	523
Edge (Other Scientific Term)	Graph (Other Scientific Term)	508
Deep learning (Method)	Machine learning (Method)	473
Deep learning (Method)	Deep learning model (Method)	449
Graph (Other Scientific Term)	Graph neural network (Method)	395
Computer vision (Task)	Deep learning (Method)	391
Adversarial attack (Other Scientific Term)	Robustness (Metric)	386

4.2. The statistical analysis of networks

We used the SciNERTopic model to extract knowledge entities from arXiv computer science domain literature and ICLR 2023 papers, respectively, and constructed the corresponding co-occurrence networks. Finally, the two networks' entity pairs were 10,391,013 and 6,535,432, respectively. The distribution of knowledge entity types for the two networks is shown in Fig. 3.

From Fig. 3, we observe that the key knowledge entities in the arXiv_cs dataset are methods, other scientific terms, and task types, while the proportion of metrics and material entities is relatively small, yet each category contains over 1,500,000 entities. This suggests that the arXiv_cs dataset contains a wider and more diverse range of knowledge entities in terms of both quantity and type. Table 4 lists the top 10 pairs of knowledge entities with the highest co-occurrence rate in the arXiv_cs network. This indicates that the arXiv_cs network is primarily focused on areas such as machine learning, deep learning, and graph neural networks, which is consistent with the themes of the ICLR 2023 conference. Considering the quantity and distribution of knowledge entity types, as well as their alignment with the relevant research domains, the knowledge entity network constructed from the arXiv_cs dataset appears to be a suitable knowledge source.

4.3. Experimental setup

We experimented with two classification tasks (as introduced previously) on the ICLR 2023 dataset using the gold standard. The experimental environment was as follows: Python 3.8 as the programming language, scikit-learn[2] for machine learning models, Windows 11 as the operating system, a 12th generation Intel Core i7-12700K CPU, an NVIDIA GeForce RTX 3070Ti GPU, and PyTorch version 1.10.

For machine learning models, we employed a range of commonly used and well-performing classifiers, including the linear model SVM, tree-based models DT and RF, distance-based model KNN, and ensemble models XGB, GBC, and ABC. These models have been widely applied in the task of paper quality assessment (Thelwall, 2022; Basuki and Tsuchiya, 2022; Kousha and Thelwall, 2023; Thelwall et al., 2023b).

[1] https://grobid.readthedocs.io/en/latest/Introduction/.

[2] https://scikit-learn.org/stable/index.html.

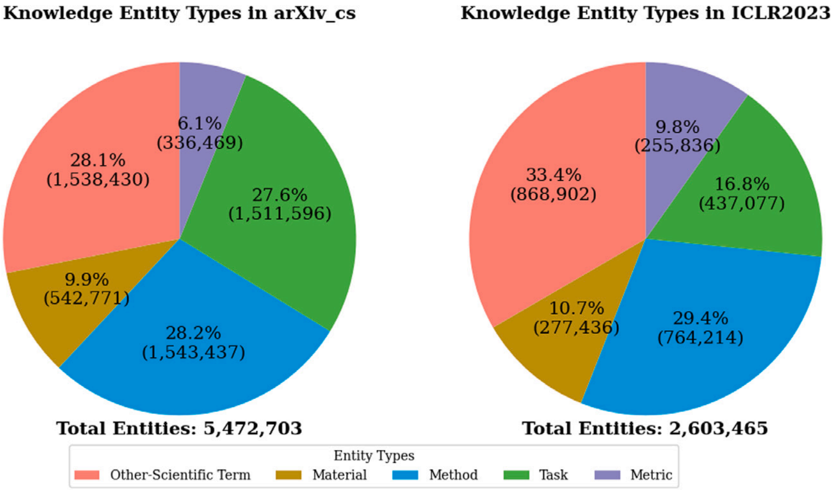


Fig. 3. The distribution of knowledge entity types for the two networks.

Regarding the hyperparameter settings, we utilized grid search to determine the optimal hyperparameters for each classifier. To address the class imbalance in the dataset, we applied the SMOTE algorithm to oversample the training set. Additionally, we performed mean and standard deviation normalization to standardize the features.

For pre-trained language models, we selected BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019), SciBERT (Beltagy et al., 2019), SPECTER (Cohan et al., 2020), SPECTER2, THExt-cs-sciBERT (La Quatra and Cagliero, 2022), and CS_RoBERTa[3] (Singh et al., 2022). These models have been widely used in text classification tasks, and the SciBERT, SPECTER, SPECTER2, THExt-cs-sciBERT, and CS_RoBERTa models have been further trained on scientific text, making them more suitable for tasks in the scientific domain. We fine-tuned these models on the title and abstract text from the ICLR2023 dataset to predict the quality labels of the papers. For the fine-tuning hyperparameters, we set the batch size to 64, employed the Adam optimizer for parameter optimization, and set the learning rate to 2e-5, the maximum sequence length to 512, and the dropout rate to 0.3 to address potential overfitting issues.

Additionally, we employed several tools in the construction of the quality features. We utilized a state-of-the-art deep learning model trained on the Multicite dataset (Lauscher et al., 2021) for citation function classification. The citation counts of the references were retrieved using the Semantic Scholar API (Kinney et al., 2023), with a data collection deadline of July 10, 2023. Readability indices for the titles and abstracts were calculated using the TextSTAT tool. In order to obtain a relevant literature corpus, we compiled a collection of 255,297 computer science papers from arXiv, including papers from previous ICLR conferences. To ensure the relevance and currency of the corpus, we restricted the timeframe to papers published between January 1, 2019, and August 31, 2022. Papers published beyond September 2022 were not included, as the ICLR 2023 conference began accepting submissions in that month.

4.4. *Evaluation metrics*

We used accuracy, precision, recall, and F1 score as the evaluation metrics for the classification models since they are the most commonly used metrics. In addition, we used two different metric calculation methods, weighted and binary, for the ADR and AR tasks, respectively, to minimize possible biases in the evaluation results.

4.5. *Experimental results and analysis*

4.5.1. *Overall results*

The overall results of the experiments are presented in Table 5, which illustrates the best performance of various models for the ADR and AR tasks. Fig. 4 depicts the confusion matrices of the RF model on the test set for the ADR task (left) and the AR task (right), respectively. Due to the input length limitation of PLMs, we only fine-tuned the models on the title and abstract of the papers. This can effectively avoid the redundant information in the full text and improve the model performance (Xue et al., 2023).

From the results, we make the following observations:

- Model performance comparison: The RF model achieved the best performance on both the ADR and AR tasks, followed by the XGB model. This suggests that tree-based methods are effective for the paper quality classification task, as they can capture the complex nonlinear relationships between various features and paper quality levels. The F1 scores of the SVM, RF, XGB, and GBC models all exceeded 0.7, with the RF model reaching 0.762, indicating that the constructed machine learning models can effectively differentiate paper quality and have potential for practical application.
- ML vs. PLMs: The traditional ML models significantly outperformed the PLMs on both the ADR and AR tasks. This suggests that the feature engineering using metadata and knowledge entity networks was successful, while the general language models may suffer from overfitting or insufficient generalization ability. Furthermore, PLMs lack interpretability in the classification process, making it unclear whether the classification is based on semantic similarity or if the models have truly learned the features indicating paper quality.
- ADR vs. AR: The three-class ADR task is more complex than the two-class AR task, and the performance metrics are generally higher for the AR task across different models. The three-class problem requires the model to learn a more complex decision boundary to distinguish between *Accepted*, *Disputed*, and *Rejected* paper quality categories. Additionally, in the real world, *Disputed* papers are those that experts find difficult to determine, and the assessment of their quality should be more cautious.

To further investigate the impact of fine-tuning PLMs on titles and abstracts for enhancing paper quality assessment, we employed the SPECTER2 model. SPECTER2 has demonstrated superior performance on the aforementioned tasks. We encoded the title and abstract text by extracting the [CLS] token output from each text. The [CLS] token,

[3] https://huggingface.co/allenai/cs_roberta_base.

Table 5
The best results of different models on the ADR and AR tasks.

	Model	Accuracy		Precision		Recall		F1 score	
		ADR	AR	ADR	AR	ADR	AR	ADR	AR
ML	SVM	0.588	0.704	0.594	0.679	0.588	0.792	0.589	0.731
	DT	0.540	0.643	0.538	0.642	0.540	0.674	0.539	0.657
	RF	**0.714**	**0.746**	**0.719**	**0.729**	**0.714**	**0.796**	**0.715**	**0.762**
	KNN	0.531	0.653	0.550	0.625	0.531	0.794	0.512	0.699
	XGB	0.678	0.734	0.685	0.726	0.678	0.766	0.680	0.745
	GBC	0.601	0.713	0.612	0.695	0.601	0.777	0.602	0.733
	ABC	0.521	0.675	0.527	0.667	0.521	0.720	0.522	0.693
PLMs	BERT	0.401	0.529	0.405	0.450	0.426	0.559	0.405	0.499
	RoBERTa	0.411	0.540	0.414	0.464	0.440	0.562	**0.416**	0.508
	SciBERT	0.394	**0.573**	0.395	**0.484**	0.400	0.515	0.397	0.499
	SPECTER	0.409	0.557	0.411	0.482	0.416	0.528	0.413	0.504
	SPECTER2	**0.417**	**0.573**	**0.420**	0.462	**0.443**	**0.622**	0.393	**0.530**
	THExt-cs-sciBERT	0.397	0.529	0.399	0.450	0.410	0.538	0.397	0.490
	CS_RoBERTa	0.404	0.557	0.406	0.483	0.426	0.431	0.399	0.456
SPECTER2 + ML	SVM	**0.560**	**0.668**	0.453	**0.584**	**0.559**	0.586	0.499	**0.585**
	DT	0.433	0.593	0.433	0.490	0.433	0.497	0.432	0.494
	RF	0.534	0.627	0.428	0.550	0.534	0.365	0.473	0.439
	KNN	0.465	0.574	0.447	0.467	0.465	0.487	0.451	0.477
	XGB	0.545	0.650	0.496	0.561	0.545	0.563	0.502	0.562
	GBC	0.556	0.659	**0.536**	0.574	0.556	0.563	**0.511**	0.568
	ABC	0.524	0.656	0.493	0.564	0.524	**0.605**	0.496	0.584

Note: Values in bold represent the best results, and underlining represents the second-best results.

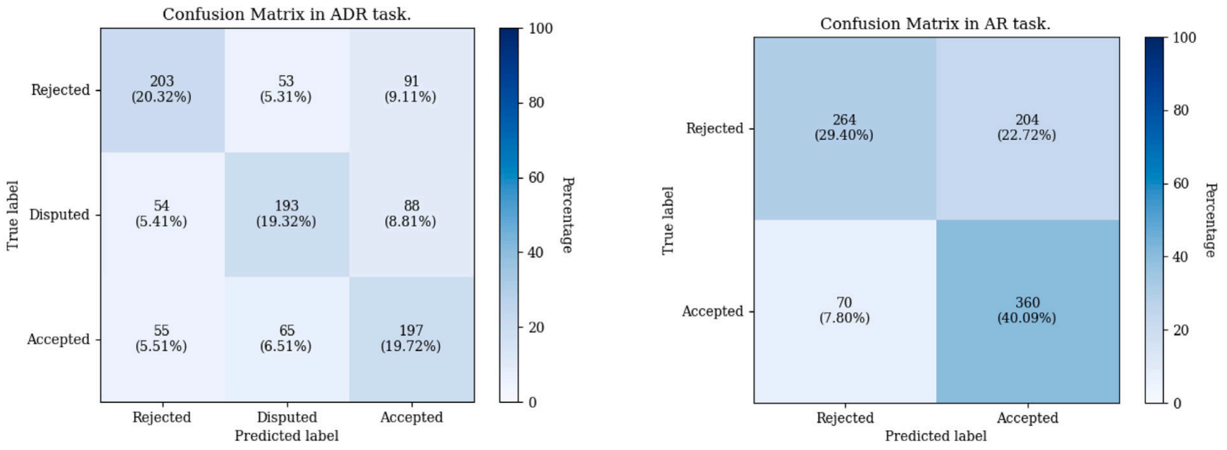


Fig. 4. The confusion matrix for the test set on the classification tasks.

inserted at the beginning of each text segment in the input sequence, captures the contextual information of the entire sequence within SPECTER2's internal representation (Gomes et al., 2023).

We then used these embedding feature vectors, along with other features, as input to ML classifiers for evaluation. Our results indicated that the performance of the SPECTER2+ML model improved compared to the model fine-tuned on titles and abstracts. However, it still did not surpass the performance of the ML model without text embeddings. This suggests that the semantic information in titles and abstracts may not fully capture the quality characteristics of a paper. Instead, titles and abstracts might reflect the classification of the research field more than the paper's quality.

4.5.2. *Feature analysis*

We conducted feature analysis to identify the features that effectively influence the evaluation of the quality of scientific papers. We eliminated the features of each dimension sequentially. Table 6 showed the performance changes of the RF model on the AR task, compared to the original best results, when different types of features were removed one by one.

The feature analysis results indicate that *Integrity*, *Clarity*, *Novelty*, and *Significance* are all important for scientific paper quality. *Integrity* had the greatest impact on the RF model's AR task performance, with its removal causing a significant decline. This suggests *Integrity* is the most crucial factor in evaluating paper quality for the AR task, followed by *Clarity*, *Novelty*, and *Significance*. Scientific articles should prioritize integrity and clarity to ensure effective communication of ideas and drive scientific progress. Novelty and Significance represent the value of scientific articles. High-quality papers should contribute novel and significant knowledge.

To further investigate the impact of each fine-grained feature on model performance, we conducted an interpretability analysis using the SHAP method on the RF model, as shown in Fig. 5. The SHAP method (Lundberg and Lee, 2017), derived from the Shapley value in game theory, assigns a value to each feature that represents its contribution to the prediction. The sum of all feature contributions equals the final prediction, making the model's predictions interpretable.

Observing Fig. 5, we can draw the following conclusions:

Table 6
Feature analysis results of the RF model on the AR task.

	Accuracy	Precision	Recall	F1 score
All features	**0.746**	**0.729**	**0.796**	**0.762**
w/o Integrity	0.699 **(−0.047)**	0.660 **(−0.069)**	0.767 **(−0.029)**	0.710 **(−0.052)**
w/o Clarity	0.705 (−0.041)	0.664 (−0.065)	0.777 (−0.020)	0.716 (−0.046)
w/o Novelty	0.722 (−0.024)	0.679 (−0.050)	0.793 (−0.003)	0.732 (−0.030)
w/o Significance	0.725 (−0.021)	0.683 (−0.046)	0.793 (−0.003)	0.734 (−0.027)

Note: w/o indicated that the features under this dimension were eliminated in the model training. Values in parentheses indicate differences from complete performance.

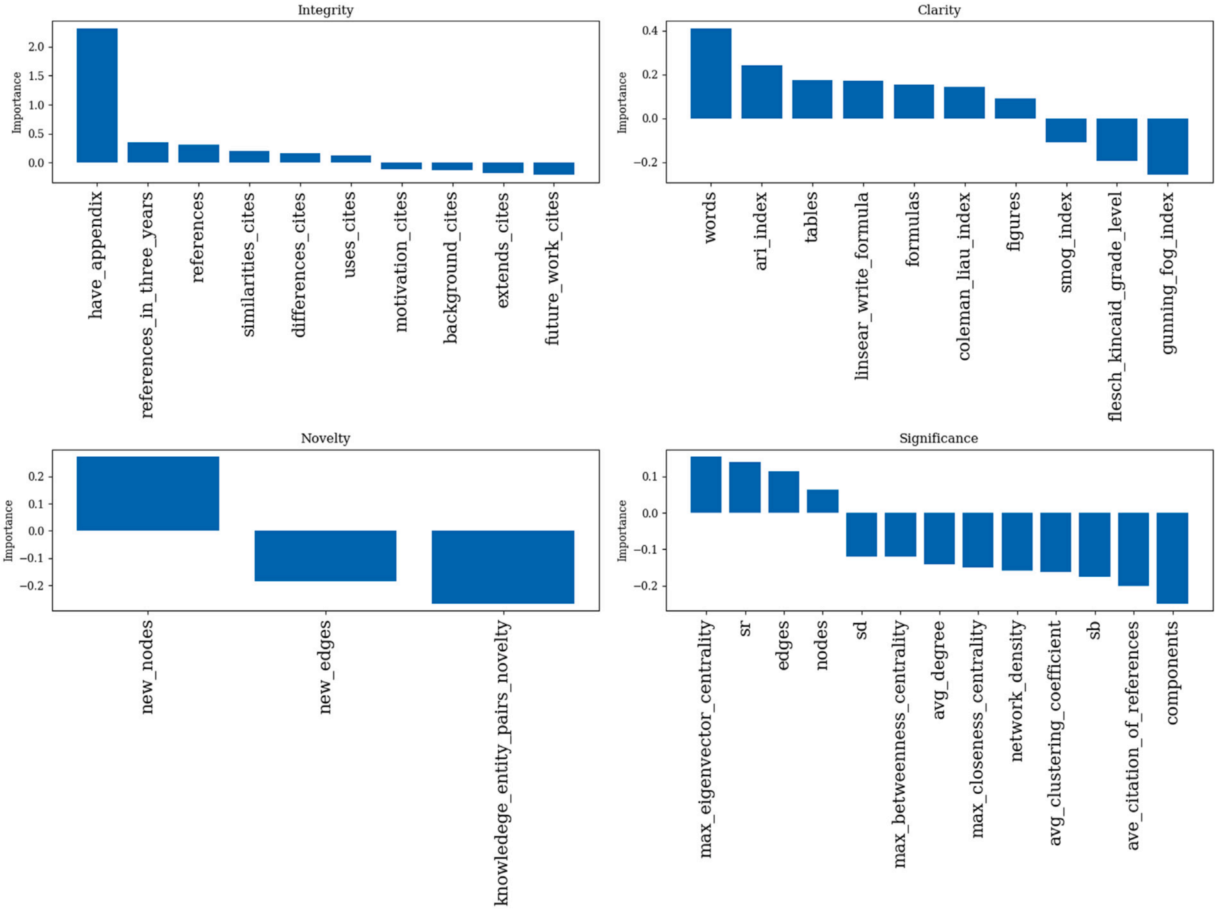


Fig. 5. SHAP-based feature importance analysis for paper quality evaluation. Appendix A.3 describes the specific meaning of each feature.

- For integrity, the inclusion of appendices, a higher proportion of recent references, and a greater number of references that engage with and compare relevant studies can positively contribute to the paper's integrity. Conversely, an overreliance on heuristic, background, extended, and future work citations may indicate a lack of the paper's own theoretical and methodological framework, compromising its independence and systematicity, and consequently impacting its integrity.
- For clarity, a greater number of words, tables, figures, and formulas in the main text can more intuitively present the research results, enhancing the paper's comprehensibility. However, the readability indices demonstrate varied influences. Metrics such as ARI index, Linsear Write formula, and Coleman-Liau index measure lexical difficulty and sentence complexity, reflecting the paper's technicality. Conversely, SMOG index, Flesch–Kincaid Grade level, and Gunning Fog index assess the colloquialism and readability of the language. The opposing effects of these two aspects suggest that simplicity does not necessarily equate to high quality. In fact, papers with greater technicality and complexity often represent the author's deeper understanding and insights in the field.
- For novelty, new nodes represent new knowledge, and the introduction of more new nodes suggests that the paper covers more novel perspectives and ideas, which can enhance its innovativeness. However, the novelty of new edges and knowledge entity pairs exhibits a negative influence, which may indicate that the paper's knowledge recombination has not reached the expected effect, or the connections with existing knowledge are insufficient, failing to fully absorb and integrate previous research, and consequently impacting the calculated novelty.
- For significance, a high eigenvector centrality indicates that the paper has core knowledge entities, while a larger number of connected components suggests a broader range of knowledge covered, implying that high-quality papers should have a certain

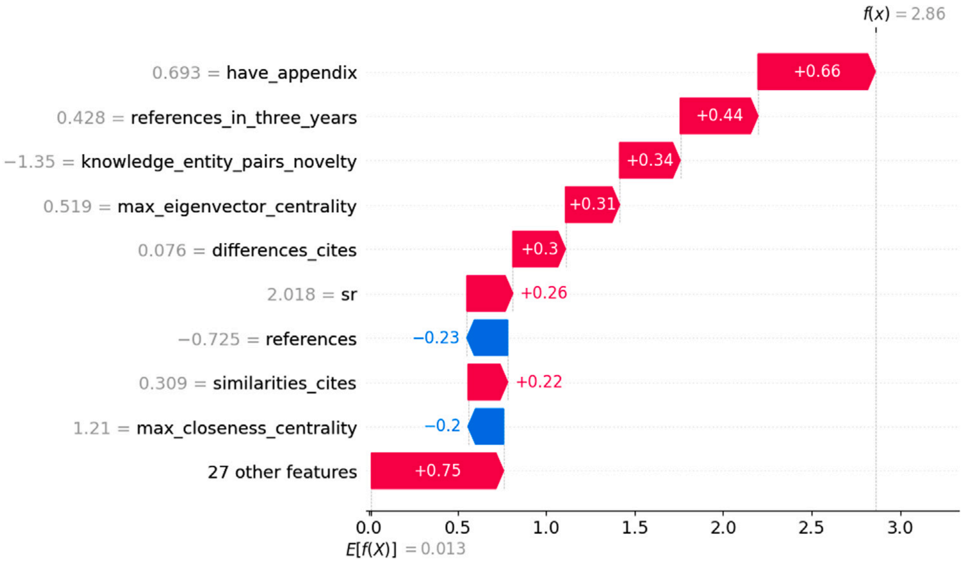


Fig. 6. SHAP feature importance analysis for case 1.

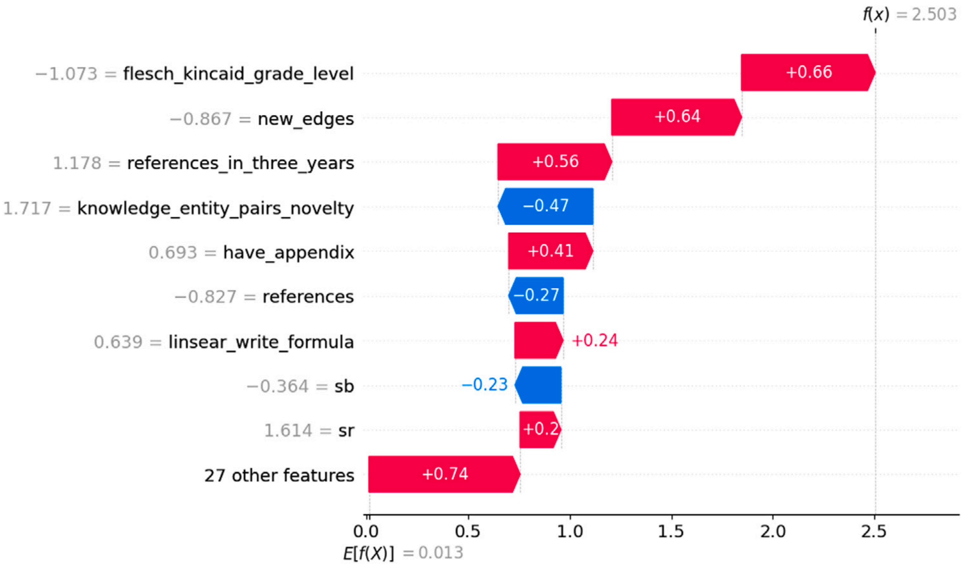


Fig. 7. SHAP feature importance analysis for case 2.

degree of focus. Regarding structural entropy, larger degree and betweenness entropy have negative impacts, as they reflect divergence from the domain's knowledge structure. Conversely, a larger structural entropy ratio implies that although the overall difference is large, the difference in degree distribution dominates, while the difference in key nodes is relatively small. This suggests that the paper may have absorbed and integrated mainstream knowledge while also supplementing and developing some new knowledge structures, forming a certain degree of innovation.

4.5.3. *Case study*

To compare how scientific papers are assessed based on various features, we randomly selected two high-quality articles (top 5%) and two low-quality ones (rejected). Table 7 outlines the details. Figs. 6, 7, 8, and 9 present the SHAP-based interpretability analysis for four case studies. $E[f(X)]$ represents the classification threshold, while $f(x)$ denotes the model's calculated result. Please note that the values in the figure have been standardized.

Table 7
Four randomly selected papers from ICLR 2023.

	Title	Paper decision
Case 1	Targeted Hyperparameter Optimization with Lexicographic Preferences Over Multiple Objectives	Accept: top5%
Case 2	Emergent World Representations: Exploring a Sequence Model Trained on a Synthetic Task	Accept: top5%
Case 3	Music-to-Text Synesthesia: Generating Descriptive Text from Music Recordings	Reject
Case 4	Policy Architectures for Compositional Generalization in Control	Reject

The first paper focused on multi-objective hyperparameter optimization, garnering strong support for its innovative approach using lexicographic preferences. The second paper explored whether transformers learned reasonable world-state representations, earning acclaim for its significant findings. The third paper on music-to-text synesthesia faced criticism for terminology use but lack of recognized related work. The

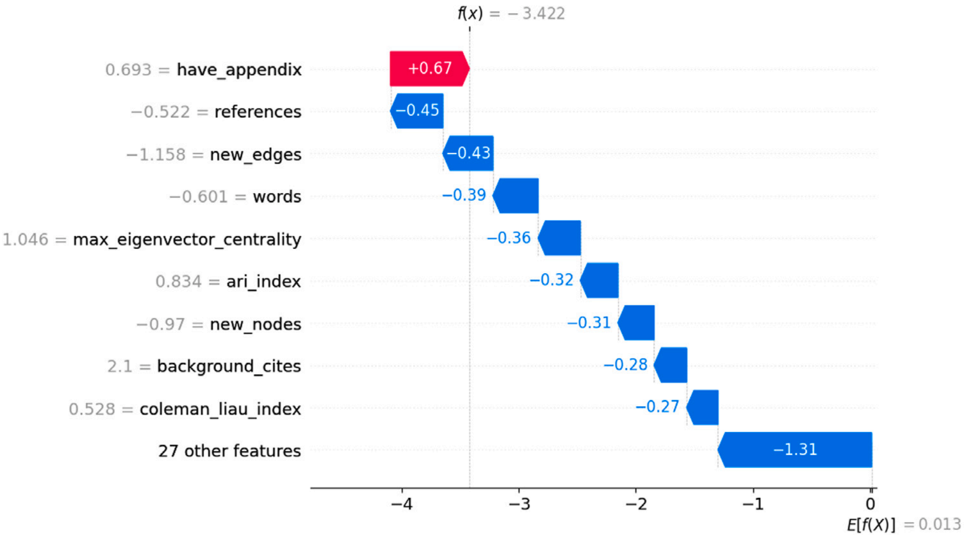


Fig. 8. SHAP feature importance analysis for case 3.

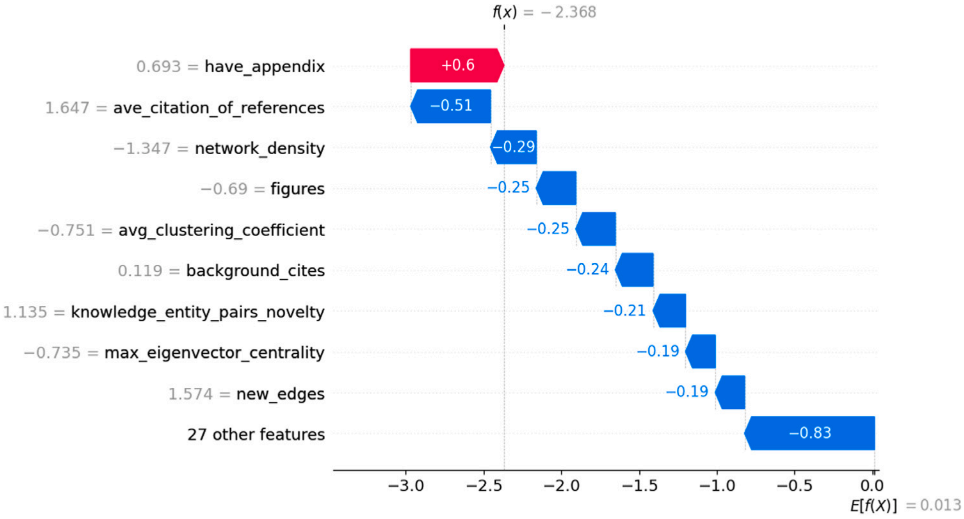


Fig. 9. SHAP feature importance analysis for case 4.

fourth paper proposed Entity-Factored Markov Decision Processes, but reviewers found similar results in existing work. From the four cases, we can see that the contribution of different features varies across each case. Multiple features interact and ultimately drive the movement of the result along the horizontal axis. This means that, in evaluating paper quality, while features like the presence of an Appendix play a significant role, the final judgment must consider a comprehensive assessment of multiple features.

5. Vision on future considerations and implication

5.1. Future considerations

Quantifying and predicting the quality of scientific papers drives scientific and technological advancements. In this context, we identify three crucial directions for future research in assessing scientific paper quality.

Firstly, quality standards for scientific papers should derive from human experts. While citation count is a common proxy for quality, it has limitations (Thelwall et al., 2023a). Highly cited articles may not necessarily reflect true quality, and some low-cited papers might be undervalued. Expert peer review, grounded in reviewers' expertise, provides a more objective assessment (Sun et al., 2022). Relying solely on citation count makes it challenging to discern specific contributions within a paper. Expert reviews reveal a paper's core contributions and diverse aspects of quality. Thus, human expert assessments should serve as the quality standard. Platforms like OpenReview and F1000 Research offer transparent review processes, storing valuable data for quality assessment across disciplines.

Secondly, the quality of scientific papers can be evaluated through knowledge networks. Papers contribute to and reshape existing knowledge networks, influencing subsequent network changes (Min et al., 2021; Xu et al., 2022). Constructing knowledge networks with semantic relationships from full-text papers, rather than co-occurrence-based networks, provides richer information (Ma et al., 2023). Feature extraction methods, such as graph embedding algorithms (Xue et al., 2023) and metrics like structural entropy (Xu et al., 2022) and the entity-based disruption index (Wang et al., 2023), enable meaningful feature extraction from knowledge networks. These strategies facilitate the measurement of scientific paper quality, offering insights into a paper's impact on domain knowledge.

Thirdly, text-generation techniques can be employed to evaluate scientific paper quality. Unlike classification tasks, using models like BART and T5, text generation can generate direct review feedback (Wang et al., 2020; Yuan et al., 2022). Although challenging, large language models like GPT-4 have shown promise in generating feedback (Liang et al., 2023). However, limitations arise due to a lack of specific knowledge. Enhancing feedback specificity and actionability requires providing detailed knowledge, including existing and novel domain knowledge. Combining knowledge networks with large language models can improve the quality of generated feedback, enabling more insightful evaluations of scientific papers.

5.2. Implication

Our research contributes in two main ways to the evaluation of scientific paper quality. Firstly, we segment the quality of scientific papers into four dimensions: integrity, clarity, novelty, and significance. We derive coarse features and network features from both the metadata of papers and the knowledge entity network. These features not only enhance the effectiveness of machine learning models in identifying paper quality but also offer a degree of interpretability. Integrity is demonstrated by the adequacy of a paper's references and citations, conveying a comprehensive understanding of scientific knowledge. Clarity is manifested in the paper's readability, conveying the author's ideas and research content through an appropriate number of figures, tables and formulas. Novelty is reflected in the paper's contribution to existing knowledge domains. Significance is gauged by the extent to which the paper's contribution reshapes existing knowledge domains.

Secondly, we emphasize the predictive value of fully exploiting features that are accessible beforehand. Our approach holds practical value as it does not require the accumulation of post-data such as citations, thus saving time during the review process. Researchers can utilize the time saved to expedite scientific output (Huisman and Smits, 2017). Additionally, our method holds implications for government agencies and technology policymakers. Early allocation of research funds and management of high-quality research outcomes is a complex task that necessitates the accurate identification and targeted support (Abramo and D'Angelo, 2020; Xu et al., 2021). By identifying high-quality research outcomes based on pre-existing data, decision-makers can make informed decisions and proactively steer the direction of future technological development.

In summary, this research will benefit all stakeholders in the peer review process and support scientific research advancement. The pre-evaluation approach and enhanced interpretability make our study valuable for evaluating the quality of scientific papers, aiding decision-making, and fostering progress in academia and scientific knowledge.

6. Conclusion

Considering the explosion of the number of scientific papers produced nowadays, automatically pre-measuring the quality of these papers can not only alleviate the pressure on peer reviewers but also maintain the fairness of scientific evaluation. However, developing an effective and explainable algorithm for the quality evaluation of scientific articles is a non-trivial task. Therefore, this study proposed a content-based quality evaluation framework for scientific papers using coarse features and knowledge entity networks. The quality evaluation criteria (i.e., integrity, clarity, novelty, and significance) were summarized from the peer review guidelines of 11 top conferences in computer science. We utilize the metadata of scientific papers and the knowledge entity network that can be accessed beforehand to construct corresponding features for the four quality aspects. We conducted an empirical evaluation on two different classification tasks on the ICLR 2023 dataset; the experiments demonstrate the effectiveness of our proposed framework, and the content-based features also provide excellent model interpretability. However, the accuracy of the proposed method might be affected by the limitation of existing tools for parsing PDFs and extracting knowledge entities. Another limitation of this study is that we did not consider semantic relationships among different knowledge entities when constructing the knowledge entity networks.

In the future, we will extend our framework to other domains by constructing discipline-specific features for evaluating and predicting paper quality. Considering the impressive capabilities of large language models (e.g., GPT-3 and GPT-4 from OpenAI, LLaMA from Meta, and PaLM2 from Google), particularly in text generation tasks within the realm of natural language processing, we will also investigate the potential integration of text generation techniques into the assessment of scientific paper quality.

CRediT authorship contribution statement

Zhongyi Wang: Writing – review & editing, Supervision, Resources, Methodology, Funding acquisition, Conceptualization. **Haoxuan Zhang:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Haihua Chen:** Writing – review & editing, Supervision, Project administration, Methodology, Formal analysis, Conceptualization. **Yunhe Feng:** Writing – review & editing, Methodology, Conceptualization. **Junhua Ding:** Writing – review & editing, Methodology, Conceptualization.

Acknowledgements

This study is supported by National Social Science Foundation of China (Grant Number 22BTQ102). The authors are are grateful to the editor and all the anonymous reviewers for their precious comments and suggestions.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix

A.1. Abbreviations and acronyms used in this article

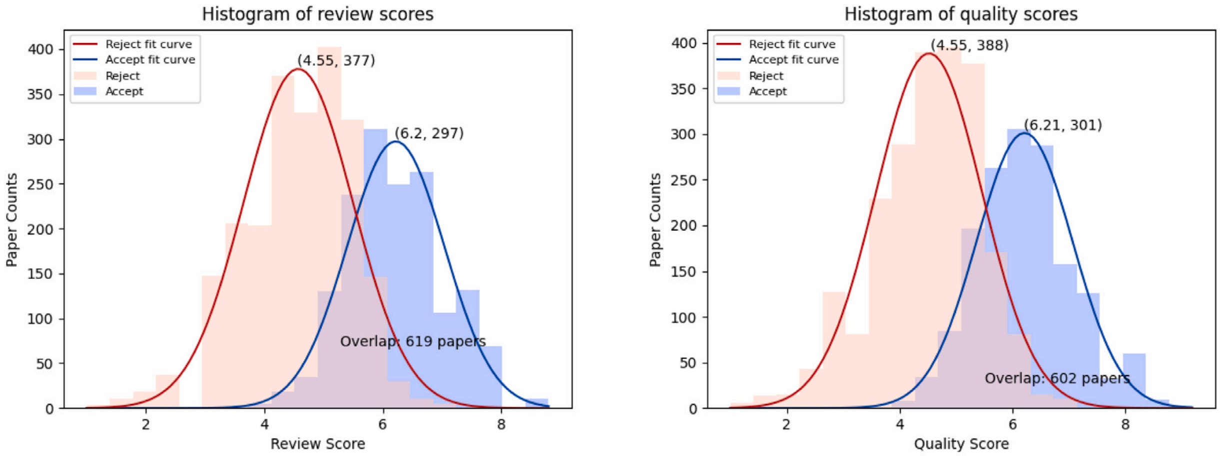
The abbreviations and acronyms used in this article are listed in Table A.1.

A.2. The calculation process of the comprehensive quality score

We designed a comprehensive quality score for each paper by combining the review score and the confidence score of each reviewing expert. Specifically, (1) we calculated recommendation weights based on the confidence scores for each expert. (2) Then, we calculated the confidence-weighted review score. (3) Finally, we derived the overall quality score of the paper by averaging all the weighted review scores. The formula for calculating the recommendation weights is shown in Eq. (A.1):

$$Weight_i = \frac{Confidence_i}{\sum_{j=1}^{n} Confidence_j} \tag{A.1}$$

In Eq. (A.1), $Weight_i$ represents the recommendation weight of the ith expert, $Confidence_i$ represents the confidence level of the ith expert, and n represents the number of experts participating in the review of



(a) The average review score

(b) The comprehensive quality score

Fig. A.1. The distribution of the two scores related to whether the paper was accepted or rejected.

Table A.1
Abbreviations and acronyms used in this article.

Abbreviation	Full name
ADR	Accepted/Disputed/Rejected
AR	Accepted/Rejected
ML	Machine Learning
PLMs	Pre-trained Language Models
CS	Computer Science
NLP	Natural Language Processing
BART	Bidirectional and Auto-Regressive Transformers
NeurIPS	The Neural Information Processing Systems Foundation
ICLR	The International Conference on Learning Representations
ICML	The International Conference on Machine Learning
ICCV	The International Conference on Computer Vision
CVPR	The IEEE/CVF Conference on Computer Vision and Pattern Recognition
ACL	Annual Meeting of the Association for Computational Linguistics
NAACL	North American Chapter of the Association for Computational Linguistics
EMNLP	Conference on Empirical Methods in Natural Language Processing
ECCV	European Conference on Computer Vision
IJCAI	International Joint Conferences on Artificial Intelligence
AISTATS	The International Conference on Artificial Intelligence and Statistics
BERT	Bidirectional Encoder Representation from Transformers
SPECTER	Scientific Paper Embeddings using Citation-informed TransformERs
SVM	Support Vector Machine
DT	Decision Trees
RF	Random Forests
KNN	K-nearest Neighbors
XGBoost	Extreme Gradient Boosting
GBC	Gradient Boosting for Classifiers
ABC	Adaptive Boosting for Classifiers
RoBERTa	Robustly optimized BERT approach
SHAP	SHapley Additive exPlanations

Table A.2
The descriptive statistical information of the two scores.

	The number of papers	Mean	Std	Min	Max
The average review scores	3810	5.25	1.20	1	8.80
The comprehensive quality score	3810	5.22	1.23	1	9.18

Note: The Spearman correlation coefficient for the two scores is 0.864 with a p-value of 0.000.

the paper. The formula for calculating the review score weighted by the confidence score is shown in Eq. (A.2):

$$Weight(Score_i) = \text{Weight}_i \times \text{Score}_i \tag{A.2}$$

In Eq. (A.2), $Weight(Score_i)$ represents the review score weighted by the confidence score, and $Score_i$ represents the review score given by the reviewers for the paper. Finally, the comprehensive quality score of the paper is shown in Eq. (A.3), $Quality(Score_i)$ represents the comprehensive quality score of the ith paper.

$$Quality(Score_i) = \frac{\sum_{j=1}^{n} Weight(Score_j)}{n} \tag{A.3}$$

To further compare the quality scores and the average review scores, we visualized each of the two with the acceptance and rejection of the paper and fitted normal curves, as shown in Fig. A.1. Descriptive statistics of the two scores are analyzed as shown in Table A.2.

Observing Fig. A.1, we found that the average review score and the comprehensive quality score were extremely similar in terms of the distribution of paper acceptance and rejection. However, in the region of overlap between the acceptance fit curve and the rejection fit curve, the number of papers measured by the comprehensive quality score was smaller relative to the average review score (602 < 619). Additionally, at the peaks of the two fitted curves, there were more papers measured by the comprehensive quality score than by the average review score (388 > 377, 301 > 297). In Table A.2, the means of the two scores were very similar and there was a significant positive correlation. However, the standard deviation and extreme values of the comprehensive quality score were slightly larger than the average review scores, implying a more dispersed distribution of the composite quality scores. Despite the nuances of these results, it can be seen that the comprehensive quality score was a more reasonable measure of a paper, was closer to the final decision of a paper, and better distinguished the quality of a paper than the average review score.

A.3. Detailed information of all features

The complete list of all the features and their detailed description are shown in Table A.3.

Table A.3
Features for evaluating scientific paper quality.

Quality standard	Feature name	Description	Feature type
Integrity	background_cites	Background citations	Coarse feature
	differences_cites	Difference citations	Coarse feature
	extends_cites	Extension citations	Coarse feature
	future_work_cites	Future work citations	Coarse feature
	motivation_cites	Motivation citations	Coarse feature
	similarities_cites	Similarity citations	Coarse feature
	uses_cites	Use citations	Coarse feature
	references	Number of references	Coarse feature
	have_appendix	Presence of appendix	Coarse feature
	references_in_three_years	Percentage of references in last 3 years	Coarse feature
Clarity	flesch_kincaid_grade_level	Readability index	Coarse feature
	smog_index	Readability index	Coarse feature
	coleman_liau_index	Readability index	Coarse feature
	ari_index	Readability index	Coarse feature
	linsear_write_formula	Readability index	Coarse feature
	gunning_fog_index	Readability index	Coarse feature
	figures	Number of figures	Coarse feature
	tables	Number of tables	Coarse feature
	formulas	Number of formulas	Coarse feature
	words	Number of words in the main text	Coarse feature
Novelty	new_nodes	Number of new nodes in the paper's knowledge network	Knowledge network
	new_edges	Number of new edges in the paper's knowledge network	Knowledge network
	knowledge_entity_pairs_novelty	Semantic novelty of the paper's knowledge entity pairs	Knowledge network
Significance	sd	Degree structure entropy	Knowledge network
	sb	Betweenness structure entropy	Knowledge network
	sr	Structure entropy ratio	Knowledge network
	ave_citation_of_references	Average citation count of references	Coarse feature
	nodes	Number of nodes in the paper's knowledge network	Knowledge network
	edges	Number of edges in the paper's knowledge network	Knowledge network
	avg_degree	Average degree in the paper's knowledge network	Knowledge network
	network_density	Density of the paper's knowledge network	Knowledge network
	avg_clustering_coefficient	Average clustering coefficient of the paper's knowledge network	Knowledge network
	max_betweenness_centrality	Maximum betweenness centrality in the paper's knowledge network	Knowledge network
	max_closeness_centrality	Maximum closeness centrality in the paper's knowledge network	Knowledge network
	max_eigenvector_centrality	Maximum eigenvector centrality in the paper's knowledge network	Knowledge network
	components	Number of connected components in the paper's knowledge network	Knowledge network

References

Abramo, G., D'Angelo, C.A., 2020. A novel methodology to assess the scientific standing of nations at field level. J. Inform. 14 (1), 100986. http://dx.doi.org/10.1016/j.joi.2019.100986.

Abrishami, A., Aliakbary, S., 2019. Predicting citation counts based on deep neural network learning techniques. J. Inform. 13 (2), 485–499. http://dx.doi.org/10.1016/j.joi.2019.02.011.

Amplayo, R.K., Hong, S., Song, M., 2018. Network-based approach to detect novelty of scholarly literature. Inf. Sci. 422, 542–557. http://dx.doi.org/10.1016/j.ins.2017.09.037.

Ante, L., 2022. The relationship between readability and scientific impact: Evidence from emerging technology discourses. J. Inform. 16 (1), 101252. http://dx.doi.org/10.1016/j.joi.2022.101252.

Basuki, S., Tsuchiya, M., 2022. The quality assist: A technology-assisted peer review based on citation functions to predict the paper quality. IEEE Access 10, 126815–126831. http://dx.doi.org/10.1109/ACCESS.2022.3225871.

Beltagy, I., Lo, K., Cohan, A., 2019. Scibert: A pretrained language model for scientific text. http://dx.doi.org/10.48550/arXiv.1903.10676, arXiv preprint arXiv:1903.10676.

Buckle, R.A., Creedy, J., 2019. The evolution of research quality in New Zealand universities as measured by the performance-based research fund process. N. Z. Econ. Pap. 53 (2), 144–165. http://dx.doi.org/10.1080/00779954.2018.1429486.

Checco, A., Bracciale, L., Loreti, P., Pinfield, S., Bianchi, G., 2021. AI-assisted peer review. Hum. Soc. Sci. Commun. 8 (1), 1–11. http://dx.doi.org/10.1057/s41599-020-00703-8.

Chen, J., Shao, D., Fan, S., 2021. Destabilization and consolidation: Conceptualizing, measuring, and validating the dual characteristics of technology. Res. Policy 50 (1), 104115. http://dx.doi.org/10.1016/j.respol.2020.104115.

Cohan, A., Feldman, S., Beltagy, I., Downey, D., Weld, D.S., 2020. Specter: Document-level representation learning using citation-informed transformers. http://dx.doi.org/10.48550/arXiv.2004.07180, arXiv preprint arXiv:2004.07180.

Devlin, J., Chang, M.-W., Lee, K., Toutanova, K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. http://dx.doi.org/10.48550/arXiv.1810.04805, arXiv preprint arXiv:1810.04805.

Ding, Y., Song, M., Han, J., Yu, Q., Yan, E., Lin, L., Chambers, T., 2013. Entitymetrics: Measuring the impact of entities. PLoS One 8 (8), e71416. http://dx.doi.org/10.1371/journal.pone.0071416.

van Dongen, T., Wenniger, G.M.d., Schomaker, L., 2020. SChuBERT: Scholarly document chunks with BERT-encoding boost citation count prediction. http://dx.doi.org/10.48550/arXiv.2012.11740, arXiv preprint arXiv:2012.11740.

Franceschini, F., Maisano, D., 2017. Critical remarks on the Italian research assessment exercise VQR 2011–2014. J. Inform. 11 (2), 337–357. http://dx.doi.org/10.1016/j.joi.2017.02.005.

Ghosal, T., Verma, R., Ekbal, A., Bhattacharyya, P., 2019. DeepSentiPeer: Harnessing sentiment in review texts to recommend peer review decisions. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 1120–1130. http://dx.doi.org/10.18653/v1/P19-1106.

Gomes, L., da Silva Torres, R., Côrtes, M.L., 2023. BERT-and TF-IDF-based feature extraction for long-lived bug prediction in FLOSS: A comparative study. Inf. Softw. Technol. 160, 107217. http://dx.doi.org/10.1016/j.infsof.2023.107217.

Hinze, S., Butler, L., Donner, P., McAllister, I., 2019. Different processes, similar results? A comparison of performance assessment in three countries. In: Springer Handbook of Science and Technology Indicators. Springer, pp. 465–484. http://dx.doi.org/10.1007/978-3-030-02511-3_18.

Hou, J., Wang, D., Li, J., 2022. A new method for measuring the originality of academic articles based on knowledge units in semantic networks. J. Inform. 16 (3), 101306. http://dx.doi.org/10.1016/j.joi.2022.101306.

Hu, Y.-H., Tai, C.-T., Liu, K.E., Cai, C.-F., 2020. Identification of highly-cited papers using topic-model-based and bibliometric features: the consideration of keyword popularity. J. Inform. 14 (1), 101004. http://dx.doi.org/10.1016/j.joi.2019.101004.

Huang, S., Huang, Y., Bu, Y., Lu, W., Qian, J., Wang, D., 2022. Fine-grained citation count prediction via a transformer-based model with among-attention mechanism. Inf. Process. Manage. 59 (2), 102799. http://dx.doi.org/10.1016/j.ipm.2021.102799.

Huisman, J., Smits, J., 2017. Duration and quality of the peer review process: the author's perspective. Scientometrics 113 (1), 633–650. http://dx.doi.org/10.1007/s11192-017-2310-5.

Kang, D., Ammar, W., Dalvi, B., Van Zuylen, M., Kohlmeier, S., Hovy, E., Schwartz, R., 2018. A dataset of peer reviews (peerread): Collection, insights and nlp applications. http://dx.doi.org/10.48550/arXiv.1804.09635, arXiv preprint arXiv:1804.09635.

Kinney, R., Anastasiades, C., Authur, R., Beltagy, I., Bragg, J., Buraczynski, A., Cachola, I., Candra, S., Chandrasekhar, Y., Cohan, A., et al., 2023. The semantic scholar open data platform. http://dx.doi.org/10.48550/arXiv.2301.10140, arXiv preprint arXiv:2301.10140.

Kousha, K., Thelwall, M., 2023. Factors associating with or predicting more cited or higher quality journal articles: An annual review of information science and technology (ARIST) paper. J. Assoc. Inf. Sci. Technol. http://dx.doi.org/10.1002/asi.24810.

La Quatra, M., Cagliero, L., 2022. Transformer-based highlights extraction from scientific papers. Knowl.-Based Syst. 252, 109382. http://dx.doi.org/10.1016/j.knosys.2022.109382.

Lauscher, A., Ko, B., Kuehl, B., Johnson, S., Jurgens, D., Cohan, A., Lo, K., 2021. MultiCite: Modeling realistic citations requires moving beyond the single-sentence single-label setting. http://dx.doi.org/10.48550/arXiv.2107.00414, arXiv preprint arXiv:2107.00414.

Li, J., Chen, J., 2022. Measuring destabilization and consolidation in scientific knowledge evolution. Scientometrics 127 (10), 5819–5839. http://dx.doi.org/10.1007/s11192-022-04479-3.

Liang, Z., Liu, F., Mao, J., Lu, K., 2021. A knowledge representation model for studying knowledge creation, usage, and evolution. In: International Conference on Information. Springer, pp. 97–111. http://dx.doi.org/10.1007/978-3-030-71292-1_9.

Liang, W., Zhang, Y., Cao, H., Wang, B., Ding, D., Yang, X., Vodrahalli, K., He, S., Smith, D., Yin, Y., et al., 2023. Can large language models provide useful feedback on research papers? A large-scale empirical analysis. arXiv preprint arXiv:2310.01783.

Lin, J., Qilin, Z., 2020. Research on academic evaluation based on fine-grain citation sentimental quantification. Data Anal. Knowl. Discov. 4 (6), 129–138. http://dx.doi.org/10.11925/infotech.2096-3467.2019.0967.

Lin, J., Song, J., Zhou, Z., Chen, Y., Shi, X., 2023. Automated scholarly paper review: Concepts, technologies, and challenges. Inf. Fusion 98, 101830. http://dx.doi.org/10.1016/j.inffus.2023.101830.

Lin, K.L., Sui, S.X., 2020. Citation functions in the opening phase of research articles: A corpus-based comparative study. In: Corpus-based Approaches to Grammar, Media and Health Discourses: Systemic Functional and Other Perspectives. Springer, pp. 233–250. http://dx.doi.org/10.1007/978-981-15-4771-3_10.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V., 2019. Roberta: A robustly optimized bert pretraining approach. http://dx.doi.org/10.48550/arXiv.1907.11692, arXiv preprint arXiv:1907.11692.

Lu, C., Bu, Y., Wang, J., Ding, Y., Torvik, V., Schnaars, M., Zhang, C., 2019. Examining scientific writing styles from the perspective of linguistic complexity. J. Assoc. Inf. Sci. Technol. 70 (5), 462–475. http://dx.doi.org/10.1002/asi.24126.

Luan, Y., He, L., Ostendorf, M., Hajishirzi, H., 2018. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. http://dx.doi.org/10.48550/arXiv.1808.09602, arXiv preprint arXiv:1808.09602.

Lundberg, S.M., Lee, S.-I., 2017. A unified approach to interpreting model predictions. Adv. Neural Inf. Process. Syst. 30.

Luo, Z., Lu, W., He, J., Wang, Y., 2022. Combination of research questions and methods: A new measurement of scientific novelty. J. Inform. 16 (2), 101282. http://dx.doi.org/10.1016/j.joi.2022.101282.

Ma, Y., Liu, J., Lu, W., Cheng, Q., 2023. From "what" to "how": Extracting the procedural scientific information toward the metric-optimization in AI. Inf. Process. Manage. 60 (3), 103315. http://dx.doi.org/10.1016/j.ipm.2023.103315.

Marsh, H.W., Bazeley, P., 1999. Multiple evaluations of grant proposals by independent assessors: Confirmatory factor analysis evaluations of reliability, validity, and structure. Multivar. Behav. Res. 34 (1), 1–30. http://dx.doi.org/10.1037/0003-066X.63.3.160.

Min, C., Bu, Y., Sun, J., 2021. Predicting scientific breakthroughs based on knowledge structure variations. Technol. Forecast. Soc. Change 164, 120502. http://dx.doi.org/10.1016/j.techfore.2020.120502.

Ribeiro, A.C., Sizo, A., Lopes Cardoso, H., Reis, L.P., 2021. Acceptance decision prediction in peer-review through sentiment analysis. In: EPIA Conference on Artificial Intelligence. Springer, pp. 766–777. http://dx.doi.org/10.1007/978-3-030-86230-5_60.

Roman Jurowetzki, H.B., 2022. SciNERTopic - NER enhanced transformer-based topic modelling for scientific text. http://dx.doi.org/10.57967/hf/0095, URL https://huggingface.co/RJuro/SciNERTopic.

Shi, X., Liu, Y., Liu, J., Cheng, Q., Lu, W., 2024. Integrity verification for scientific papers: The first exploration of the text. Expert Syst. Appl. 237, 121488. http://dx.doi.org/10.1016/j.eswa.2023.121488.

Singh, A., D'Arcy, M., Cohan, A., Downey, D., Feldman, S., 2022. SciRepEval: A multi-format benchmark for scientific document representations. http://dx.doi.org/10.48550/arXiv.2211.13308, arXiv preprint arXiv:2211.13308.

Spezi, V., Wakeling, S., Pinfield, S., Fry, J., Creaser, C., Willett, P., 2018. "Let the community decide"? The vision and reality of soundness-only peer review in open-access mega-journals. J. Doc. 74 (1), 137–161. http://dx.doi.org/10.1108/JD-06-2017-0092.

Sun, M., Barry Danfa, J., Teplitskiy, M., 2022. Does double-blind peer review reduce bias? Evidence from a top computer science conference. J. Assoc. Inf. Sci. Technol. 73 (6), 811–819. http://dx.doi.org/10.1002/asi.24582.

Thelwall, M., 2022. Can the quality of published academic journal articles be assessed with machine learning? Quant. Sci. Stud. 3 (1), 208–226. http://dx.doi.org/10.1162/qss_a_00185.

Thelwall, M., Kousha, K., Stuart, E., Makita, M., Abdoli, M., Wilson, P., Levitt, J., 2023a. In which fields are citations indicators of research quality? J. Assoc. Inf. Sci. Technol. http://dx.doi.org/10.1002/asi.24767.

Thelwall, M., Kousha, K., Wilson, P., Makita, M., Abdoli, M., Stuart, E., Levitt, J., Knoth, P., Cancellieri, M., 2023b. Predicting article quality scores with machine learning: The UK research excellence framework. Quant. Sci. Stud. 4 (2), 547–573. http://dx.doi.org/10.1162/qss_a_00258.

Uzzi, B., Mukherjee, S., Stringer, M., Jones, B., 2013. Atypical combinations and scientific impact. Science 342 (6157), 468–472. http://dx.doi.org/10.1126/science.1240474.

Vincent-Lamarre, P., Larivière, V., 2021. Textual analysis of artificial intelligence manuscripts reveals features associated with peer review outcome. Quant. Sci. Stud. 2 (2), 662–677. http://dx.doi.org/10.1162/qss_a_00125.

Wang, S., Ma, Y., Mao, J., Bai, Y., Liang, Z., Li, G., 2023. Quantifying scientific breakthroughs by a novel disruption indicator based on knowledge entities. J. Assoc. Inf. Sci. Technol. 74 (2), 150–167. http://dx.doi.org/10.1002/asi.24719.

Wang, K., Wan, X., 2018. Sentiment analysis of peer review texts for scholarly papers. In: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. pp. 175–184. http://dx.doi.org/10.1145/3209978.3210056.

Wang, Z., Wang, K., Liu, J., Huang, J., Chen, H., 2022. Measuring the innovation of method knowledge elements in scientific literature. Scientometrics 127 (5), 2803–2827. http://dx.doi.org/10.1007/s11192-022-04350-5.

Wang, Q., Zeng, Q., Huang, L., Knight, K., Ji, H., Rajani, N.F., 2020. ReviewRobot: Explainable paper review generation based on knowledge synthesis. http://dx.doi.org/10.48550/arXiv.2010.06119, arXiv preprint arXiv:2010.06119.

Wenniger, G.M.d., van Dongen, T., Aedmaa, E., Kruitbosch, H.T., Valentijn, E.A., Schomaker, L., 2020. Structure-tags improve text classification for scholarly document quality prediction. http://dx.doi.org/10.18653/v1/2020.sdp-1.18, arXiv preprint arXiv:2005.00129.

Wilsdon, J., 2016. The metric tide: Independent review of the role of metrics in research assessment and management. Metric Tide 1–192. http://dx.doi.org/10.4135/9781473978782.

Wu, L., Wang, D., Evans, J.A., 2019. Large teams develop and small teams disrupt science and technology. Nature 566 (7744), 378–382. http://dx.doi.org/10.1038/s41586-019-0941-9.

Xu, L., Ding, K., Lin, Y., Zhang, C., 2023. Does citation polarity help evaluate the quality of academic papers? Scientometrics 1–23. http://dx.doi.org/10.1007/s11192-023-04734-1.

Xu, S., Hao, L., Yang, G., Lu, K., An, X., 2021. A topic models based framework for detecting and forecasting emerging technologies. Technol. Forecast. Soc. Change 162, 120366. http://dx.doi.org/10.1016/j.techfore.2020.120366.

Xu, H., Luo, R., Winnink, J., Wang, C., Elahi, E., 2022. A methodology for identifying breakthrough topics using structural entropy. Inf. Process. Manage. 59 (2), 102862. http://dx.doi.org/10.1016/j.ipm.2021.102862.

Xue, Z., He, G., Liu, J., Jiang, Z., Zhao, S., Lu, W., 2023. Re-examining lexical and semantic attention: Dual-view graph convolutions enhanced BERT for academic paper rating. Inf. Process. Manage. 60 (2), 103216. http://dx.doi.org/10.1016/j.ipm.2022.103216.

Yang, P., Sun, X., Li, W., Ma, S., 2018. Automatic academic paper rating based on modularized hierarchical convolutional neural network. http://dx.doi.org/10.48550/arXiv.1805.03977, arXiv preprint arXiv:1805.03977.

Yuan, W., Liu, P., Neubig, G., 2022. Can we automate scientific reviewing? J. Artificial Intelligence Res. 75, 171–212. http://dx.doi.org/10.1613/jair.1.12862.

Zhang, Q., Li, M., 2022. A betweenness structural entropy of complex networks. Chaos Solitons Fractals 161, 112264. http://dx.doi.org/10.1016/j.chaos.2022.112264.

Zhang, Y., Wang, M., Saberi, M., Chang, E., 2020. Knowledge fusion through academic articles: a survey of definitions, techniques, applications and challenges. Scientometrics 125, 2637–2666. http://dx.doi.org/10.1007/s11192-020-03683-3.

Zhang, Y., Wu, M., Miao, W., Huang, L., Lu, J., 2021. Bi-layer network analytics: A methodology for characterizing emerging general-purpose technologies. J. Inform. 15 (4), 101202. http://dx.doi.org/10.1016/j.joi.2021.101202.

Zhao, Q., Feng, X., 2022. Utilizing citation network structure to predict paper citation counts: A deep learning approach. J. Inform. 16 (1), 101235. http://dx.doi.org/10.1016/j.joi.2021.101235.